


# Acquiring Domain-Specific Knowledge for WordNet from a Terminological Database

Alberto Simões 

2Ai-Polytechnic Institute of Cávado and Ave, 4750-810 Barcelos, Portugal  
asimoes@ipca.pt

Xavier Gómez Guinovart 

SLI-TALG, Universidade de Vigo, Vigo, Galiza  
xgg@uvigo.gal

---

## Abstract

---

In this research we explore a terminological database (Termoteca) in order to expand the Portuguese and Galician wordnets (PULO and Galnet) with the addition of new synset variants (word forms for a concept), usage examples for the variants, and synset glosses or definitions.

The methodology applied in this experiment is based on the alignment between concepts of WordNet (synsets) and concepts described in Termoteca (terminological records), taking into account the lexical forms in both resources, their morphological category and their knowledge domains, using the information provided by the WordNet Domains Hierarchy and the Termoteca field domains to reduce the incidence of polysemy and homography in the results of the experiment.

The results obtained confirm our hypothesis that the combined use of the semantic domain information included in both resources makes it possible to minimise the problem of lexical ambiguity and to obtain a very acceptable index of precision in terminological information extraction tasks, attaining a precision above 89% when there are two or more different languages sharing at least one lexical form between the synset in Galnet and the Termoteca record.

**2012 ACM Subject Classification** Computing methodologies → Language resources; Computing methodologies → Natural language processing

**Keywords and phrases** WordNet, Terminology, Lexical Resources, Natural Language Processing

**Digital Object Identifier** 10.4230/OASICS.SLATE.2019.6

**Funding** *Alberto Simões*: This research has been carried out thanks to the “Programa IACOBUS”, coordinated by Interreg España-Portugal, CCDRN Portugal and Xunta de Galicia.

*Xavier Gómez Guinovart*: This research has been carried out thanks to the project DeepReading (RTI2018-096846-B-C21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government and the European Fund for Regional Development (MCIU/AEI/FEDER,UE).

## 1 Introduction

Princeton WordNet (PWN) [12, 26] is undoubtedly one of the most successful resources ever built. Even though it was not developed specifically for natural language processing (NLP), its usage in this field is indispensable. The relevance of this resource in NLP led scientists from all around the world to work on the creation of similar resources for their languages. The main problem on creating those kind of resources is the amount of specialised labour required for this purpose. While the PWN for English was created manually from scratch, most wordnets for other languages are created using automatic methods, followed by a more superficial or in-depth manual analysis.



© Alberto Simões and Xavier Gómez Guinovart;  
licensed under Creative Commons License CC-BY

8th Symposium on Languages, Applications and Technologies (SLATE 2019).

Editors: Ricardo Rodrigues, Jan Janoušek, Luís Ferreira, Luísa Coheur, Fernando Batista, and Hugo Gonçalo Oliveira; Article No. 6; pp. 6:1–6:13



OpenAccess Series in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Galnet [19]<sup>1</sup> and PULO [37]<sup>2</sup>, two wordnets built semi-automatically for Galician and Portuguese, have been enlarged during the last years with a set of experiments of lexical acquisition that explore different resources and methods, extracting data for inclusion in the knowledge databases after expert review and validation [3, 15, 38, 40, 41]

While there is only one wordnet project for the Galician language – Galnet –, for the Portuguese language there are other projects aiming at the creation of WordNet-based linked lexical resources, like Open Multilingual Wordnet (OMW), OpenWordNet-PT, Ufes WordNet or Onto.PT [11].

This document presents a further experiment in order to enrich Galnet and PULO not just with new synset variants, but also with variant usage examples and synset glosses. This experiment explores Termoteca [14], a terminological database that includes records for different areas, ranging from medicine to tourism, and including data for Galician, English, Portuguese, Spanish and French.

The experiment is based on the alignment between concepts included in WordNet (synsets) and concepts described in Termoteca (terminological records). Each entry in Termoteca contains terms that refer to a specific concept, and these concepts are very much like synsets. This alignment between WordNet and Termoteca concepts was performed taking into account the lexical forms already existing in both resources, their morphological category and their knowledge domains.

Our hypothesis is that using specific domain terms, aligned by form, field and category, will reduce the incidence of lexical polysemy (and homography) and will therefore contribute to raising the quality (and precision) level of the extracted information.

The document is organised as follows: Section 2 presents related work, reviewing a number of similar approaches based on lexical resources used to enlarge different languages wordnets. Section 3 describes the specific lexical resources used in the experiment and Section 4 discusses the algorithm designed for their exploitation. In Section 5 the results are evaluated, analysed and commented. The article concludes with Section 6 where some final remarks and future work is presented.

