Using Context to Improve the Spanish WordNet Translation

Alfonso Methol, Guillermo López, Juan Miguel Álvarez, Luis Chiruzzo, Dina Wonsever
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay

Abstract

We present some strategies for improving the Spanish version of WordNet, part of the MCR, selecting new lemmas for the Spanish synsets by translating the lemmas of the corresponding English synsets. We used four simple selectors that resulted in a considerable improvement of the Spanish WordNet coverage, but with relatively lower precision, then we defined two context based selectors that improved the precision of the translations.

1 Introduction

This paper presents an approach at the expansion of the lexical database WordNet in Spanish using an automatic translation processes. We implemented some previously proposed strategies for improving the coverage of the lexical database in Spanish, then we analyzed the results that these strategies produced and finally we designed new strategies in order to improve the quality of the translated lemmas.

The rest of the paper is structured as follows: section 2 presents the lexical database we aim to improve and describes related work in the area, section 3 describes the translation sources we used and how they were prepocessed, section 4 details the different strategies implemented for translation, section 5 shows the results obtained by the strategies and their evaluation, finally section 6 shows our conclusions and some future research directions.

2 Background

The Multilingual Central Repository, MCR (González-Agirre et al., 2012), is a multilingual lexical database that contains linked WordNet versions for English and five languages spoken in the Iberian peninsula: Spanish, Catalan, Basque, Galician and Portuguese. The same Princeton WordNet synsets structure is used for all languages. The central component of this lexical database is the Inter-Lingual-Index (ILI), which allows the mapping of concepts of different languages through the use of identifiers. The identifiers are composed of four values: language, version of MCR, synset offset and part of speech.

Synsets in different languages that have the same meaning share the offset, version and part of speech, varying the language. For example, "house" (eng-30-03544360-n) corresponds to "casa" (spa-30-03544360-n) and both synsets are related through the ILI code "ili-30-03544360-n".

The first attempts at building a Spanish version of WordNet are described in (Atserias et al., 1997), using bilingual English-Spanish dictionaries and a large monolingual Spanish dictionary. A different approach is proposed in (Oliver and Climent, 2011) for Spanish and Catalan, using machine translation systems to translate the semantically annotated SemCor (Miller et al., 1993) corpus and select the translations for variants based on the relative frequencies of words in the corpus with the following strategies:

- Algorithm A: Order the English synsets by frequency in the original corpus. Starting with the most frequent synset, build a subset of the automatically translated corpus with the sentences that contain a member of the synset. Choose the most frequent lemma from the translated corpus that has the same POS as the original synset. This process is repeated for each synset in order of frequency.
- Algorithm B: The same as algorithm A, but choose a lemma only if its frequency is at least twice the frequency of the next lemma. This process has considerably better precision than the previous one.

In (Pradet et al., 2014) the authors present a method for improving the French version of WordNet. They compile a collection of possible translations for the variants from several bilingual sources and design strategies for selecting the appropriate translation, these strategies are called "selectors". A selector is a heuristic strategy that takes a synset and a set of candidate lemmas in the target language, and returns the most appropriate lemma that should be associated to the synset. A similar approach was followed by (Herrera et al., 2016) for the expansion of Spanish WordNet, defining five selectors and obtaining good results for a subset of synsets from Princeton WordNet (92% accuracy for simple selectors and 74% accuracy for the distributional selector). The selectors were only applied on a subset of the synsets due to the long execution times, also some problematic synsets (such as multiword expressions) were not considered, which might explain in part the high accuracy of the simple selectors.

The authors of (Oliver, 2016) also use a dictionary based approach, combining several linguistic resources in a variety of languages for improving the WordNet translation in each of those languages.

3 Translation sources

Translation sources are key elements in the process of building WordNet in Spanish. They provide, for the English lemmas, the lemmas in Spanish that will be used by the selectors as translation candidates.

Two types of sources were used: dictionaries and statistical machine translators. The dictionaries are made up of tuples *[English word, Spanish word, POS]*. They are generated manually so they are very reliable, but with a limited volume of translations. The machine translators used are statistical systems that allow to translate words and also complete sentences taking the context into consideration, a property that will be exploited by some of the selectors.

