# Terminology extraction from English-Portuguese and English-Galician parallel corpora based on probabilistic translation dictionaries and bilingual syntactic patterns

*Alberto Simões*

Department of Computer Science
Universidade do Minho
ambs@di.uminho.pt

*Xavier Gómez Guinovart*

Department of Translation and Linguistics
Universidade de Vigo
xgg@uvigo.es

## Abstract

This paper presents a research on parallel corpora-based bilingual terminology extraction based on the occurrence of bilingual morphosyntactic patterns in the probabilistic translation dictionaries generated by NATools. To evaluate this method, we carried out an experiment in which both the level of lexical cohesion of the term candidates and their specificity with respect to a non-terminological corpus of the target language were taken into account. The evaluation results show a high degree of accuracy of the terminology extraction based on probabilistic translation dictionaries complemented by bilingual syntactic patterns.

**Index Terms**: bilingual terminology extraction, probabilistic translation dictionaries

## 1. Introduction

This paper presents a research on parallel corpora-based bilingual terminology extraction based on the occurrence of bilingual morphosyntactic patterns in the probabilistic translation dictionaries generated by NATools. NATools[1] is an open source workbench for parallel corpora processing which includes a sentence aligner, a probabilistic translation dictionaries extractor, a word aligner, a terminology extractor, and a set of other tools to study the aligned parallel corpora. To evaluate the method used by NATools, we carried out an experiment in which both the level of lexical cohesion of the term candidates and their specificity with respect to a non-terminological corpus of the target language were taken into account. Testing was conducted for the language pairs English-Galician and English-Portuguese using the corpus of the Unesco Courier and the JRC-Acquis, respectively. The evaluation results show a high degree of accuracy of the terminology extraction based on probabilistic translation dictionaries complemented by bilingual syntactic patterns.

## 2. Terminology Extraction

The extraction algorithm used by NATools is based on translation patterns containing the most commonly found grammatical bilingual combinations for terminological units. As a help to detect the term relevance, we calculate the log-likelihood ratio for each term and the translation probability in the corpus for each candidate pair of bilingual terminological equivalents.

[1] http://natools.sourceforge.net/

### 2.1. Extraction Algorithm

The terminology extraction algorithm used in this study is based on NATools probabilistic translation dictionaries [1]. These dictionaries are extracted automatically from sentence aligned parallel corpora. The resulting dictionaries are mappings from words in a language to a set of probable translations in other language. Each of these translations have a probabilistic measure of translatability.

This information enables to create an alignment matrix for any translation unit, either from that same corpora or from a different one. These translation matrixes include in each cell the mutual translation probability for each word combination (from the source/target language). [2] provides a detailed explanation of the matrix construction, and how it can be used to extract simple translation examples.

These same matrixes can be used to extract bilingual terminology using translation patterns. These patterns specify how word order in the source language changes after translation takes place.

| | Human | Rights |
|---|---|---|
| Direitos | | $X$ |
| do | | |
| Homem | $X$ | |

Figure 1: Example of translation pattern: `A "de" B = B A`

Figure 1 illustrates an alignment pattern and its visual representation. This pattern can be read as: $\mathcal{T}(A \cdot \text{“de”} \cdot B) = \mathcal{T}(B) \cdot \mathcal{T}(A)$ Each $X$ in the table represents an anchor: it corresponds to a high translation probability.

These patterns are searched in the translation matrix, matching on anchor cells, as shown in figure 2. These cells need to have a probability value higher than 20% of the remaining column and row cells to be considered anchor cells.

Translation patterns may include morphological restrictions defining the morphological categories allowed for the words matching the pattern. Each variable on the right side is followed by a morphological restriction in square brackets `[...]`. NATools relies on external morphological analyzers to validate the morphological restrictions.

There are several morphological analyzer engines and, sometimes, different languages require different morphological analyzers. For instance, for our experiments we needed a morphological analyzer for Portuguese and for Galician. While

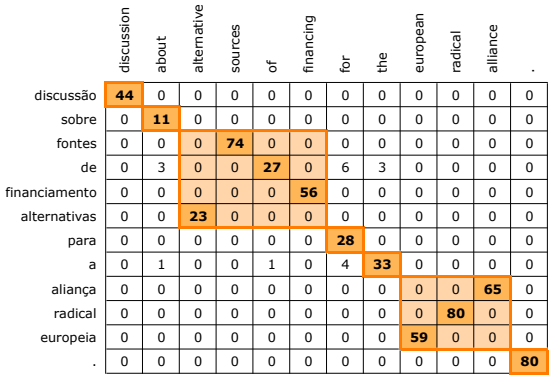


Figure 2: Alignment matrix for a Portuguese–English translation unit with marked patterns.

jSpell [3] has a dictionary for Portuguese, it lacks a dictionary for Galician. In the same way, FreeLing [4] has a dictionary for Galician, but it does not include a good Portuguese one.

