

# Enriching a Portuguese WordNet using Synonyms from a Monolingual Dictionary

Alberto Simões<sup>1,3</sup>, Xavier Gómez Guinovart<sup>2</sup>, José João Almeida<sup>3</sup>

<sup>1</sup>Centro de Estudos Humanísticos, Universidade do Minho, Portugal

<sup>2</sup>Universidade de Vigo, Spain

<sup>3</sup>Centro Algoritmi, Universidade do Minho, Portugal

ambs@ilch.uminho.pt, xgg@uvigo.es, jj@di.uminho.pt

## Abstract

In this article we present an exploratory approach to enrich a WordNet-like lexical ontology with the synonyms present in a standard monolingual Portuguese dictionary. The dictionary was converted from PDF into XML and senses were automatically identified and annotated. This allowed us to extract them, independently of definitions, and to create sets of synonyms (synsets). These synsets were then aligned with WordNet synsets, both in the same language (Portuguese) and projecting the Portuguese terms into English, Spanish and Galician. This process allowed both the addition of new term variants to existing synsets, as to create new synsets for Portuguese.

**Keywords:** WordNet, lexical ontology, synonym dictionary, translation, lexical acquisition

## 1. Introduction

There are various initiatives aiming at the construction of a Portuguese lexical ontology, with different goals and some substantial differences in their building process. Their authors have been working together in order to reuse the other authors' results to enrich their own resources (Gonçalo Oliveira et al., 2015). Nevertheless, each one of these initiatives continue working independently. This article presents an experiment being performed by one of these projects: PULO.

PULO<sup>1</sup> (Simões and Gómez Guinovart, 2014), that stands for *Portuguese Unified Lexical Ontology*, aims at integrating linguistic resources of diverse types into a WordNet-like structure aligned with Princeton WordNet (Miller, 1995). Previous work has been done in order to expand PULO (Simões and Almeida, 2015), but it lacks both quality and coverage.

We will discuss two different methodologies to enlarge and, when possible, refine PULO, using synonyms information obtained from a standard monolingual dictionary (Casteleiro, 2006). This dictionary include synonyms aligned with each word sense, which means that it is possible to extract, from the same dictionary entry, different synonyms for different senses. Thus, when creating sets of synonyms for each sense we obtain something very similar to the synsets of WordNet.

These synsets were then aligned with the synsets already present in PULO, allowing the addition of new variants of the sense<sup>2</sup>. Moreover, these synsets were also translated into Galician, English and Spanish, and their translations were aligned with the synsets of the wordnets of these languages gathered in the Multilingual Central Repository (MCR) (Gonzalez-Agirre and Rigau, 2013). The fact that PULO is aligned with these other wordnets through the InterLingual Index (ILI) allows the creation of new synsets and the addition of new variants to existing synsets.

A small evaluation on the results presents about 80% of precision for variants in new synsets, and about 70% of precision for new variants for existing synsets.

This article is structured as follows: Section 2. discusses similar approaches, both for other languages and for PULO itself. Later, on Section 3. we describe the different resources used for our experiment. The experiment is detailed in Section 4. Section 5. does a brief evaluation on the obtained results. We conclude in Section 6. analyzing our experiment and presenting future research directions.

## 2. Similar initiatives and approaches

Under the Galnet project aimed at the construction of a Galician wordnet<sup>3</sup> (Gómez Guinovart and Solla Portela, 2015), different synonyms extraction experiments were designed from the revision and extension of a classic synonyms dictionary for the Galician language (Gómez Guinovart and Simões, 2013), which had led to the publication of the new *Dicionario de Sinónimos do Galego* (Gómez Clemente et al., 2015). The methodology used for the extraction is based on the matching of lexical forms among the variants of Galnet synsets and the variants of dictionary synsets. In this way, and with many nuances and human validation, the variants of a dictionary synset can become variants of a Galnet synset if there is a formal matching between any of the variants included in these two synsets.