3.1 Dictionaries

- Apertium: It is a rule based machine translation system (Forcada et al., 2011) developed with the joint financing of the Spanish government and the Generalitat de Catalunya at the University of Alicante. The software as well as the linguistic data is free and it is released under the terms of the GNU GPL license. A dictionary was created from the ".dix" file of Apertium corresponding to the translations from English into Spanish. The version used has 26,643 translations, and covers 42,996 WordNet lemmas, which accounts for 20.67% of it.

- Wiktionary[1]: It is a project of the Wikimedia foundation that aims to create a free multilingual dictionary, based on the massive collaboration of volunteers through the wiki technology for the elaboration of its content. It is currently available in more than 170 languages and has more than 15 million entries. Because of the considerable volume of its data and its well defined structure, it is particularly useful for our processing. The version used contains 40,166 possible translations into Spanish for 47,982 lemmas, covering the 23.06% of WordNet lemmas.

- Eurovoc: Published by the Publications Office of the European Union, it is a multidisciplinary thesaurus focused on the terminology used in the different areas of activity of the European Union (Maciá, 1995), and it covers the 23 official languages of the region. Due to the scope of the thesaurus, this translation source has few general terms, which considerably restricts its broad applicability in this project, but it contains specific data that can be very useful for translation of diplomatic documents. Out of 6945 lemmas contained in EuroVoc, 2032 appear in WordNet, which represents 1.38% of the lemmas.

3.2 Machine translators

- Google Translate: It is a statistical machine translation system capable of translating texts, speech, images, websites among more than 100 languages. Provides a free access web tool[2] as well as a service included in Google Cloud Platform.

- Microsoft Translator: It is a statistical machine translation web service[3] provided by Microsoft, which can be used through an API that provides translation of text, voice and text to speech.

[1] https://www.wiktionary.org/
[2] https://translate.google.com
[3] https://www.bing.com/translator

- Yandex: They offer a statistical machine translator[4] for many pairs of languages, including Spanish and English. The translator uses a combination of dictionaries of words and expressions with probabilistic information and also linguistic rules. It can be queried using a web API.

3.3 Cleaning sources

To solve some of the limitations and reduce the costs of access to the selected translation sources, a single format was defined and stored in the same database. For each translation source a table was created with the following columns:

- English word
- Spanish translation
- Part of Speech

The dictionaries did not need any extra processing and only these fields are stored. The tables corresponding to machine translation systems have another field:

- Snapshot date

We decided to take a snapshot of the translation of all WordNet lemmas by each of the machine translation systems at a specific time. This was motivated by the different limitations in the use of online APIs and their response times. Using the snapshot approach, we can use the machine translation systems as if they were just another dictionary. Although we might not have completely up to date information in each run, we consider the translations we use should not vary much in time and the execution time is greatly improved respect to the online execution of the APIs. The snapshot date is stored, so we can later on take a new snapshot, compare the differences and adjust the methods accordingly.

None of the three machine translation used systems return the POS along with the translation, so we used FreeLing (Padró and Stanilovsky, 2012) for POS tagging. We detected many translation errors, where a different POS was returned because of the lack of context, so we did some improvements to the translation heuristic, such as adding the prefix "to" to verbs in English in order to force the translator to consider them as verbs. We also used FreeLing dependency parser to assign the POS in multiword expressions.

[4]https://translate.yandex.com/

3.4 Coverage

We analyzed the coverage of MCR over a corpus of 850 million words of news text in Spanish (Bonanata and Stecanella, 2013)

The coverage before our process is shown in the following table:

POS	Lemmas in corpus	Lemmas in MCR
Adj	42,604	5,592 (13.12%)
Adv	10,676	523 (4.90%)
Noun	104,811	11,523 (10.99%)
Verb	37,522	8,821 (23.51%)
All	195,613	26,459 (13.53%)

Table 1: MCR Coverage over news corpus

We can observer a low coverage of the corpus MCR. This is due in part to the number of lemmas available in Spanish.

4 Translation process

We first implemented some of the already defined selectors and applied them to the whole collection of synsets. As these selectors resulted in poor precision, we created new selectors that exploit contextual information in order to improve the precision of the translation.

4.1 Simple selectors

Following the strategies of (Pradet et al., 2014) and (Herrera et al., 2016), we reimplemented some of the selectors that have been previously executed for only a fraction of the English synsets and applied them to all the synsets.