In order to help integrate NATools with external morphological analyzers we need to create an interface tool for each morphological analyzer. This tool should be able to receive words (one per line) and to return an analysis of such words (one per word and on a single line).

For instance, when calling the interface to the JSpell Portuguese dictionary with the word *pode* (an ambiguous word), the interface returns:

```
[{CAT=>'v',T=>'p',N=>'s',P=>'3',rad=>'poder'},
 {CAT=>'v',T=>'i',N=>'s',P=>'2',rad=>'poder'},
 {CAT=>'v',T=>'pc',N=>'s',P=>'1_3',rad=>'podar'},
 {CAT=>'v',T=>'i',N=>'s',P=>'3',rad=>'podar'}]
```

This output should appear on a single line, and its syntax should be correct (it should be a valid Perl data-structure). The keys are completely irrelevant for NATools as far as they are the same ones used in the translation pattern definition.

For each variable containing a morphological restriction the system will invoke the morphological analyzer and ask for the specific word analysis. If any of the analysis match the required restrictions the system will continue validating words.

If the pattern matches (anchor cells exist in the specified position) and the morphological analysis are adequate, that block is marked as *used*, and the string pair presented.

### 2.2. Terminology metrics

#### 2.2.1. Translation Probability

We calculate a translation probability measure for each candidate pair of bilingual terminological equivalents. This value is based on the translation probabilities for each word pair, discarding probabilities for stop-words translation.

Considering the previous pattern example, `A "de" B = B A` the translation probability is measured as the average of the mutual translation probability of the words matching the variables $A$ and $B$.

#### 2.2.2. Log-likelihood

There are different well-known techniques for scoring the candidate terms [5]. Following many other works on term extraction based on [6], we score each candidate using the log-likelihood measure, which is computed using the `Text::NSP` Perl module.[2]

Considering that the module only supports bigrams and trigrams, for bigger terms this measure is computed as the minimum value for the partial trigrams [7].

## 3. Experiments

Our experiments focused on two language pairs: English–Galician and English–Portuguese. This choice can be explained by the proximity of the two target languages. Moreover, the availability of bigger corpora for the English–Portuguese language pair made the evaluation more relevant.

### 3.1. Parallel corpora and exclusion corpora

This section describes the parallel corpora used for the terminology extraction, and the monolingual corpora used for word bi- and trigrams exclusion, and extraction evaluation.

#### 3.1.1. Parallel Corpora

For the terminology extraction experiments we used two pairs of parallel corpora, English–Galician and English–Portuguese, of very different sizes.

| Corpus | Unesco | JRC-Acquis |
|---|---|---|
| Trans. Units | 47 903 | 1 315 907 |
| Source Tokens | 1 057 556 | 37 605 596 |
| Target Tokens | 1 019 886 | 51 075 535 |
| Source Forms | 50 866 | 283 061 |
| Target Forms | 66 515 | 295 923 |

Table 1: Used Parallel corpora

The *Unesco Corpus* is a collection of 30 issues (from the period 1998-2001) of the *Unesco Courier*[3] in four languages (English, Galician, French and Spanish) which is part of the CLUVI Parallel Corpus[4] [8]. Created in August 1947, the *Unesco Courier* is a monthly publication which reflects Unesco's concerns and thoughts in articles from around the world. Each issue consists of a thematic dossier that treats one of Unesco's scientific and cultural concerns, as endangered languages, world heritage, immigration, bioethics or the spell of sport. As a whole, the *Unesco Courier* contains a high density of terminological units from the fields of sociology and social sciences.

The *JRC-Acquis* is the total body of European Union law applicable in the EU Member States. This parallel corpus in 22 languages is maintained by the Language Technology group of the European Commission's Joint Research Centre. This collection of legislative text changes continuously and currently comprises selected texts written between the 1950s and the present time. For the purpose of this work we used JRC-Acquis v3 [9], the latest version available, for the English–Portuguese language pair.

#### 3.1.2. Exclusion Corpora

Two literary corpora were used in the evaluation process, particularly for bigrams and trigrams exclusion.

