# Automatic creation of WordNets from parallel corpora

**Antoni Oliver, Salvador Climent**

Universitat Oberta de Catalunya
Barcelona - Catalonia - (Spain)
aoliverg@uoc.edu, scliment@uoc.edu

### Abstract

In this paper we present the evaluation results for the creation of WordNets for five languages (Spanish, French, German, Italian and Portuguese) using an approach based on parallel corpora. We have used three very large parallel corpora for our experiments: DGT-TM, EMEA and ECB. The English part of each corpus is semantically tagged using Freeling and UKB. After this step, the process of WordNet creation is converted into a word alignment problem, where we want to align WordNet synsets in the English part of the corpus with lemmata on the target language part of the corpus. The word alignment algorithm used in these experiments is a simple most frequent translation algorithm implemented into the WN-Toolkit. The obtained precision values are quite satisfactory, but the overall number of extracted synset-variant pairs is too low, leading into very poor recall values. In the conclusions, the use of more advanced word alignment algorithms, such as Giza++, Fast Align or Berkeley aligner is suggested.

**Keywords:** WordNet, Expand Model, parallel corpus

## 1. Introduction

Vossen (1998) described two basic approaches for encoding semantic relations when building the WordNets for the languages in the EuroWordNet project: the *Merge Model* and the *Expand Model*. By the former, after having defined a set of nodes (*synsets*) for the ontology, they were connected internally by semantic relations; in a second phase, the WordNet thus created could be appropriately linked to other languages' WordNets. The other way round, the Expand Model took advantage of the fact that there already was a fully developed WordNet available and connected with relations: the original English WordNet developed in Princeton University (PWN). By this approach, initially followed only by the Spanish WordNet, a set of synsets were first selected in PWN which were then translated using bilingual dictionaries. Proceeding this way, the semantic-relation structure of PWN was fully exported to the Spanish WordNet so that both internal structure and cross-linguistic connection to PWN *came for free*.

As time went by, WordNets established as the most popular lexical-semantic ontology in the world, dozens of new WordNets were built for new languages and the Expand and the Merge Model were assumed to be not just two ways of encoding relations but the two paradigms for building and connecting WordNets semiautomatically. Both have been used for this purpose, as well as mixed methods. Besides, a number of WordNets of different languages are being connected in a free global grid (Horák et al., 2008).

It is often assumed that the Merge Model is the most accurate since it allows to better reflect language-specific lexical semantics. Unfortunately, it has two important drawbacks. First, the internal semantic linking has mainly to be performed manually, therefore it is extremely time and resource consuming. Second, once built the local ontology, it is absolutely not trivial to link it appropriately to other language ontologies (Ngai and Fung, 2002).

Severe limitation in resources and manpower is more a rule than an exception in research nowadays, hence many groups and projects such as BalkaNet and MultiWordNet accept the cost of having a somewhat biased WordNet in exchange for ease of development (Erjavec and Fišer, 2006). Moreover, it has to be noticed that the Expand method presents several other benefits beyond being quicker and cheaper. First, one gets a WordNet not dramatically different in size to PWN or other large WordNets, so comparable tasks can be carried out. Second, the new WordNet need not necessarily be the final release: it can be further developed and improved either manually or automatically; it can be readily used for a number of tasks such as information retrieval or summarization (De Melo and Weikum, 2008). And last but not least, the new WordNet has steady access to a number of important semantic resources linked to PWN such as SUMO (Niles and Pease, 2003), the Top Concept Ontology (Álvez et al., 2008), WordNet Domains (Magnini and Cavaglia, 2000), WordNet Base Concepts (Izquierdo et al., 2007) and to other language's WordNets via the Global WordNet Grid so that many more sophisticated monolingual and multilingual tasks (e.g. Cross-Linguistic Information Retrieval) can be immediately performed.

This paper presents a methodology and algorithm to create WordNets from parallel corpora. We also present a toolkit that implements this methodology along with methodologies based on dictionaries for the creation of WordNets. These programs have been successfully used in the projects Know2 and SKATER for the creation and improvement of the Catalan and Spanish WordNet 3.0. The toolkit is published under the GNU-GPL license and can be freely downloaded from http://lpg.uoc.edu/wn-toolkit.