- Monosemy

  This strategy works with English lemmas which appear in a single synset regardless of their part of speech. The assumption behind this is that this uniqueness condition implies the meaning of the lemma is unambiguous. The translations of all the sources for each compliant lemma are assigned to the corresponding synset.

  For example: Consider the English lemma "advisable" which only appears in the English synset "eng-30-00067038-a". The selector then assigns all of the lemmas translations to the corresponding Spanish synset "spa-30-00067038-a", in this case: "aconsejable", "recomendable" and "conveniente".

- Single Translation

  This selector takes into account only those lemmas that have a unique translation into Spanish. This translation is added in all the synsets in Spanish corresponding to the synsets in English that contain this lemma.

  For example: Consider the lemma "flavor", which occurs in the synsets "eng-30-14526182-n", "eng-30-05715864-n" and "eng-30-05844282-n". There is a unique translation for this lemma that is "sabor". This translation is selected for the corresponding synsets in Spanish. However, for the lemma "play" occuring in the synsets "eng-30-01072949-v", "eng-30-02370650-v" and "eng-30-01725051-v" (among other 35 synsets in total), our translation sources give four possible lemmas: "jugar", "reproducir", "tocar" and "interpretar". Because of this the selector discards these translations.

- Factorization

  Unlike previous selectors, this one runs at synset level. For each lemma of a synset it obtains all its translations and generates a translation set. Once the sets of translations of each lemma of the synset are obtained, the selector keeps those translations common to all sets.

  For example: The synset "eng-30-00011516-r" contains the lemmas "poorly", "badly" and "ill", where their translations are:

  - poorly: mal, pobremente.
  - badly: mal, malamente.
  - ill: mal, enfermo.

  In this example, the only translation common to the three lemmas that is selected for the corresponding synset in Spanish is "mal", the remaining translations ("pobremente", "malamente" and "enfermo") are discarded.

- Derived Adverb

  This selector is executed for the adverb synsets and is the only one that uses a semantic relation of those defined in MCR, the `is_derived_from` relation. From an adverb synset, look up with which adjective synsets it is related. For each adjective obtain the translations, and use morphological derivation rules to convert them into possible adverbs.

  The morphological rules applied are as follows:

  - If the adjective ends with the letter "o", it is replaced by the sequence "amente", for example, for "rápido" the result is "rápidamente".
  - If the adjective ends with the letter "r" or "n", the sequence "amente" is attached, for example, for "alentador" the result is "alentadoramente".
  - If the adjective does not fit into the above categories, only the sequence "mente" is attached, for example, for "vil" the result is "vilmente".

  As these rules are heuristics, not all results obtained after the process are valid adverbs. For example, when applying the rules to the adjective "rojo" we get the adverb "rojamente", which does not exist as a valid word in the Spanish language. To solve this problem the adverbs generated were validated against a list of adverbs that occur in a corpus (Bonanata and Stecanella, 2013).

  For example: The synset "eng-30-00010466-r" has the lemmas "fully", "full" and "to_the_full" and is related to the adjective synset "eng-30-00522885-a" containing the lemmas "total" and "full". The translations for the adjectives are: "pleno", "repleto", "lleno", "completo" and "total". Applying the morphological rules we get: "plenamente", "repletamente", "llenamente", "completamente" and "totalmente". These are checked using the corpus and added to the corresponding synset.

4.2 Selectors based on contextual information

After analyzing the performance of the original selectors, which will be shown in section 5, we realized that many of the errors happened because these selectors do not take in consideration the context the words could be used in. We defined two new selectors that try to use the context provided by the examples of the synsets to improve the quality of the candidate translations.

We translate all the examples contained in WordNet using Google Translate and generated a parallel corpus associated with synsets. 27.71% of the MCR English synsets have examples, adding up to 41,305 candidates, which gives us an upper bound to the number of synsets we might translate with these strategies.