The *BiVir Corpus*[5] is a Galician literary corpus containing 30 fiction works (namely romans) from the Virtual Library of

[2] http://ngram.sourceforge.net/
[3] http://www.unesco.org/courier/
[4] http://sli.uvigo.es/CLUVI/
[5] http://www.bivir.com/

| Corpus | BiVir | Compara |
|---|---|---|
| Tokens | 1 008 125 | 1 714 523 |
| Bigrams | 361 547 | 544 274 |
| Trigrams | 641 349 | 1 243 356 |

Table 2: Exclusion corpora

**EN-GL patterns using FreeLing tags**

```
[R1] A B = B[CAT<-/^NC/] A[CAT<-/^AQ0/];
[R2] A B = B[CAT<-/^NC/] "de"|"do"|"da"|
  "dos"|"das" A[CAT<-/^NC/];
[R3] A "of"|"in"|"for" B = A[CAT<-/^NC/]
  "de"|"do"|"da"|"dos"|"das" B[CAT<-/^NC/];
[R4] A B C = C[CAT<-/^NC/] A[CAT<-/^AQ0/]
  B[CAT<-/^AQ0/];
```

**EN-PT patterns using JSpell tags**

```
[R1] A B = B[CAT<-/nc/] A[CAT<-/(a_nc|adj)/];
[R2] A B = B[CAT<-/nc/] "de"|"do"|"da"|"dos"|
  "das" A[CAT<-/(a_nc|nc)/];
[R3] A "of"|"in"|"for" B = A[CAT<-/nc/] "de"|
  "do"|"da"|"dos"|"das" B[CAT<-/nc/];
[R4] A B C = C[CAT<-/nc/] A[CAT<-/(adj|a_nc)/]
  B[CAT<-/(adj|a_nc)/];
```

Figure 3: EN-GL and EN-PT bilingual syntactic patterns

Universal Literature in Galician language and mantained by the Association of Galician Translators.

*Compara*[6] [10] is a large human-edited English-Portuguese parallel corpus whose sentence alignment, sentence separation, lemmatization and POS tagging have been revised by human annotators (in fact, lemmatization and tagging have been checked and corrected by hand only for Portuguese so far). Compara contains 75 fiction texts and their translations, corresponding to approximately 1.5 million words in each language.

### 3.2. Translation Patterns

In order to evaluate the precision of the NATools-based term extraction algorithm, four translation patterns have been extracted, using the morphological analyzer of FreeLing for Galician and Jspell for Portuguese.

Translation patterns for Galician (with FreeLing analyzer) and for Portuguese (with JSpell analyzer) are shown in figure 3. Tables 3 and 4 show the top occurring entries extracted using these rules.

| English (and LLR) | | Galician (and LLR) | | Prb | Oc. |
|---|---|---|---|---|---|
| united states | 4 701 | estados unidos | 9 286 | 53.7 | 265 |
| human rights | 4 942 | dereitos humanos | 3 904 | 68.3 | 207 |
| united nations | 2 462 | nacións unidas | 5 130 | 47.4 | 125 |
| world bank | 1 490 | banco mundial | 1 809 | 60.0 | 114 |
| security council | 467 | consello de seguridade | 1 023 | 69.2 | 26 |
| street children | 342 | nenos da rúa | 700 | 60.7 | 22 |
| market economy | 268 | economía de mercado | 492 | 67.7 | 19 |
| life expectancy | 304 | esperanza de vida | 852 | 51.6 | 18 |

Table 3: EN-GL top-occurring term candidates from the Unesco Corpus

## 4. Filtering and evaluation

Different methods are used for filtering the results of term extraction: identification of unlikely term candidates because of their similarity with a lexical pattern, ranking of candidates by virtue of some score of lexical association, and assessment of term specificity with respect to some kind of non-terminological corpus of the language, among others [11].