## **2 Related Work**

The Galnet and PULO wordnets have been created from PWN 3.0, following the expand model [42], where the variants associated with the PWN synsets are obtained through different strategies. This model has also been used in the development of the wordnets for Italian [30], Indonesian [33], Hungarian [25], Croatian [34], French – WOLF [36] and WoNeF [32] wordnets – and Kurdish [2]. The same approach has been taken in the MCR framework [20] for the creation of the wordnets of Spanish [6], Catalan [8] and Basque [31].

In the expand model, the main methodology used to extend a wordnet coverage from the variants associated with the PWN synsets is the acquisition of their translations from existing lexical resources. We have applied that methodology in previous phases of the Galnet and PULO developments.

On the one hand, we have used the WN-Toolkit [28] – a set of Python programs for the creation or enlargement of wordnets – to expand the Galnet first distributions from different existing bilingual English–Galician resources: Wikipedia (whose Galician version is known as

---

<sup>1</sup> <http://sli.uvigo.gal/galnet/>

<sup>2</sup> <http://wordnet.pt>

Galipedia)<sup>3</sup>, the English–Galician CLUVI Dictionary [4]<sup>4</sup>, the Apertium English-Galician dictionary<sup>5</sup>, the Galizionario (the Galician Wiktionary)<sup>6</sup>, Babelnet 2.0<sup>7</sup>, the multilingual dictionary OmegaWiki<sup>8</sup>, the database of toponyms GeoNames<sup>9</sup>, and the catalogue of species Wikispecies<sup>10</sup> [17]. Due to the difficulty of this task, the use of automatic extraction techniques was complemented with an arduous process of human revision where the variant candidates identified by the extraction tool were either approved or rejected one by one by a team of reviewers, but no comparable evaluation measures for precision were provided.

On the other hand, we have designed several lexical extraction experiments aimed to enlarge the coverage of Galnet and PULO from the lexical information contained in classical monolingual dictionaries for the Galician and Portuguese language, using the *Dicionario de Sinónimos do Galego* [13]<sup>11</sup>, the *Dicionário da Língua Portuguesa Contemporânea* (DLPC) [10] and the *Dicionário Aberto*<sup>12</sup>.

In the first case, the methodology used for the extraction from the *Dicionario de Sinónimos do Galego* was based on the matching of lexical forms among the variants of Galnet synsets and the variants of dictionary synsets – i.e. the lexical forms included in each dictionary entry [15, 18, 41]. 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. The highest precision obtained with this method is about 65%, selecting the new candidates among the variants appearing only once in the dictionary of synonyms and in Galnet.

In the second experiment, we present an exploratory approach to enrich the PULO lexical ontology with the synonyms present in the DLPC [40]. 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 which are then aligned with the WordNet synsets. We also project the Portuguese terms into English, Spanish and Galician. This process allowed both the addition of new term variants to existing synsets, as the creation of new synsets for Portuguese. An evaluation of synonym extraction based on the selection of 100 random synsets gave a precision of 62% for intersections with two variants in both synsets and 76% for intersections with three variants.

Following a similar methodology, the third experiment [39] explores the entries of the *Dicionário Aberto* in order to extract synsets from its entries. This dictionary includes an interesting way to present definitions based on synonyms. This was exploited to extract synsets from the dictionary (bags of words presented as synonyms). These synsets were intersected with existing synsets in PULO. The experiment got an accuracy of 58% for intersections of two variants from the evaluation of 200 random synsets.

Other related experiments which use bilingual lexical resources to enhance existing wordnets from existing bilingual English dictionaries by intersecting lemmas are conducted for Sanskrit [9], Bengali [35], Croatian [22] or Moroccan Darija [27]. In all the cases, the automatic extraction results are subjected to human revision and validation. In the case of Croatian, the only of these experiments with a true evaluation, the reported precision of the automatic process is about 30%.

<sup>3</sup> <http://gl.wikipedia.org>

<sup>4</sup> <http://sli.uvigo.gal/dicionario/>

<sup>5</sup> <http://sourceforge.net/projects/apertium/>

<sup>6</sup> <http://gl.wiktionary.org>

<sup>7</sup> <http://babelnet.org>

<sup>8</sup> <http://www.omegawiki.org>

<sup>9</sup> <http://www.geonames.org>

<sup>10</sup> <http://species.wikimedia.org>

<sup>11</sup> <http://sli.uvigo.gal/dicionario/>

<sup>12</sup> <http://dicionario-aberto.net>

### 3 Resources

This section describes the three main resources used for this experiment: the terminological database Termoteca, the English, Portuguese and Galician wordnets, as well as the WordNet Domains Hierarchy.

#### 3.1 Termoteca

The Termoteca is a corpus-based Galician-centred multilingual terminological database based on the monolingual and parallel specialty texts collected in the CLUVI Parallel Corpus [16]<sup>13</sup> and in the CTG Galician Technical Corpus [1]<sup>14</sup>.