- Filtering selector

  This selector works by analyzing which of the generated translations of the present lemma in an example in English, are in the translation of the example to Spanish. The check of occurrences of both the lemmas in the example in English, and their translations in the translated example is done in two stages. It is called filtering because it leverages the information from the dictionaries, trying to filter which of the candidates are present in the example and its translation. The procedure is as follows: First check if lemma and translation occur in the example and the translated example. If this does not happen, apply FreeLing to the text to obtain the lemma and POS of each word. Then iterate them by re-checking the occurrences. This second stage tries to detect words or translations that occur in the examples in a conjugated form, as is the case for many verbs. Otherwise we would be losing many valid translations. This is done as a second step because using FreeLing to get the lemmas is an expensive process.

  For example: When we apply the selector to the example "his last words" associated with the synset "eng-30-00004296-a", it detects that the only lemma of the synset ("last") occurs directly in the example. Once this is detected, the translations are obtained. In this case, the lemma "último" is the only translation candidate.

  The translated example is "sus últimas palabras". The candidate lemma is not present so FreeLing is used to obtain the lemmas and POS of the translated example, getting the following information: "[(su, D), (último, A), (palabra, N)]". Since we are dealing with an adjective synset, we compare to the adjectives returned by FreeLing and we get a match with the lemma "último", which is selected as the translation to the corresponding synset ("spa-30-00004296-a") in Spanish.

- Structure based selector

  This selector focuses on the use of translated examples as a parallel corpus where it is possible to align the different parts of the sentences in both languages. We use the path from the root to the word in a dependency parse tree, and try to match the corresponding path in the tree of the translated example. In this way, we use the internal structure of the sentences and the relative positions of the words, such as their location within a subject or a predicate.

  We begin by obtaining the dependency parses of the example and its translation using FreeLing. This construction allows the analysis of the different components of sentences and their relationships. Using the dependency structures, we identify the lemma to be translated from the sentence and its syntactic (subject or predicate) location, and take note of the labels belonging to the shortest path from the root of the tree. The same path is followed in the translated example, taking in consideration the differences in label names for both languages, and we return a lemma if it is in the appropriate position in the tree and has the expected POS.

  Example: We want to translate the lemma "bond" for the English synset "eng-30-13792183-n" using the example: "their friendship constitutes a powerful bond between them". The dependency tree for this sentence is shown in figure 1.

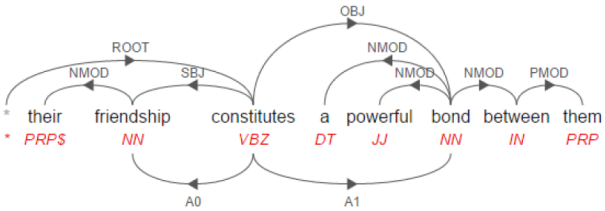

Figure 1: Dependency parsing of "their friendship constitutes a powerful bond between them"

The corresponding translation for this example in Spanish is "su amistad constituye un poderoso vínculo entre ellos", whose dependency tree is shown in figure 2. In this tree we find the lemma "vínculo" in the correspond-

Selector	Generated	MCR	Intersection	Overlap	New
Monosemy	183386	146501	47632	32.51%	74.03%
Single Transl.	81058	146501	38505	26.28%	52.50%
Factorization	111919	146501	34400	23.48%	69.26%
Derived Adv.	5161	3583	1907	53.22%	63.05%
All Simple	256852	146501	72674	50.39%	71.71%
Filtering	22401	146501	12680	8.66%	43.40%
Structure	12168	146501	6857	4.68%	43.65%
All Context	25223	146501	13291	9.07%	47.31%
All	264105	146501	75416	51.48%	71.44%

Table 2: Number of generated lemmas, overlap with MCR lemmas and generated lemmas that are new by selector.

ing position, so this lemma gets selected for the Spanish synset "spa-30-13792183-n".

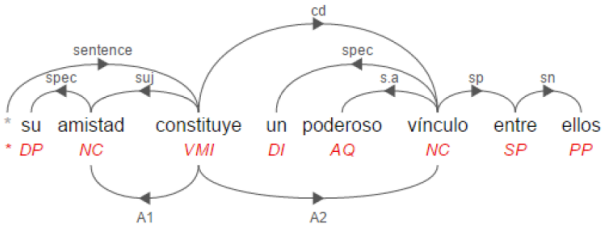

Figure 2: Dependency parsing of "su amistad constituye un poderoso vínculo entre ellos"

In this case the lemma and its translation was easy to locate both in the original sentence and the translation: it is a single name located in the direct object of the sentence in both cases, so it quickly follows that the translation of "bond" is "vínculo".