In a first experiment (Gómez Guinovart, 2014), the matching was limited to synsets sharing a hapax, i.e. a variant with a unique occurrence both in the thesaurus and in Galnet. This allowed to limit incorrect extractions due to polysemy, offering an accuracy of 65%. In the second experiment (Solla Portela and Gómez Guinovart, 2014), the focus was centered on the synsets of the dictionary and Galnet sharing three or more variants, which provided 6,335 new variants with an accuracy of 35%. In order to improve the accuracy of the results, the experiment was repeated by removing the dictionary senses with five or more synonyms. The result was 856 candidate variants with an accuracy of

<sup>1</sup><http://wordnet.pt/>

<sup>2</sup>We use the term *variant* to refer to each one of the synonyms in a synset.

<sup>3</sup><http://sli.uvigo.gal/galnet/>

60%. In a third approach to the question (Gómez Guinovart and Solla Portela, 2014), the authors analyzed the results from these two experiments in order to discover and adjust further lexicographic parameters useful to improve the methodology applied in this task.

A previous approach to the enlargement of a particular wordnet based on the acquisition of synonyms from a dictionary was adopted in the development of the Turkish wordnet as part of the Balkanet project (Bilgin et al., 2004). In that work, the synonyms (along with the antonyms and the hyponyms) were automatically extracted by using pattern recognition from the lexical entries in a machine-readable monolingual dictionary of Turkish.

Another related approach was used to construct the Danish wordnet (DanNet) from a Danish monolingual dictionary (containing synonyms). In this case, the dictionary was used not as a source to expand an existing reduced wordnet, but as a strictly monolingual starting point (Pedersen et al., 2009).

A similar work of enriching a Portuguese lexical ontology like wordnet from lexical resources has been made in the Onto.PT project<sup>4</sup> (Gonçalo Oliveira and Gomes, 2014), where synsets were obtained from thesauri and other textual sources (corpus, encyclopedias and monolingual dictionaries) using several indexes of similarity for lexical extraction (Gonçalo Oliveira, 2013).

### 3. Resources

In this section we briefly describe the resources used in our experiments:

- DLPC, the *Dicionário da Língua Portuguesa Contemporânea*, used as source for the extraction of new synsets;
- The PT, GL, ES and EN wordnets in MCR used for the synsets enlargement and language triangulation;
- The translation dictionaries used for the synsets triangulation, both statistical translation dictionaries and dictionaries from the Apertium machine translation software (Forcada, 2009).

#### 3.1. Synsets from DLPC

The experiments reported in this article were motivated by a distinct work: the process of reverse-engineering the *Dicionário da Língua Portuguesa Contemporânea* (DLPC) from PDF to TEI-XML (Vanhoutte, 2004), for later revision and enlargement by human linguists.

During the reverse-engineering process, it was noted that the synonyms suggested in the dictionary were very clearly marked and, more interestingly, they were associated perfectly with the different senses of each term. An example is presented in figure 1, where the word *florir* entry is shown. This entry has two different senses (marked by the sense number, as usual), and three of these senses end with a list of synonyms, after the congruency symbol ( $\cong$ ), in small capitals.

Having these entries properly encoded in XML-TEI made possible the extraction of some sort of synsets. These

**florir** [flurír]. *v.* (Do lat. tard. *florĭre*, por *florĕre*). **1.** Dar flor; estar em flor.  $\cong$  FLORESCER. *Esta planta costuma florir no Verão. No jardim, floriam hortensias e camélias.* **2.** Enfeitar com flores. **3.** Ter origem ou aparecer.  $\cong$  BROSTAR, DESPONTAR, FLORESCER, NASCER. *Palavras que floriram dos seus lábios.* **4.** Dar ou ter vitalidade, alegria, frescura... *Florir a vida.* **5.** Manifestar-se ou surgir com encanto, esplendor.  $\cong$  BRILHAR, RAIAR. *Um sorriso floriu nos seus lábios.* **6.** *funç. subst.* O florir das amendoeiras.