This terminological database is freely accessible<sup>15</sup> and freely downloadable<sup>16</sup> on the web under a CC-BY 4.0 license. It contains 8,085 records with information about 16,387 terms in Galician (8,172 terms), Spanish (3,257), English (3,031), Portuguese (1,112) and French (815) documented in the CLUVI and CTG corpora, and belonging to the areas of law (1,681 records), sociology (1,145), economy (1,268), ecology (1,673), computer science (564), medicine (1,155) and tourism (1,176)<sup>17</sup>.

The information extracted from the corpora and collected in the Termoteca includes the terms, their contexts, and their intra- and inter-linguistic formal variants together with their frequencies of use. Additionally, it includes their definition and their semantic relations (antonyms, holonyms, hyperonyms, etc.) with other terms, when they are explicitly coded in the textual corpora. Finally, all the terminological records are catalogued according to their thematic field, with reference to a conceptual ontology hierarchy.

#### 3.2 PWN, MCR, Galnet and PULO

PWN is a lexical database of the English language, organised as a semantic network where the nodes are concepts represented as sets of synonyms and the links between nodes are semantic relations between lexical concepts. These nodes contain nouns, verbs, adjectives and adverbs grouped by synonymy. In WordNet terminology, a set of synonyms is called a *synset*. The term *variant* applied to WordNet refers to each synonym in a synset, which is considered a lexical variant of the same concept. Thus, each synset represents a distinct lexicalised concept and includes all the synonymous variants of this concept. Additionally, each synset must contain (at least in PWN) a brief definition or gloss, which is common to every variant in the synset, and, in some cases, one or more examples of the use of the variants in context.

Both Galnet for Galician and PULO for Portuguese wordnets are part of the Multilingual Central Repository (MCR) [20]<sup>18</sup>, a database that currently integrates wordnets from six different languages (English, Spanish, Catalan, Galician, Basque and Portuguese) with PWN 3.0 as Interlingual Index (ILI). Table 1 provides the number of synsets and variants for the different languages gathered in this repository, and their percentage of development with respect to the PWN.

<sup>13</sup><http://sli.uvigo.gal/CLUVI/>

<sup>14</sup><http://sli.uvigo.gal/CTG/>

<sup>15</sup><http://sli.uvigo.gal/termoteca/>

<sup>16</sup>[http://sli.uvigo.gal/download/SLI\\_Termoteca/](http://sli.uvigo.gal/download/SLI_Termoteca/)

<sup>17</sup>In the Termoteca, each record could be assigned to more than one thematic domain. This is why the total number of assigned domains (8,665) is slightly superior to the number of records (8,085).

<sup>18</sup><http://adimen.si.ehu.es/web/MCR/>

■ **Table 1** Current coverage relative to English of wordnets in MCR.

	English (PWN 3.0)		Galician (Galnet 3.0.28)	
	variants	synsets	variants	synsets
Total	206,941	117,659	70,030	43,043
%	100%	100%	33.8%	36.6%
	Spanish (MCR 2016)		Portuguese (MCR 2016)	
Total	146,501	78,995	32,604	17,942
%	70.8%	67.1%	15.8%	15.2%
	Catalan (MCR 2016)		Basque (MCR 2016)	
Total	100,793	60,956	50,037	30,263
%	48.7%	51.8%	24.2%	25.7%

It is also worth noting that the concepts contained in the MCR are categorised into domain hierarchies and ontologies, such as the WordNet Domains [7], the Suggested Upper Merged Ontology (SUMO) [29] and the Top Concept Ontology [5], which allows the various applications benefiting from these semantic categorisations to make better use of the resource.

### 3.3 WordNet Domains

The WordNet Domains hierarchy (WDH) is a freely available<sup>19</sup> lexical resource created in a semi-automatic way by augmenting WordNet synsets with one or more domain labels selected from an original set of 165 hierarchically organised semantic fields. These domains are mainly based on the subject field codes used in lexicography, and on the subject codes from the Dewey Decimal Classification (DDC), a general taxonomy used worldwide for library organisation. For the purposes of this experiment, we use the version of WDH distributed with the MCR resource.

The exploitation of WDH permits reducing word polysemy and grouping the synsets by domain. For instance, the noun *tongue* has eight senses in PWN 3.0, but the first sense *tongue#1* (included in the PWN 3.0 synset with the inter-linguistic offset 05301072-n) is labelled with the domain label ANATOMY, the second sense *tongue#2* (06904171-n) with the label LINGUISTICS, *tongue#3* (13918387-n) with the label FACTOTUM, *tongue#4* (07082198-n) with ART, *tongue#5* (09442595-n) with *Geography*, *tongue#6* (07652995-n) with GASTRONOMY and *tongue#7* (04450994-n) with FASHION.

WDH has been used in several NLP tasks, as word-sense disambiguation [21, 23] or text categorisation [24]. In this experiment, we use WDH to intersect the Termoteca with WordNet by means of the semantic field domains coded in both lexical resources, in order to acquire terminological information from the Termoteca to be incorporated in WordNet.