However, this is not always the case. Among the most common errors in the execution of this selector are situations in which the root of the example in English changes considerably when translated. This is because in many cases the English and Spanish parsers use different criteria. That is the case of the sentence: "Can you read Greek?" (figure 3), whose translation is "¿Puede usted leer griego?" (figure 4). The lemma that we want to translate is "read", and is located in the sentence predicate in the original version, but becomes the root of the tree in the translated version. Even though both sentences have similar structure in English and Spanish, the parsing process treats them differently.

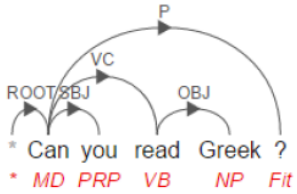

Figure 3: Dependency parsing of "Can you read greek?"

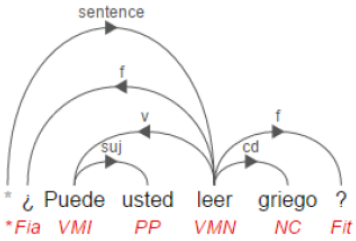

Figure 4: Dependency parsing of "¿Puede usted leer griego?"

5 Evaluation

Evaluation was one of the hardest tasks due to the complexity of the evaluation of some semantic notions, as well as the volume of data involved. Because of this, we decided to use two methods of evaluation: evaluation by overlap and evaluation by sampling.

5.1 Overlap evaluation

The overlap evaluation consists in comparing the translations generated with those already found in Spanish MCR. This could be seen as a kind of recall, giving an idea of how good our heuristics are at capturing the information we already knew. The overlap by phase and selector is shown in table 2.

Notice that the lemmas translated using the con-

text based selectors are fewer than the ones generated with the simple translators. This was an expected result, because these selectors use the synset examples. Not all synsets have examples, and even the ones that contain examples do not necessarily have them for every lemma. This coverage could be greatly improved using more data.

5.2 Sample evaluation

Due to the large volume of translations generated we could not evaluate the correctness of each one of the terms. For this reason we carried out a sampling evaluation consisting of taking a random sample of 3,000 synsets and evaluating them manually. For the initial phase, 750 synsets by POS were selected, in the contextual information phase, 1500 were selected per selector (375 by POS). We built a special tool that aids in the process of evaluating the correctness of the sampled translations. The result of this method of evaluation is an estimation of the precision for each selector and each phase. The precision is shown in table 3.

Selector	Sampled	Correct
Monosemy	3,603	2,367 (65.70%)
Single Transl.	2,471	1,927 (73.65%)
Factorization	3,193	2,057 (64.42%)
Derived Adv.	1,164	852 (73.20%)
All simple	10,431	7,203 (69.05%)
Filtering	1,695	1,424 (83.96%)
Structure	1,674	1,361 (81.30%)
All contextual	3,369	2,785 (82.67%)

Table 3: Precision by selector, showing the number of tested lemmas and the number of correct ones for each selector.

Table 4 shows the precision achieved for each POS, separated in the two phases: simple selectors and selectors with contextual information.

POS	Simple Sel.	Contextual Sel.
Adj	74,89%	87.34%
Adv	73,65%	88.42%
Noun	57,51%	80.24%
Verb	52,47%	74.12%

Table 4: Precision by POS, showing the overall precision for simple selectors and selectors with contextual information.

As we can see, the precision for the initial selectors was lower than the one reported in (Herrera et al., 2016). There are several causes for this, first of all we transformed the whole collection of synsets and took a larger evaluation sample, even considering multiword expressions and their translations. In one of the cases the precision only for simple lemmas got 81%, while for multiword expressions it dropped to 66%. Also, on occasions the machine translation systems returned results that contained an unnecessary determinant (e.g. translating "immigration" as *"la inmigración"*). However, at many times the error was caused by selecting a translation that would be unfit for the context, for example it translated "ring" from synset "eng-30-07391863-n" ("the sound of a bell ringing") as *"anillo"*, which is an appropriate translation for the other sense in synset "eng-30-04092609-n" ("jewelry consisting of a circlet of precious metal..."). The low precision of these methods motivated the contextual information approach, which obtained fewer translations but with better precision for all parts of speech.