Figure 1: Dictionary entry for *florir*.

synsets are comprised of the head word of the entry and the listed synonyms. For the presented example, the extracted synsets would be three:

$$S_1 = \{florir, florescer\}$$

$$S_2 = \{florir, brostar, despontar, florescer, nascer\}$$

$$S_3 = \{florir, brilhar, raiar\}$$

DLPC contains about 69,350 entries, that resulted into 43,036 different synsets, ranging from small synsets with only 2 synonyms, up to lists of 9 synonyms with an average of 2.79 synonyms per sense. Figure 2 present the distribution of the number of synonyms per number of entries.

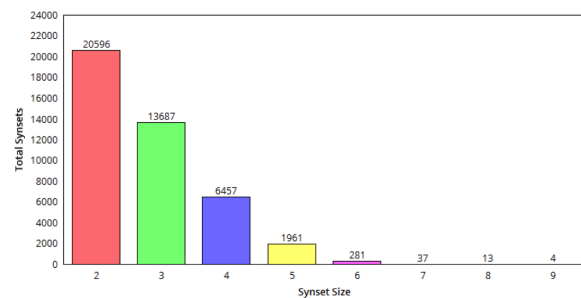


Figure 2: Distribution of DLPC synsets accordingly with their cardinality.

#### 3.2. Wordnets

As already stated, our main goal is to enrich PULO (Simões et al., 2015), both with new variants and new synsets. Although we are aware that the quality of PULO is not the best, as it was only built using automatic methods and no human validation was performed up to the moment, its current contents were used for this experiment.

We also used the Princeton WordNet, for the English language, the Galician wordnet (Galnet) and the Spanish wordnet. All these were retrieved from the MCR.

Table 1 shows both the original size of PULO and the sizes of the other three used wordnets, sorted by size. Note that each one of these wordnets grow everyday, as these are ongoing active projects.

#### 3.3. Translation dictionaries

It is not easy to find quality bilingual translation dictionaries available for free in open formats. When found, they lack in coverage, both on the source and target languages.

<sup>4</sup><http://ontopt.dei.uc.pt/>

Synsets					
	Nouns	Adject.	Verbs	Adv.	Total
EN	82,889	18,156	13,769	3,621	118,435
ES	26,594	5,180	6,251	677	38,702
GL	18,850	5,092	1,541	349	25,832
PT	9,999	3,556	3,776	523	17,854

Variants					
	Nouns	Adject.	Verbs	Adv.	Total
EN	147,358	30,004	25,051	5,580	207,993
ES	39,142	6,967	10,829	1,051	57,989
GL	25,205	8,050	4,145	420	37,820
PT	14,088	4,836	6,189	598	25,711

Table 1: Total number of synsets and variants from the used wordnets.

In order to reduce this problem, our work used both translation dictionaries from the Apertium machine translation project and probabilistic translation dictionaries (Simões et al., 2013) from the Per-Fide project (Almeida et al., 2014).

### 3.3.1. Machine translation dictionaries

The better translation dictionaries we found between Portuguese and Spanish, English and Galician were from the Apertium machine translation system. The main problem about machine translation dictionaries is that they lack coverage. First, they include most used words only (as they are usually bootstrapped from specific corpora). Second, the amount of possible translations for each word is small, in order to diminish the amount of possible translations, and making the system work faster.

To overcome this problem we included data from some other smaller dictionaries that were collected from the Internet. We also generated an particular Portuguese–Galician dictionary based on a study on the possibility of rewriting Portuguese words into Galician words with a relatively high success index (Simões and Gómez Guinovart, 2013).

The data from all these sources were merged together in a single translation dictionary for each language pair. Table 2 presents the number of entries in the dictionary, and the average number of translations per entry. Note that these dictionaries include some multiword expressions.