## 4 Methodology

As pointed out earlier, the experiment is based on the alignment between the synsets in WordNet and the terminological records in Termoteca, taking into account the lexical forms, their morphological category and their knowledge domains. Compared with other techniques,

<sup>19</sup><http://wndomains.fbk.eu>

■ **Table 2** Mapping from Termoteca domains to WordNet Domains.

Termoteca	WordNet Domains Hierarchy
ACHEGA (economy)	ADMINISTRATION, COMMERCE, INDUSTRY
AUGA (environment)	ENVIRONMENT, BIOLOGY, AGRICULTURE, EARTH
GALEX (law)	LAW, ADMINISTRATION, POLITICS, ECONOMY
MEDIGAL (medicine)	MEDICINE, PSYCHOLOGY, HEALTH
TURIGAL (tourism)	FOOD, TOURISM, TRANSPORT
UNESCO (sociology)	ART, SOCIAL SCIENCE, TELECOMMUNICATION, POLITICS, SEXUALITY, PSYCHOLOGICAL FEATURES
XIGA (computing)	COMPUTER SCIENCE, TELECOMMUNICATION, ENGINEERING

the methodology applied in this research is characterised by the use of domain information, which is rarely found in the lexical resources employed to extend wordnets.

In order to trace the alignments between the concepts in Termoteca and WordNet, three different experiments were performed (using the same algorithm, but different data):

**Experiment 1:** Treat each domain independently, both in WordNet and in Termoteca. For this purpose, a rough mapping between the Termoteca domains and the WordNet Domains was created, as presented in Table 2). Note that we just present the top classes of both ontologies, although all sub-domains were used. As an example, the MEDICINE WDH domain includes all their child domains, like DENTISTRY or SURGERY. In this experiment, for each mapping, we take into account all the WordNet synsets and all the Termoteca records pertaining to any domain in the considered mapping.

**Experiment 2:** Use the synsets from WordNet for the domains included in Table 2 as a whole, as well as all the Termoteca records. This experiment may produce some relevant inter-domain alignments, but the resulting precision will diminish.

**Experiment 3:** Use the synsets from WordNet independently of their domain, as well as the whole Termoteca.

Some of the WordNet synsets were discarded from all these experiments:

- All synsets composed exclusively by terms starting with an upper case letter were not considered. The main reason for this is the amount of proper names available in WordNet, ranging from toponyms to anthroponym. As Termoteca does not include proper names, their analysis is not relevant.
- All adverbs were also discarded, as our experiment is focused on terminological information and adverbs are not usually considered in terminology work.

Table 3 shows some statistics about the data and results of the three experiments just mentioned.

The first line includes the number of considered synsets. The next eight lines show the synsets coverage per language.

For instance, looking up the ACHEGA column, the number of considered synsets is 5 172. From these, only 575 synsets include a variant in the four languages, while 2 725 synsets just have variants in the English language.

For each of these experiments, each selected WordNet synset was compared with each selected Termoteca record. An alignment was defined between a synset and a record if at least one variant (for any of the four languages considered) is shared and if their morphological category is the same (verb, noun or adjective).

**Table 3** Statistics per experiment: (i) number of considered synsets, (ii) synset coverage per languages, (iii) aligned synsets and alignment score (maximum  $\uparrow$  and average  $\bar{x}$ ), and (iv) variants, examples and gloss extraction metrics.