5.3 Impact over MCR

The contribution to Spanish MCR is shown in Table 5.

POS	Spanish MCR Lemmas	New Lemmas	Increase
Adjectives	6,967	19,140	274.72%
Adverb	1,051	8,689	826.74%
Noun	39,142	183,880	469.78%
Verb	10,829	21,355	197.20%

Table 5: Contribution to Spanish MCR

Reanalyzing the coverage of MCR over the news text based corpus (Bonanata and Stecanella, 2013) including the newly generated lemmas we obtained the new coverage shown in table 6.

6 Conclusions

We implemented four simple selectors and two contextual based selectors for the translation of English WordNet synsets to Spanish, in order to expand the Spanish version of WordNet present in MCR. Using the simple selectors, we obtained 182,051 nouns, 19,683 verbs, 17,384 adjectives and 8,436 adverbs with 69.05% precision. The precision of these selectors was lower than the one reported in previous works, probably because in our case we evaluated the whole collection

POS	Lemmas in corpus	Lemmas in MCR	MCR + new lemmas
Adj	42,604	5,592 (13.12%)	18,063 (42,40%)
Adv	10,676	523 (4.90%)	7,105 (66,55%)
Noun	104,811	11,523 (10.99%)	35,535 (33,90%)
Verb	37,522	8,821 (23.51%)	22,427 (59,77%)
All	195,613	26,459 (13.53%)	83,130 (42,50%)

Table 6: Coverage of MCR with new lemmas.

of synsets, even processing multiword lemmas. In order to improve this precision, we designed and implemented two new selectors that use the contextual information, whose execution obtained 5,339 nouns, 4,441 verbs, 6,444 adjectives and 1,747 adverbs with 82.67% precision. The context based selectors yield much fewer results because they depend on the existence of examples in the corresponding WordNet synsets.

During the course of the project we detected several directions that could be explored in the future. First of all, we would need to analyze the cases in which the simple selectors did not give any results. This could mean expanding the set of translation sources in order to cover all the vocabulary of the original WordNet, as this coverage is the upper bound to what we might be able to translate.

For the contextual information selectors, we could obtain a larger parallel corpus of examples. One possibility is using the SemCor corpus that has been used in other projects, another possibility would be performing word sense disambiguation over a large parallel corpus, taking into account that this process would probably not select the correct synset every time. The structure selector is particularly interesting to analyze and extend, because this selector applies syntactic notions and heuristic rules that could be expanded and improved in order to add coverage and accuracy.

It would also be interesting to design new selectors based on the notions of distributed semantics, such as the use of word embeddings. The relations contained in WordNet could be used to guide the selection of new lemmas given the word embeddings property that words close in the vector space tend to have similar or related meanings.

References

Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodriguez. 1997. Combining multiple methods for the automatic construction of multilingual wordnets. In *Recent Advances in Natural Language Processing II. Selected papers from RANLP*, volume 97, pages 327–338.

Jairo Bonanata and Rodrigo Stecanella. 2013. Extracción de opiniones de prensa. Proyecto de grado, Ingeniería en Computación, Facultad de ingeniería, Universidad de la República, Uruguay.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Aitor González-Agirre, Egoitz Laparra, German Rigau, and Basque Country Donostia. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *GWC 2012 6th International Global Wordnet Conference*, page 118.

Matías Herrera, Javier González, Luis Chiruzzo, and Dina Wonsever. 2016. Some strategies for the improvement of a spanish wordnet. In *Proceedings of the Global WordNet Conference*.

Mateo Maciá. 1995. El tesauro eurovoc. *Revista General de Información y Documentación*, 5(2):265–284.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

Antoni Oliver and Salvador Climent. 2011. Construcción de los wordnets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente. Sociedad Española para el Procesamiento del Lenguaje Natural.

Antoni Oliver. 2016. Extending the wn-toolkit: dealing with polysemous words in the dictionary-based strategy. In *Proceedings of the Global WordNet Conference*.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Quentin Pradet, Gaël de Chalendar, and Jeanne Baguenier Desormeaux. 2014. Wonef, an improved, expanded and evaluated automatic french translation of wordnet. *Volume editors*, page 32.