## 2. The WN-Toolkit

The WN-Toolkit (Oliver, 2014) is a set of programs written in Python for the creation of WordNets following the Expand Model. The toolkit also provides some free language resources. The main goal of the toolkit is to facilitate the creation of WordNets for new languages as well as to expand existing WordNets.

The toolkit is divided into the following parts:

- Programs
  - Miscellaneous tools
  - Dictionary based strategy
  - Babelnet based strategy
  - Parallel corpora based strategies
  - Evaluation tools
- Resources
  - Dictionaries
    - Apertium dictionaries
    - Wiktionary dictionaries
    - Wikipedia dictionaries
  - Parallel corpora
    - Semcor 3.0
    - Princeton WordNet Gloss Corpus
    - DGT-TM-release2013
    - EMEA-03 Corpus
    - UNCorpus
    - ECB Corpus

The resources published with the toolkit are preprocessed allowing a direct an easy use with the toolkit.

In this paper we will present the results and evaluation for the parallel corpus based strategies using some of the algoritms and resources published in the WN-Toolkit.

## 3. Creation of WordNets from parallel corpora

In some previous works we presented a methodology for the construction of WordNets based on the use of parallel bilingual corpora. These corpora have to be semantically tagged, the tags being PWN synsets, at least in the English part. As this kind of corpus is not easily available we explored two strategies for the automatic construction of these corpora:

- By machine translation of sense tagged corpora (Oliver and Climent, 2011), (Oliver and Climent, 2012a)
- By automatic sense tagging of bilingual corpora (Oliver and Climent, 2012b).

Once we have created the parallel corpus, we need a word alignment algorithm in order to create the target WordNet. Fortunately, word alignment is a well-known task and several freely available algorithms are available. In previous works we have used Berkeley Aligner (Liang et al., 2006). In this paper, we present the results using (i) a very simple word alignment algorithm based on the most frequent translation, (ii) parallel corpora between English and several target languages, and (iii) automatic sense tagging of the English part of the corpora.

### 3.1. Automatic sense-tagging of parallel corpora

In our experiments we have used Freeling and UKB (Padró et al., 2010) to semantically tag the English part of the parallel corpora.

Let's observe an example:

We have the English sentence:

```
Protocol adjusting the Agreement on
the European Economic Area
```

We can use Freeling+UKB to obtain the synsets in this sense and transform the original sentence into a sentence where the English words are replaced by their correspondig synsets:

```
06665108-n 00150287-v the 13971065-n
on the European_Economic_Area
```

As we are working with parallel corpora we have the translation into the target language (Spanish in this example):

```
Protocolo por el que se adapta el
Acuerdo sobre el Espacio Económico
Europeo
```

Using any available tagger we can tag the target sentence. Please, note that we need very simple tags (n for nouns, v for verbs, a for adjectives, r for adverbs and any other tag for the rest of the part-of-speech), as in the example:

```
protocolo|n por|x el|x que|x se|x
adaptar|v el|x acuerdo|n sobre|x
el|x espacio_económico_europeo|n
```

Now, using a word alignment algorithm we can align the synsets in the English part with their corresponding Spanish words, thus obtaining a subset of the target language WordNet.

```
06665108-n   protocolo
00150287-v   adaptar
13971065-n   acuerdo
```

As we are using statistical word alignment algorithms, we can cope with word order changes and other linguistic phenomena.

## 4. Experimental settings and evaluation

In this section we present the results and evaluation figures for the WordNets created for Spanish, French, German, Italian and Portuguese using the following corpora:

- DGT-TM (Steinberger et al., 2012)
- EMEA (European Medicines Agency) (Tiedemann, 2009)

|  | en-es |  | en-fr |  | en-de |  | en-it |  | en-pt |  |
|---|---|---|---|---|---|---|---|---|---|---|
|  | Segments | Words | Segments | Words | Segments | Words | Segments | Words | Segments | Words |
| DGT-TM | 460,071 | 7,869,668 | 444,698 | 7,850,196 | 454,010 | 7,803,279 | 456,781 | 7,858,212 | 454,792 | 7,787,947 |
| EMEA | 292,905 | 4,210,487 | 297,670 | 4,243,719 | 290,686 | 4,146,055 | 288,033 | 4,132,762 | 288,180 | 4,132,762 |
| ECB | 89,884 | 2,421,166 | 156,947 | 4,527,453 | 87,907 | 2,396,739 | 153,183 | 4,361,019 | 161,069 | 4,595,816 |

Table 1: Sizes of the corpora

- ECB (European Central Bank) (Tiedemann, 2009)

In table 1 we can see the size of the corpora used in the experiments (figures of words are given for the English part of each corpus).

The evaluation of the results is performed in an automatic way using the existing versions of WordNet for each language. We compare the variants obtained for each synset in the target languages. If the reference WordNet for the given languages has the same variant for the synset, the result is evaluated as correct. If the reference WordNet does not have any variant for the synset, this result is not evaluated. If the reference WordNet has some variants for this synset, but not the one obtained by our algorithm, the result is evaluated as incorrect. This evaluation method has a major drawback: since the existing WordNets for the target languages are not complete (some variants for a given synset may be not registered), some correct proposal can be evaluated as incorrect.

### 4.1. Setting the optimal values for the parameters of the synset-word alignment algorithm

The alignment algorithm used in the WN-Toolkit is a very simple one, based on the most frequent translation. For a given synset the algorithm look at all the target language sentences corresponding to the English sense tagged sentences having this synset. The algorithm calculates the most frequent word for the same POS and takes it as a target variant for the synset. To refine the results, the algorithm uses two parameters:

- parameter $i$: The minimum value of the frequency of 1st alignment candidate divided by frequency of 2nd alignment candidate. That is, the minimum number of times the frequency of the 1st alignment candidate should be bigger than the frequency of the 2nd alignment candidate.
- parameter $f$ : The maximum number of times the synset frequency can be higher than the variant frequency. That is, if the frequency of the given synset is higher than $f$ times the frequency of the alignment candidate, this extracted variant is rejected.

| i | f | Variants | Precision |
|---|---|---|---|
| 1 | 1 | 13,133 | 23.10 |
| 2.5 | 1 | 850 | 77.14 |
| 5 | 1 | 149 | 85.53 |
| 1 | 2.5 | 13,975 | 38.61 |
| 2.5 | 2.5 | 2,123 | 79.73 |
| 5 | 2.5 | 524 | 85.26 |
| 1 | 5 | 14,404 | 41.55 |
| **2.5** | **5** | **2,433** | **79.17** |
| 5 | 5 | 612 | 84.03 |

Table 2: Evaluation results for several combinations of parameters $i$ and $f$ (values for Spanish)

In table 2 the number of extracted variants and the precision for several combination of the parameters $i$ and $f$ for Spanish can be observed. For the rest of the experiments we will use $i = 2.5$ and $f = 5$.

As expected, the more restrictive the parameters $i$ and $f$, the higher the precision and the lower the number of extracted variants. Using $i = 2.5$ and $f = 5$ we can obtain 2,434 variants with a precision of 79.17%. We must keep in mind that the precision values are calculated in an automatic way and the real values are expected to be higher. In previous experiments we have manually revised the results of an extraction process from a fragment of the DGT-TM corpus and we have obtained a value of corrected precision of 88.94% while we obtained a precision of 79.71% from the automatic evaluation.

### 4.2. Full evaluation results

In table 3 we can observe the precision and the number of extracted variants for $i = 2.5$ and ”$f = 5$ for all five languages. Best precision values are obtained for French (85.03% for the ECB corpus) and worse precision values are obtained for German (45.64% for the EMEA corpus). Please, note that while we are obtaining better precision values for French, we are also obtaining less variants compared with Spanish, Italian and Portuguese. On the other hand, we are obtaining worse results of precision for German and also a smaller number of variants. These differences on the results can be produced by several factors:

- The completeness of the reference WordNet. As the evaluation is performed by a comparison of the obtained results with the reference WordNet, if this is incomplete and contains some but not all variants for a given synset, the automatic precision values can be lower than real values.