	Experiment 3			Experiment 2						Experiment 1								
	Whole WordNet	Relevant Domains	Synsets	ACHEGA	AUGA	GALEX	MEDICAL	TURIGAL	UNESCO	XIGA	ACHEGA	AUGA	GALEX	MEDICAL	TURIGAL	UNESCO	XIGA	
	101 162	43 936	5 172	20 920	3 160	4 710	5 607	6 110	2 024									
EN ES GL PT	10 913	3 275	575	922	462	357	404	872	180	8,9%	7,5%	4,4%	14,6%	7,6%	7,2%	14,3%	8,9%	
EN ES GL	6 667	4 310	185	3 110	148	314	242	306	228	11,3%	9,8%	4,4%	4,7%	6,7%	4,3%	5,0%	11,3%	
EN GL PT	638	209	61	61	53	31	19	48	9	0,4%	0,5%	0,3%	1,7%	0,7%	0,3%	0,8%	0,4%	
EN ES PT	5 212	1 554	271	295	196	248	227	443	80	4,0%	3,5%	1,4%	6,2%	5,3%	4,1%	7,3%	4,0%	
EN GL	17 883	9 789	686	6 448	385	1 343	502	734	327	16,2%	22,3%	30,8%	12,2%	28,5%	9,0%	12,0%	16,2%	
EN PT	394	124	5	19	9	61	6	34	2	0,1%	0,3%	0,3%	0,3%	1,3%	0,11%	0,6%	0,1%	
EN ES	10 832	4 745	664	1 588	475	469	923	845	281	13,9%	10,8%	7,6%	15,0%	10,0%	16,5%	13,8%	13,9%	
EN	48 623	19 930	2 725	8 477	1 433	1 887	3 284	2 828	917	45,3%	45,4%	40,1%	45,4%	40,1%	58,6%	46,3%	45,3%	
Aligned Synsets	5 750	2 358	294	195	233	63	172	282	76	3,8%	5,4%	0,9%	7,4%	1,3%	3,1%	4,6%	3,8%	
Total Alignments	7 654	3 212	343	202	279	67	250	315	81									
Align Scores	$\uparrow=4.5$ $\bar{x}=1.38$	$\uparrow=4.5$ $\bar{x}=1.37$	$\uparrow=3.0$ $\bar{x}=1.17$	$\uparrow=3.0$ $\bar{x}=1.01$	$\uparrow=3.0$ $\bar{x}=1.44$	$\uparrow=2.0$ $\bar{x}=1.07$	$\uparrow=4.0$ $\bar{x}=1.34$	$\uparrow=4.0$ $\bar{x}=1.73$	$\uparrow=4.5$ $\bar{x}=1.83$									
New GL variants	3 984	1 405	77	26	138	40	61	186	126									
New GL examples	8 292	3 402	357	229	317	116	109	370	177									
New PT variants	1 198	479	—	—	—	—	196	—	—									
New PT examples	1 619	667	—	—	—	—	264	—	—									
New GL glosses	652	381	—	27	—	61	—	6	9									



Each alignment was then scored according to its alignment strength. A score of 1.0 was assigned for each different language sharing at least one variant. A score of 0.5 was added for each extra variant shared per language. Thus, if two variants for a language are shared, that language alignment will score 1.5. Table 3 presents three lines regarding alignment metrics. The first shows the number of synsets aligned with at least one Termoteca record. The second line shows the total number of alignments. For example, if a synset can be aligned with two different Termoteca records, there are two possible alignments. Finally, the third line presents the maximum score achieved in this experiment, as well as their average.

Every possible alignment is then processed in order to understand what information from Termoteca can be used to enrich PULO or Galnet. Termoteca includes three different kinds of relevant data:

- Portuguese and Galician variants that are not present in PULO or Galnet;
- Galician definitions for synsets missing a Galician gloss<sup>20</sup>;
- Galician or Portuguese usage examples, when none is available<sup>21</sup>.

Table 3 also presents the amount of items extracted for each one of these categories. The lack of Portuguese suggestions for all experiments – except for those that include TURIGAL – is justified by the fact that Termoteca only includes Portuguese terms and examples for the tourism domain.

## 5 Results and Evaluation

Considering the number of alignments in the different experiments, and taking into account the amount of extracted information, for the evaluation we only considered the experiment in the tourism domain. The main reason is that it is the only experiment with contributions for both the Portuguese and Galician languages. Although no glosses are suggested by the Termoteca in this knowledge area, our evaluation is based on the concept alignment quality and not in each specific information item extracted. That is, if a form  $w$  from terminology entry  $\mathcal{T}$  is suggested to be added to the synset  $\mathcal{S}$ , and even if that form might seem adequate to be added to  $\mathcal{S}$ , it will only be considered correct if synset  $\mathcal{S}$  and entry  $\mathcal{T}$  refer to the same concept  $\mathcal{C}$ . Thus, our evaluation is not based exactly on which information items are contributed to Galnet or PULO, but rather on the alignment of the concepts. When a specific alignment is correct all this record information should be correct given that Termoteca was manually produced.

All the 250 alignments obtained for this domain were manually evaluated. Whereas the number of results is low, it should be noted that there are other alignments from different domains that can be used for Galician and that, if the methodology proves effective, the same methodology can be applied to other terminological resources. Each concept alignment was classified as being (i) correct, (ii) incorrect or (iii) incorrect due to mistakes found in the lexical resources used in the experiment:

### (i) Correct Alignments

From the 250 alignments, 160 were classified as correct alignments. Table 4 splits this evaluation by score, showing that when there are two or more common languages between the synset in Galnet and the Termoteca record, the alignment has a precision above 89%.

<sup>20</sup> PULO has machine translation glosses for every synset and therefore no definitions were considered for the Portuguese language. Also, Termoteca includes just about 10 records with a Portuguese definition.

<sup>21</sup> As PULO does not include any example, all examples in the Portuguese language were considered.



■ **Table 4** Evaluation metric results for the tourism domain.