[6] http://www.linguateca.pt/COMPARA/

| English (and LLR) | | Portuguese (and LLR) | | Prob. | Oc. |
|---|---|---|---|---|---|
| european union | 231 965 | união europeia | 311 030 | 65.24 | 12 465 |
| european parliament | 205 297 | parlamento europeu | 267 379 | 63.31 | 13 066 |
| european community | 136 471 | comunidade europeia | 224 132 | 57.48 | 18 251 |
| european communities | 265 877 | comunidades europeias | 284 409 | 53.51 | 19 545 |
| council decision | 43 760 | decisão do conselho | 398 348 | 58.80 | 1 665 |
| commission decision | 32 322 | decisão da comissão | 264 191 | 43.73 | 2 215 |
| basic regulation | 61 569 | regulamento de base | 103 700 | 63.75 | 3 390 |
| management committee | 36 170 | comité de gestão | 83 014 | 69.79 | 3 549 |

Table 4: EN-PT top-occurring term candidates from the JRC-Acquis Corpus

With the first filtering method, term candidates beginning or ending with any of the words of a list of stop words are removed from the list. This is the approach used by the Corpógrafo [12]. This method, however, does not apply to the results of NATools complemented with bilingual syntactic patterns, since term candidates generated by NATools match the patterns specified by particular morphosyntactic rules, which means that they never begin or end with a stop word.

Another well-known method for filtering the results of term extraction consists of calculating the lexical association of candidates in the corpus using one of the possible scores to test the strength of this attraction, such as the Mutual Information [13] and the log-likelihood ratio [6]. One of the most widely used scores for terminology extraction is the log-likelihood ratio, which is the score calculated by the term extractor in NATools. However, this score does not carry any significance as a discriminatory factor when assessing the outcome of the terminology extraction by NATools with bilingual syntactic patterns, presumably because the quality of selection based on a probabilistic translation dictionary derived from the parallel corpus and filtered with patterns ensures a fairly high minimum cohesion between the components of the candidate terms.

Thus, we decided to check the accuracy of the term extraction of NATools with bilingual syntactic patterns using a non-terminological corpus of exclusion as a filter. The exclusion corpus will determine the identification (and exclusion) of unlikely term candidates. Literary corpora, unlike corpora of news articles, for instance, usually contain very few terminological units. A literary corpus, as a corpus of exclusion for term extraction, represents a very safe filter. When using a literary corpus as a filter, there are more false candidates identified as such than correct candidates wrongly identified as false ones. We created lists of word n-grams from the exclusion corpora BiVir and Compara (see above), and applied these lists as criteria for filtering and evaluation of NATools-based terminology extraction. The results are discussed in the next section.

### 4.1. Experiment Results

The evaluation results (table 5) point to a high precision of the NATools-based extraction algorithm. As shown in the first column of the table, the 12,689 translation equivalences (TE) identified in the Unesco Corpus using NATools with the EN-GL bilingual syntactic patterns depicted in figure 3 represent 7,250 candidate bilingual term pairs (term candidates or TC) (57% of TE) after eliminating repeated TE. When filtering that list of TC with the list of word bi- and trigrams from the BiVir Cor-

pus, we obtain a list of 6,949 Galician terms from TC (corresponding to 96% of TC) which are not present in the exclusion corpus, and a complementary list of 301 Galician term candidates (only 4% of TC) identified as erroneous term candidates due to their presence in the exclusion corpus. Thus, these scores show a precision of 96% in the NATools-based term extraction from the Unesco Corpus.

As for the experiments with the JRC-Acquis Corpus, the 717,293 TE identified with the EN-PT bilingual syntactic patterns shown in figure 3 represent 72,952 TC (only 10.2% of TE) after eliminating repeated TE. Differences between the TE/TC ratio of the Unesco Corpus and and that of the JRC-Acquis Corpus (57% vs. 10.2%) lie in the lexical density (percentage of different words in a text) of the two corpora. When filtering that list of TC with the list of n-grams from the Compara Corpus, we get a list of 63,744 Portuguese terms from TC (corresponding to 87.4% of TC) which are not present in the exclusion corpus, and a complementary list of 6,949 Portuguese term candidates (12.6% of TC) identified as unlikely term candidates because of their presence in the exclusion corpus. Differences in the precision scores of term extraction between the Unesco Corpus and the JRC-Acquis Corpus (96% vs. 87.4%) lie in the different size of the corpora (and of the exclusion corpora) and also in their level of lexical density and terminological specificity.

| **Corpora** | **Unesco** | **JRC-Acquis** |
|---|---|---|
| Language | GL | PT |
| Trans. Equiv. | 12 689 | 717 293 |
| Term Cand. | 7 250 (57%) | 72 952 (10.2%) |
| Excluded TC | 301 (4%) | 9 208 (12.6%) |
| Not-excl. TC | 6 949 (96%) | 63 744 (87.4%) |

Table 5: Extraction results

Moreover, the evaluation undergone has shown that the log-likelihood ratio (LLR) may be significant as a score to rank the "terminological quality" of candidates belonging to a language from one corpus. However, the LLR cannot be used for comparing the quality of the term candidates extracted from two different sized corpora, being LLR dependent upon the size of corpora, as shown in table 6.