Dictionary	Entries	Avg. nr. Transl.
PT–EN	196,223	1.38
PT–ES	70,455	6.08
PT–GL	158,013	1.02

Table 2: Number of entries for each dictionary, and average number of translations per entry.

### 3.3.2. Probabilistic Translation Dictionaries

Parallel corpora alignment can be used to create an association between each occurrence of a word with its respective translation, or to obtain translation dictionaries, through the computation of how many times a specific form is translated other in the target language. There are different tools

available to perform this task like Giza++ (Och and Ney, 2003) or NATools (Simões and Almeida, 2003).

Although NATools is not able to do a real word-alignment task (explicitly aligning each occurrence of a word with its translation), it is able to compute Probabilistic Translation Dictionaries with acceptable quality and without the limitations on corpora size imposed by Giza++ (both limited by the algorithm and the amount of memory needed to compute the dictionaries for huge corpora).

A PTD is, basically, a mapping from words in the source language to a group of possible translations in the target language. Together with this translation association, they also comprise a probabilistic measure of how likely is the source word translated by each one of its possible translations.

The used PTDs were extracted from parallel corpora: from the Per-Fide corpora (Almeida et al., 2014) for the PT–EN and PT–ES language pairs, and from the CLUVI corpus<sup>5</sup> (Gómez Guinovart, 2015) for the PT–GL pair. The sizes of these dictionaries are presented in table 3, including the number of entries in the source language, the average number of translations per entry, and the average of the best translation probabilities.

Dictionary	Entries	Avg. nr. Transl.	Avg. Prob. for Best Transl.
PT–EN	548,220	7.72	52%
PT–ES	541,240	7.65	52%
PT–GL	3,204	2.92	68%

Table 3: Sizes of the used Probabilistic Translation Dictionaries.

## 4. Experiment description

Our experiment can be described, in a simplistic way, as the alignment of synsets extracted from DLPC with synsets from WordNet, using two distinct approaches:

- direct alignment in the Portuguese language with PULO, allowing the detection of new variants to be added to already existing synsets (in the Portuguese language);
- triangulation through the English, Spanish and Galician language (using the InterLanguage Index — ILI), aligning translations of the synsets extracted from DLPC with synsets in these languages wordnets. This approach will both allow the addition of new variants to existing synsets and the creation of new synsets to the Portuguese language.

Both these approaches will be explained in the next two subsections.

### 4.1. Intersecting DLPC with PULO

As explained before, DLPC synsets are sets of words extracted from the synonyms of the dictionary. In the other hand, PULO synsets are sets of words (variants) already existing.

<sup>5</sup><http://sli.uvigo.gal/CLUVI/>

Consider  $S^{DLPC}$  all the synsets obtained from DLPC, and  $S^{PULO}$  all PULO existing synsets. We look for a pair of synsets,  $S_i^{DLPC}$  and  $S_j^{PULO}$  which intersection is not empty:

$$S_i^{DLPC} \cap S_j^{PULO} \neq \emptyset.$$

In order to guarantee a better alignment, instead of looking for non-empty intersections, the algorithm computed the cardinality of the set intersection and reported all intersections with cardinality higher than  $n$ , being  $n$  a threshold defined by us:

$$|S_i^{DLPC} \cap S_j^{PULO}| \geq n.$$

We performed two different runs, one with  $n = 2$  and another with  $n = 3$ . As expected, the first resulted in a larger number of expanded synsets (8,926 synsets) in a total of 15,773 new variants. From the second run resulted only 573 expanded synsets, in a total of 846 new variants.

In section 5. we perform a brief evaluation on the results, and present some examples of good and bad alignments.

## 4.2. Intersecting DLPC by triangulation

The approach proposed in the previous section allows only the expansion of existing synsets. Also, it only works for synsets that already include two or three variants (so that the intersection meets our selection criteria).

As PULO is aligned with the other languages' wordnets available in the MCR through their InterLingual Index (ILI), it allows the exploitation of these wordnets for our experiments.