Score	Alignments	Correct	Incorrect	Error in resource	Precision
4.0	6	5 (83%)	1 (17%)	0	0.83
3.5	2	2 (100%)	0	0	1.00
3.0	6	5 (83%)	1 (17%)	0	0.83
2.5	4	4 (100%)	0	0	1.00
2.0	39	33 (85%)	4 (10%)	2 (5%)	0.89
1.5	2	2 (100%)	0	0	1.00
1.0	191	109 (57%)	74 (39%)	4 (2%)	0.58

Although both a bilingual link (two variants from different languages) and three monolingual links (three variants from a single language) yield the same score, there are only eight alignments sharing two variants in the same language – and no alignment shares more than two variants in the same language. Furthermore, there are six alignments sharing at least one variant in four different languages, and ten alignments sharing at least one variant in three different languages.

As the number of links decreases to just one language, the precision goes down to 58%, but still comparable with the experiments referred in Section 2, as the evaluations given for those experiments (with precision values of 65%, 62% and 58%) are based on alignments with at least two shared variants. Table 5 shows an example of a correct alignment (with score 1.0).

#### (ii) Incorrect Alignments

Incorrect alignments are due to the polysemy of lexical forms that are present in two or more different concepts, although sharing the same terminological domain. This is specially true in the tourism knowledge domain, as it intersects with different areas, like history, architecture, leisure, etc. Table 6 shows one such problem, where the form “*nave*” is used to refer both to a type of boat and to a section of a church or cathedral.

#### (iii) Errors in the Resources

There are three main types of errors found in the used resources: (i) synsets that are classified in the wrong domain (as WDH was expanded automatically to cover all WordNet, some mistakes exist) (ii) wrong variants in MCR wordnets and (iii) Termoteca records with incomplete information. One example of an entry from a different domain is presented in Table 7, where “*colina*” is used both as a biological term and a geographic term. On the other hand, Table 8 shows an example of an incomplete record in Termoteca. In this case, although the alignment is correct, the context of use for the term gathered in Termoteca is not valid.

## 6 Final Remarks

We presented a methodology to align a terminological database with WordNet at the concept level, with the objective of acquiring domain specific terms, usage examples and definitions to enrich the Portuguese and Galician wordnets.

To validate the proposed methodology we used the MCR (Multilingual Central Repository) wordnets – that include the English, Spanish, Galician and Portuguese wordnets linked at the concept level –, the WordNet Domains hierarchy (WDH) and the Termoteca, a multilingual terminological database focusing on different fields of knowledge.

■ **Table 5** Correct alignment with score 1.0.

<b>ILI:</b> 03594945-n	<b>Termoteca ID:</b> 2220741
<b>WordNet gloss:</b>	a car suitable for travelling over rough terrain
<b>New PT variants:</b>	jipe, viatura todo-o-terreno, TT, 4x4, veículo todo-o-terreno, jeep
<b>New PT Examples:</b>	
veículo todo-o-terreno	Só é aconselhável a veículos todo-o-terreno porque a calçada, apesar de curta, é de meados do século XIX e apresenta alguns troços em mau estado.
4x4	Circuito Megalítico de Barbacena (em 4x4).
jeep	Este, pode ser o início de um dos muitos circuitos que, a pé ou de jeep, levam à descoberta da Serra da Lousã.
TT	Acessibilidade: Boa para TT e BTT.
viatura todo-o-terreno	Partindo do Centro de Recepção de Muxagata e de Castelo Melhor, os visitantes do Parque Arqueológico serão transportados em viaturas todo-o-terreno para apreciar pormenorizadamente todo o ciclo artístico.
jipe	O trajecto só se aconselha a quem viajar de jipe ou não se importar de meter o automóvel por maus caminhos.

■ **Table 6** Alignment error.

<b>ILI:</b> 04194289-n	<b>Termoteca ID:</b> 2220366
<b>WordNet gloss:</b>	a vessel that carries passengers or freight
<b>New PT Example:</b>	
nave	No interior, a catedral é composta por três naves principais e três capelas.
<b>New GL Example:</b>	
nave	Como manda o canon, son rexos edificios de cantaría, cada un dividido en tres naves, a central abovedada, e cuxa estrutura ternaria queda tamén reflectida nas fachadas das frontes, que teñen tres portas e tres ventás apoiadas nos oportunos arcos de medio punto peraltados.

■ **Table 7** Incorrect alignment due to an error in the WDH.

<b>ILI:</b> 14810561-n	<b>Termoteca ID:</b> 2220018
<b>WordNet gloss:</b>	a B-complex vitamin that is a constituent of lecithin; essential in the metabolism of fat
<b>New PT Example:</b>	
colina	O Maciço desenvolve-se em duas grandes unidades morfológicas: a primeira, situada no lado oriental e dominada por um conjunto de colinas dolomíticas formadas a partir dos 300 metros de altitude e onde se distingue a depressão do Rabaçal.

■ **Table 8** Incorrect alignment due to an incomplete record in the Termoteca.