| | **Log-likelihood ratio** | | | |
|---|---|---|---|---|
| | *Unesco* | | *JRC-Acquis* | |
| | EN | GL | EN | PT |
| min | 0 | 0 | 0 | 0 |
| max | 4 942 | 9 286 | 448 664 | 613 529 |
| mean | 75 | 153 | 3 486 | 9 734 |
| stddev | 238 | 415 | 10 666 | 32 238 |

Table 6: Log-likelihood ratio statistics

| | **Translation prob.** | |
|---|---|---|
| | *Unesco* | *JRC-A.* |
| | EN–GL | EN–PT |
| min | 10.25 | 10.03 |
| max | 85.57 | 94.53 |
| mean | 49.60 | 53.00 |
| stddev | 13.92 | 13.02 |

Table 7: Translation probability statistics

Finally, differences in translation probabilities of bilingual term candidates from the two parallel corpora (mean values of 49.60 vs. 53) point to a highly homogenous extraction of candidates with respect to translation probability, in spite of the highly heterogenous characteristics of corpora, in table 7.

# 5. Conclusions

Bilingual terminology extraction from parallel corpora based on probabilistic translation dictionaries and complemented with bilingual syntactic patterns shows high rates of accuracy. In the experiments described here this ratio is between 87.4% and 96% depending on the characteristics of the corpus. Considering that this method of extraction is dependent on POS-taggers accuracy, an erroneous tagging may lead to false candidates. Thus, improvement in tagging results brings about an improvement in the performance of terminology extraction.

# 6. Acknowledgments

This work has been funded by the Ministerio de Educación y Ciencia and the Fondo Europeo de Desenvolvemento Rexional (FEDER) within the project "Diseño e implementación de un servidor de recursos integrados para el desarrollo de tecnologías de la lengua gallega (RILG)" (HUM2006-11125-C02-01/FILO), and by the Consellaría de Innovación e Industria da Xunta de Galicia within the project "Desenvolvemento e aplicación de recursos integrados da lingua galega" (ref. INCITE08PXIB302185PR).

# 7. References

[1] Simões, A. and Almeida, J.J., "NATools: A Statistical Word Aligner Workbench", Procesamiento del Lenguaje Natural, 31, 217-224, 2003.

[2] Simões, A. and Almeida, J.J., "Combinatory Examples Extraction for Machine Translation", in Proc. of the 11th Annual Conference of the European Association for Machine Translation, 27-32, 2006.

[3] Almeida, J.J. and Pinto, U. "Jspell: um módulo para análise léxica genérica de linguagem natural, in Actas do X Encontro da Associação Portuguesa de Linguística, 1-15, 1994.

[4] Atserias, J. et al., "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library", in Proc. of the 5th International Conference on Language Resources and Evaluation, 48-55, 2006.

[5] Daille, B., "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology", in J. Klavans and P. Resnik [Ed] The Balancing Act: Combining Symbolic and Statistical Approaches to Language, 49-66, The MIT Press, 1996.

[6] Dunning, T., "Accurate methods for the statistics of surprise and coincidence", Computational Linguistics, 19, 61-74, 1993.

[7] Patry, A. and Langlais, P., "Corpus-Based Terminology Extraction", in Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, 313-321, 2005.

[8] Gómez Guinovart, X. and Sacau, E., "Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo), in Proceedings of the 4th International Conference on Language Resources and Evaluation, 1179–1182, 2004.

[9] Steinberger, R. et al., "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in Proc. of the 5th International Conference on Language Resources and Evaluation, 2006.

[10] Frankenberg-Garcia, A. and Santos, D., "Introducing COMPARA, the Portuguese-English parallel translation corpus, in F. Zanettin, S. Bernardini and D. Stewart [Ed] Corpora in Translation Education, St. Jerome Publishing, 71-87, 2003.

[11] Munpyo Hong, M., Fissaha, S. and Haller, J., "Hybrid filtering for extraction of term candidates from German technical texts, in Proceedings of Terminologie et Intelligence Artificielle, 2001.

[12] Sarmento, L. et al.,"Corpógrafo V3: From Terminological Aid to Semi-automatic Knowledge Engine", in Proc. of the 5th International Conference on Language Resources and Evaluation, 1502-1505, 2006.

[13] Church, K. and Hanks, P., "Word association norms, mutual information, and lexicography", Computational Linguistics, 16(1), 22-29, 1990.