Thus, our idea is based on the assumption that translating DLPC entries will allow the application of the approach described in the previous section, computing intersections between the translated DLPC sets and other languages synsets (in our experiment, Galician, Spanish and English).

Figure 3 tries to represent this process. The dashed arrow is the desired alignment. It is obtained by the DLPC words' translation to a language  $\mathcal{L}_i$ , and its alignment with that same language wordnet. Then, by means of the ILI, we are able to know which PULO synset should be created or expanded.

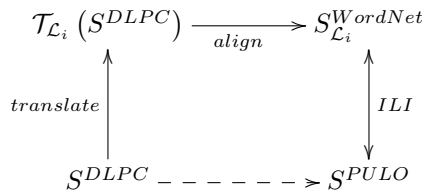


Figure 3: Schematic of the triangulation approach.

The translation was performed by the use of the dictionaries presented in section 3.3. As it can be seen by the dictionaries sizes and translation codomains, the translation will, surely, over-generate. For example, if a DLPC synset has two terms, and each one of these terms has 8 possible different translations (and the translations do not intersect), the translated synset would include 16 different terms.

Nevertheless we expect (and our algorithm depends on that) the intersection not being empty. Looking to a real example, consider the following DLPC synset:

$$S^{DLPC} = \{\text{evidente, óbvio}\}$$

Using the translation dictionaries to translate this synset to English we obtain a list with 14 entries, shown bellow (in fact, we removed some low scored entries).

As a human, we know the first three are correct. Nevertheless, we need to find a way to allow the algorithm to cut the list in the correct place.

Nevertheless, as the remaining terms are so diverse, with no semantic proximity, even if we consider them during our intersection with the English WordNet, it is very unlikely there will exist a synset containing more than one bad translation (or a bad translation and a good one, for example). This means the cut line is important to exist, in order to diminish the algorithm complexity, but mostly irrelevant for the results.

<b>461.76</b>	<b>obvious</b>
<b>400.00</b>	<b>clear</b>
<b>261.76</b>	<b>evident</b>
200.00	response
200.00	particularly
200.00	obtain
200.00	direction
50.00	sure
50.00	assumptive
11.76	striking
[...]	

The presented score is computed taking into account different aspects of the translation process:

- the probability present in the probabilistic translation dictionaries;
- the number of possible translations for each word;
- the number of terms from the DLPC synset that were translated into the same target word;

These values are accumulated, and multiplied by a hundred, for readability.

After computing the scores for all possible translations of a specific DLPC synset, the target language wordnet is searched, looking up which synsets gets a higher matching score. As an example, WordNet synset containing “obvious” would get a score of 461, while synsets containing “clear” would get only a score of 400 (as they do not share any other word from the list).

These alignments, together with the matching scores, are then consolidated for the three languages. In order to raise results' quality, the direct alignment from the previous section are also added to these alignments.

This experiment resulted in:

- 424 alignments that did not add any new information to PULO existing synsets. This measure was interesting, and gave us some confidence on the overall results;

- 28,427 synset alignments that suggest new variants to already existing synsets;
- 10,012 synset alignments that suggest new synsets, not yet available in PULO (but that can be created given the ILI identifier referred earlier).

In the next section we perform a brief evaluation for both new variant proposals, and new synsets creation.

## 5. Evaluation

In this section we describe a brief evaluation of the two experiments shown in Section 4.: the first one using only the Portuguese language (direct intersection between DLPC synsets and PULO synsets), and the second one regarding the DLPC triangulation through English, Galician and Spanish, and consequent intersection with those languages wordnets.

We are aware of the real *naive* evaluation procedure, both using only one evaluator and only one sample for each evaluation task. As the results looked good enough, we decided to incorporate the results into PULO, and devise new approaches to validate the final resource (as previous enrichment approaches were not totally reliable).

### 5.1. DLPC and PULO intersection

As explained before, we performed two experiments on intersecting DLPC and PULO, using different intersection sizes: a first one where we required two variants to be present in both synsets for the alignment to take place, and second one, where we were more restrict, requiring an intersection cardinality of three elements.