<b>ILI:</b> 03093427-n	<b>Termoteca ID:</b> 2220920
<b>WordNet gloss:</b>	diplomatic building that serves as the residence or workplace of a consul
<b>New PT Example:</b>	
consulado	Consulado

The alignment was mainly based on the semantic domain information included in WDH and the Termoteca, thus minimising the problems derived from the lexical ambiguity of the terms. The evaluation of the results in the tourism domain shown that the methodology has a very valuable precision, even when using no more than one term to link the concepts. Currently, the obtained resources are being manually validated and added to their respective wordnets.

Taking advantage of the high precision obtained in this approach, we are planning the alignment of WordNet with other relevant terminological databases, like bUSCatermos<sup>22</sup> for Galician and IATE<sup>23</sup> for Portuguese.

---

## References

- 1 Rodrigo Agerri, Xavier Gómez Guinovart, German Rigau, and Miguel Anxo Solla Portela. Developing New Linguistic Resources and Tools for the Galician Language. In *Eleventh International Conference on Language Resources and Evaluation (LREC)*, pages 2322–2325, 2018.
- 2 Purya Aliabadi, Mohamed Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. Towards building KurdNet, the Kurdish WordNet. In *7th Global Wordnet Conference (GWC)*, pages 1–6, 2014.
- 3 María Álvarez de la Granja, Xosé María Gómez Clemente, and Xavier Gómez Guinovart. Introducing Idioms in the Galician WordNet: Methods, Problems and Results. *Open Linguistics*, 2(1):253–286, 2016. doi:10.1515/opli-2016-0012.
- 4 Alberto Álvarez Lugrís and Xavier Gómez Guinovart. Lexicografía bilingüe práctica basada en corpus: planificación y elaboración del Diccionario Moderno Inglés-Galego. In *Lexicografía de las lenguas románicas: Aproximaciones a la lexicografía moderna y contrastiva*, pages 31–48, 2014. doi:10.1515/9783110310337.31.
- 5 Javier Álvez, Jordi Atserias, Jordi Carrera, Salvador Climent, Antoni Oliver, and German Rigau. Consistent Annotation of EuroWordNet with the Top Concept Ontology. In *4th Global WordNet Conference (GWC)*, 2008.
- 6 Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodriguez. Combining multiple methods for the automatic construction of multilingual WordNets. In *Recent Advances in Natural Language Processing II. Selected papers (RANLP)*, volume 97, pages 327–338, 1997.
- 7 Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *COLING Workshop on Multilingual Linguistic Resources*, pages 101–108, 2004.
- 8 Laura Benítez, Sergi Cervell, Gerard Escudero, Mònica López, German Rigau, and Mariona Taulé. Methods and tools for building the Catalan WordNet. In *ELRA Workshop on Language Resources for European Minority Languages*, 1998.
- 9 Sudha Bhingardive, Tanuja Ajojkar, Irawati Kulkarni, Malhar Kulkarni, and Pushpak Bhattacharyya. Semi-Automatic Extension of Sanskrit Wordnet using Bilingual Dictionary. In *7th Global Wordnet Conference (GWC)*, pages 324–329, 2014.
- 10 João Malaca Casteleiro, editor. *Dicionário da Língua Portuguesa Contemporânea*. Academia das Ciências de Lisboa, Verbo, 2006.
- 11 Valeria de Paiva, Livy Real, Hugo Gonçalo Oliveira, Alexandre Rademaker, Cláudia Freitas, and Alberto Simões. An overview of Portuguese WordNets. In Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, and Piek Vossen, editors, *8th Global WordNet Conference (GWC2016)*, pages 74–81, 2016.