| | Spanish | | French | | German | | Italian | | Portuguese | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Variants | P. | Variants | P. | Variants | P. | Variants | P. | Variants | P. |
| DGT-TM | 2,434 | 79.17 | 1,771 | 83.01 | 1,585 | 51.69 | 1,959 | 79.49 | 2,076 | 78.7 |
| EMEA | 1,474 | 81.78 | 1,058 | 83.86 | 974 | 45.64 | 1,176 | 81.71 | 1,304 | 76.64 |
| ECB | 744 | 78.79 | 358 | 85.03 | 666 | 53.15 | 450 | 81.17 | 793 | 75.73 |

Table 3: Full results of the experiments for $i = 2.5$ and "$f = 5$

- The quality of the tagger used to preprocess the corpora for a given language.
- The quality of the sense tagging algorithm (Freeling + UKB in our experiments).
- The percent of monosemic variants in the corpus, that is, the number of variants assigned to a single synset. These variants don't need disambiguation so they are less prone to errors. In table 4 we can observe an estimation of the percent of ambiguous variants in each corpus.

| | **Non ambiguous** | **Ambiguous** |
|---|---|---|
| **DGT-TM** | 17.54 | 82.46 |
| **EMEA** | 26 | 74 |
| **ECB** | 18.77 | 81.23 |

Table 4: Percent of ambiguous and non ambiguous variants in each corpus

Precision values can be considered good enough, but the number of extracted variants is very low compared with the size of the corpora. It is worth to compute the number of synsets present in the English part of each corpus. In table 5 we can see its values for each target language. Please, note that the size of each corpus is different for each language, as stated in table 1, and figures are given for the English part of the corpus. In table 6 we can observe the values of precision, recall and F1 for all the experiments.

## 5. Conclusions

In this paper we have presented the results of automatic creation of WordNets for five languages using an Expand Model technique that performs automatic sense tagging of the English part of parallel corpora. As these corpora are also available in other languages it is possible to replicate these experiments for new languages. This technique, along with techniques based on dictionaries, are implemented in the freely available WN-toolkit and have been successfully used for the expansion of the Catalan and Spanish WordNets under the Know2 project[1]. The WordNets and the toolkit itself are being improved under the Skater Project[2]. The successful use of this toolkit has been also reported for the Galician WordNet (Gómez Guinovart and Simões, 2013).

The WN-Toolkit is freely available for download at http://lpg.uoc.edu/wn-toolkit.

## 6. Future work

The major drawback of this methodology is the very low recall values obtained. This is mainly due to the use of a very simple word-alignment algorithm, based on the most frequent translation. In future experiments we plan to use more sophisticated word alignment algorithms as:

- Giza++ (Och and Ney, 2003)
- Berkeley Aligner (Liang et al., 2006)
- Fast Align (Dyer et al., 2013)

The algorithms used in this paper are part of the WN-Toolkit, a freely available toolkit for the creation of WordNets using the Expand Model. As future work we plan to expand the WN-Toolkit with new features. We also plan to publish new freely available language resources along with the toolkit.

## 7. Acknowledgements

This research has been carried out thanks to the Project *Adquisición de escenarios de conocimiento a través de la lectura de textos: Lingüística y Cognición*, SKATER (TIN2012-38584-C06-01 and TIN2012-38584-C06-06) supported by the Ministry of Economy and Competitiveness of the Spanish Government.

## 8. References

G. De Melo and G. Weikum. 2008. On the utility of automatically generated wordnets. In *Proc. Global WordNet Conference*.

C. Dyer, V. Chahuneau, and N.A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the NAACL*.

T. Erjavec and D. Fišer. 2006. Building slovene WordNet. In *Proceedings of the 5th International Conference on Language Resources and Evaluation LREC*, volume 6, pages 1678–1683, Genoa, Italy.

X. Gómez Guinovart and A. Simões. 2013. Retreading dictionaries for the 21st century. In *Proceedings of the 2nd Symposium on Languages, Applications and Technologies (SLATE'13)*, page 115–126.