Our quick evaluation was based on the selection of 100 random synsets (for each experiment), and classifying them in one of three different classes: correct alignment, incorrect alignment, and ambiguous alignment. This last was used when some variants from DLPC were acceptable, but some other, not. Table 4 summarizes the obtained results (in this table,  $n$  is the intersection cardinality).

$n$	Correct	Incorrect	Ambiguous
2	62	27	11
3	76	11	13

Table 4: Evaluation on DLPC and PULO synset alignment.

### 5.2. Intersecting DLPC by triangulation

This second evaluation is based on the same principles as the previous one. We selected 100 random synsets for two different situations:

- when there is a new variant that is suggested to be added to an existing PULO synset (this experiment is named *add*, in the results table);
- when there is a full DLPC synset that is suggested to be added as completely new PULO synset, using ILI information (this experiment is named *new*, in the results table).

Table 5 present the results obtained for these two experiments. As before, the synsets were classified as correct, incorrect or ambiguous. The table also shows the number of variants those 100 synsets include, showing that bigger synsets are usually correct (check also the percentage of correct variants).

Type	Score	Correct	Incorrect	Ambig.	Correct Variants
add	100	67 (166 v)	20 (35 v)	13 (29 v)	72%
new	50	83 (267 v)	6 (19 v)	11 (29 v)	85%

Table 5: Brief evaluation of the results using DLPC synsets projection to English, Spanish and Galician.

## 6. Conclusions

This article presents two experiments on enlarging a WordNet through the use of a conventional monolingual dictionary that includes synonyms information, and the use of translation dictionaries for alignment between these synonyms and WordNet senses.

The idea is not completely new: reusing available resources to enlarge lexical ontologies. Nevertheless, the results of this experiment are promising and might be usable for other language projects, when aiming at the enlargement of their own WordNet.

### 6.1. Future work

In some situations, the current algorithm aligns a PULO synset with more than one DLPC synset. The main reason is, we think, the reduced amount of variants and synsets existing in PULO. This might be fixed if we make the algorithm use some of the following procedures:

1. create multisets of the addition suggestions;
2. use the multisets cardinality as yet another confidence measure;
3. use a morphological analyzer to match each term part-of-speech with the synset morphological category.

Other than this future work on this specific enlargement experiment, we are working on algorithms to validate variants and synsets already existing in PULO. That is, in the next tasks we will prioritize the task of validating what we already have, changing variants confidence measures and, possibly, removing some existing mistakes.

## 7. Acknowledgments

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – *Fundação para a Ciência e Tecnologia* within the Project Scope: UID/CEC/00319/2013; and thanks to the Project SKATeR (TIN2012-38584-C06-04) supported by the Ministry of Economy and Competitiveness of the Spanish Government.

## 8. Bibliographical references

- Almeida, J. J., Araújo, S., Carvalho, N., Dias, I., Oliveira, A., Santos, A., and Simões, A. (2014). The Per-Fide corpus: A new resource for corpus-based terminology, contrastive linguistics and translation studies. In Tony Berber Sardinha et al., editors, *Working with Portuguese Corpora*, pages 177–200. Bloomsbury Publishing, April.
- Bilgin, O., Çetinoglu, O., and Oflazer, K. (2004). Building a wordnet for Turkish. *Romanian Journal of Information Science and Technology*, 7(1-2):163–172.
- João Malaca Casteleiro, editor. (2006). *Dicionário da Língua Portuguesa Contemporânea*. Academia das Ciências de Lisboa, Verbo.
- Forcada, M. L. (2009). Apertium: traducció automàtica de codi obert per a les llengües romàniques. *Linguamática*, 1(1):13–23, May.
- Gómez Clemente, X., Gómez Guinovart, X., and Simões, A. (2015). *Dicionario de sinónimos do galego*. Xerais, Vigo.
- Gómez Guinovart, X. and Simões, A. (2013). Retreading dictionaries for the 21st century. In José Paulo Leal, et al., editors, *2nd Symposium on Languages, Applications and Technologies*, pages 115–126, Saarbrücken. Dagstuhl Publishing.
- Gómez Guinovart, X. and Solla Portela, M. A. (2014). O dicionario de sinónimos como recurso para a expansión de wordnet. *Linguamática*, 6(2):69–74.
- Gómez Guinovart, X. (2014). Do dicionario de sinónimos á rede semántica: fontes lexicográficas na construción do WordNet do galego. In Ana Gabriela Macedo, et al., editors, *XV Colóquio de Outono - As humanidades e as ciencias: disjunções e confluências*, pages 331–358, Braga. CEHUM, Universidade do Minho.
- Gonçalo Oliveira, H. and Gomes, P. (2014). ECO and Onto.PT: a flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation*, 48(2):373–393.
- Gonçalo Oliveira, H., de Paiva, V., Freitas, C., Rademaker, A., Real, L., and Simões, A. (2015). As Wordnets do Português. In Alberto Simões, et al., editors, *Linguística, Informática e Tradução: Mundos que se Cruzam*, volume 7, pages 397–424. March.
- Gonçalo Oliveira, H. (2013). *Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*. Tese de doutoramento, Universidade de Coimbra.
- Gonzalez-Agirre, A. and Rigau, G. (2013). Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository. *Linguamática*, 5(1):13–28, Julho.
- Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.
- Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.
- Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). Danned: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.
- Simões, A. M. and Almeida, J. J. (2003). NATools – a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–224, September.
- Simões, A. and Almeida, J. J. (2015). Experiments on enlarging a lexical ontology. In José-Luís Sierra-Rodríguez, et al., editors, *IV Symposium on Languages, Applications and Technologies*, pages 137–142.
- Simões, A. and Gómez Guinovart, X. (2013). Dictionary Alignment by Rewrite-based Entry Translation. In José Paulo Leal, et al., editors, *2nd Symposium on Languages, Applications and Technologies*, volume 29 of *OpenAccess Series in Informatics (OASICs)*, pages 237–247, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- Simões, A. and Gómez Guinovart, X. (2014). Bootstrapping a Portuguese WordNet from Galician, Spanish and English wordnets. *Advances in Speech and Language Technologies for Iberian Languages*, 8854:239–248.
- Simões, A., Almeida, J. J., and Carvalho, N. R. (2013). Defining a probabilistic translation dictionaries algebra. In Luís Correia, et al., editors, *XVI Portuguese Conference on Artificial Intelligence - EPIA*, pages 444–455, Angra do Heroísmo, Azores, September.
- Solla Portela, M. A. and Gómez Guinovart, X. (2014). Ampliación de WordNet mediante extracción léxica a partir de un diccionario de sinónimos. In L. Alfonso Ureña López et al., editor, *Actas de las V Jornadas de la Red en Tratamiento de la Información Multilingüe y Multimodal*, volume 1199, pages 29–32, Aachen. CEUR Workshop Proceedings (CEUR-WS.org).
- Vanhoutte, E. (2004). An Introduction to the TEI and the TEI Consortium. *Literary and Linguistic Computing*, 19(1):9–16.

## 9. Language resource references

- Xavier Gómez Guinovart and Solla Portela, Miguel Anxo. (2015). *Galnet (Galician WordNet)*. SLI: University of Vigo / MCR: Multilingual Central Repository, ISLRN 544-286-653-437-9.
- Xavier Gómez Guinovart. (2015). *CLUVI Parallel Corpus*. SLI: University of Vigo, ISLRN 910-993-402-072-9.
- Alberto Simões and Xavier Gómez Guinovart and José João Almeida. (2015). *PULO: Portuguese Unifed Lexical Ontology*. MCR: Multilingual Central Repository, 20150502, ISLRN 113-808-555-300-0.