[1] KNOW2. Language understanding technologies for multilingual domain-oriented information access. Ministry of Science and Innovation (Spain). TIN2009-14715-C04-01 (2008-2011)

[2] SKATeR. Scenario Knowledge Acquisition by Textual Reading. Ministry of Economy and Competitivity (Spain). TIN2012-38584-C06-01. (2013-2015)

| | **Spanish** | **French** | **German** | **Italian** | **Portuguese** |
|---|---|---|---|---|---|
| **DGT-TM** | 23,067 | 23,254 | 22,884 | 22,953 | 22,788 |
| **EMEA** | 13,455 | 13,419 | 13,441 | 13,129 | 13,166 |
| **ECB** | 10,570 | 12,204 | 11,051 | 12,099 | 11,911 |

Table 5: Number of synsets on the English part of each corpus

| | Spanish | | | French | | | German | | | Italian | | | Portuguese | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P. | R | F1 | P. | R | F1 | P. | R | F1 | P. | R | F1 | P. | R | F1 |
| DGT-TM | 79.17 | 10.55 | 18.62 | 83.01 | 7.62 | 13.95 | 51.69 | 6.93 | 12.12 | 79.49 | 8.53 | 15.42 | 78.7 | 9.11 | 16.33 |
| EMEA | 81.78 | 10.96 | 19.32 | 83.86 | 7.88 | 14.41 | 45.64 | 7.25 | 12.51 | 81.71 | 8.96 | 16.15 | 76.64 | 9.9 | 17.54 |
| ECB | 78.79 | 7.04 | 12.92 | 85.03 | 2.93 | 5.67 | 53.15 | 6.03 | 10.83 | 81.17 | 3.72 | 7.11 | 75.73 | 6.66 | 12.24 |

Table 6: Precision, recall and F1 for the experiments for $i = 2.5$ and ”$f = 5$

A. Horák, K. Pala, and A. Rambousek. 2008. The global WordNet grid software design. In *Proceedings of the Fourth Global WordNet Conference, University of Szegéd*, page 194–199.

R. Izquierdo, A. Suárez, and G. Rigau. 2007. Exploring the automatic selection of basic level concepts. In *Proceedings of RANLP*, volume 7.

P. Liang, B. Taskar, and D. Klein. 2006. Alignment by agreement. In *Proceedings of the HLT-NAACL ’06*.

B. Magnini and G. Cavaglia. 2000. Integrating subject field codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation LREC 2000.*

M. Ngai, G.and Carpuat and P. Fung. 2002. Identifying concepts across languages: A first step towards a corpus-based approach to automatic ontology alignment. In *Proceedings of the 19th international conference on Computational Linguistics*, page 1–7.

I. Niles and A. Pease. 2003. Linking the lexicon and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering. Las Vegas. Nevada.*

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

A. Oliver and S. Climent. 2011. Construcción de los wordnets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente. In *Proceedings of the 27th Conference of the SEPLN, Huelva, Spain*.

A. Oliver and S. Climent. 2012a. Building wordnets by machine translation of sense tagged corpora. In *Proceedings of the Global WordNet Conference, Matsue, Japan*.

A. Oliver and S. Climent. 2012b. Parallel corpora for wordnet construction. In *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (Cicling 2012). New Delhi (India).*

A. Oliver. 2014. WN-Toolkit: Automatic generation of wordnets following the expand model. In *Proceedings of the 7th Global WordNetConference*, Tartu, Estonia.

L. Padró, S. Reese, E. Agirre, and A. Soroa. 2010. Semantic services in freeling 2.1: Wordnet and UKB. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010).*

R. Steinberger, A. Eisele, S. Klocek, S. Pilos, and P. Schlüter. 2012. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, page 454–459.

J. Tiedemann. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, page 237–248.

P. Vossen. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Springer.

J. Álvez, J. Atserias, J. Carrera, S. Climent, E. Laparra, A. Oliver, and G.F Rigau. 2008. Complete and consistent annotation of WordNet using the top concept ontology. In *Proceedings of the 6th Language Resources and Evaluation Conference LREC.*