---

<sup>22</sup> <https://aplicacions.usc.es/buscatermos/publica/index.htm>

<sup>23</sup> <https://iate.europa.eu>

- 12 Christiane Fellbaum, editor. *WordNet: An electronic lexical database*. MIT Press, Cambridge, 1998.
- 13 Xosé María Gómez Clemente, Xavier Gómez Guinovart, and Alberto Simões. *Dicionario de sinónimos do galego*. Xerais, Vigo, 2015.
- 14 Xavier Gómez Guinovart. A hybrid corpus-based approach to bilingual terminology extraction. In *Encoding the past, decoding the future: corpora in the 21st Century*, pages 147–175, 2012.
- 15 Xavier Gómez Guinovart. Do dicionario de sinónimos á rede semántica: fontes lexicográficas na construción do WordNet do Galego. In *XV Colóquio de Outono: As humanidades e as ciencias: disjunções e confluências*, pages 331–358, 2014.
- 16 Xavier Gómez Guinovart. Enriching parallel corpora with multimedia and lexical semantics: From the CLUVI Corpus to WordNet and SemCor. In *Parallel Corpora for Contrastive and Translation Studies: New resources and applications*, pages 141–158. John Benjamins, Amsterdam, 2019. doi:10.1075/sc1.90.09gom.
- 17 Xavier Gómez Guinovart and Antoni Oliver. Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit. *Procesamiento del Lenguaje Natural*, 53:43–50, 2014.
- 18 Xavier Gómez Guinovart and Miguel Anxo Solla Portela. O dicionario de sinónimos como recurso para a expansión de WordNet. *Linguamática*, 6(2):69–74, 2014.
- 19 Xavier Gómez Guinovart and Miguel Anxo Solla Portela. Building the Galician wordnet: methods and applications. *Language Resources and Evaluation*, 52(1):317–339, 2018. doi:10.1007/s10579-017-9408-5.
- 20 Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. Multilingual Central Repository version 3.0. In *8th International Conference on Language Resources and Evaluation (LREC)*, pages 2525–2529, 2012.
- 21 Sopan Govind Kolte and Sunil G. Bhirud. Word Sense Disambiguation Using WordNet Domains. In *1st International Conference on Emerging Trends in Engineering and Technology*, pages 1187–1191, July 2008. doi:10.1109/ICETET.2008.231.
- 22 Matea Filko Krešimir Šojat and Antoni Oliver. Further expansion of the Croatian WordNet. In *9th Global WordNet Conference (GWC)*, 2018.
- 23 Wei Jan Lee and Edwin Mit. Word Sense Disambiguation by using domain knowledge. In *International Conference on Semantic Technology and Information Retrieval*, pages 237–242, June 2011. doi:10.1109/STAIR.2011.5995795.
- 24 Angela Locoro, Daniele Grignani, and Viviana Mascardi. When You Doubt, Abstain: From Misclassification to Epoché in Automatic Text Categorisation. In *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 209–212, August 2011. doi:10.1109/WI-IAT.2011.65.
- 25 Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. Methods and results of the Hungarian wordnet project. In *4th Global WordNet Conference*, pages 387–405, 2008.
- 26 George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.
- 27 Khalil Mrini and Francis Bond. Building the Moroccan Darija WordNet (MDW) using Bilingual Resources. In *International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*, 2017.
- 28 Antoni Oliver. WN-Toolkit: Automatic generation of wordnets following the expand model. In *7th Global Wordnet Conference (GWC)*, pages 7–15, 2014.
- 29 Adam Pease, Ian Niles, and John Li. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, 2002.
- 30 Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet. developing an aligned multilingual database. In *1st International WordNet Conference*, pages 293–302, 2002.

- 31 Elisabete Pociello, Eneko Agirre, and Izaskun Aldezaba. Methodology and Construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2):121–142, 2011. doi:10.1007/s10579-010-9131-y.
- 32 Quentin Pradet, Gaël de Chalendar, and Jaume Baguenier Desormeaux. WoNeF, an improved, expanded and evaluated automatic French translation of WordNet. In *7th Global WordNet Conference (GWC)*, pages 32–39, 2014.
- 33 Desmond Darma Putra, Abdul Arfan, and Ruli Manurung. Building an Indonesian WordNet. In *2nd International MALINDO Workshop*, 2008.
- 34 Ida Raffaeli, Bekavac Božo, Željko Agić, and Marko Tadić. Building Croatian WordNet. In *4th Global WordNet Conference (GWC)*, pages 349–359, 2008.
- 35 K.M. Tahsin Rahit, Tabin Hasan, Md.Al Amin, and Zahiduddin Ahmed. BanglaNet: Towards a WordNet for Bengali language. In *9th Global WordNet Conference (GWC)*, 2018.
- 36 Benoît Sagot and Darja Fišer. Building a free French wordnet from multilingual resources. In *OntoLex*, pages 14–19, 2008.
- 37 Alberto Simões and Xavier Gómez Guinovart. Bootstrapping a Portuguese WordNet from Galician, Spanish and English wordnets. In *Advances in Speech and Language Technologies for Iberian Languages*, volume 8854 of *Lecture Notes in Computer Science*, pages 239–248, 2014.
- 38 Alberto Simões and Xavier Gómez Guinovart. Extending the Galician wordnet using a multilingual Bible through lexical alignment and semantic annotation. In *7th Symposium on Languages, Applications and Technologies (SLATE 2018)*, volume 62 of *OpenAccess Series in Informatics (OASISs)*, pages 14:1–14:13, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/OASISs.SLATE.2018.14.
- 39 Alberto Simões and José João Almeida. Experiments on Enlarging a Lexical Ontology. In *Languages, Applications and Technologies*, volume 563 of *Communications in Computer and Information Science*, pages 49–56. Springer International Publishing, 2015. doi:10.1007/978-3-319-27653-3\_5.
- 40 Alberto Simões, Xavier Gómez Guinovart, and José João Almeida. Enriching a Portuguese WordNet using Synonyms from a Monolingual Dictionary. In *9th International Conference on Language Resources and Evaluation (LREC)*, May 2016.
- 41 Miguel Anxo Solla Portela and Xavier Gómez Guinovart. Ampliación de WordNet mediante extracción léxica a partir de un diccionario de sinónimos. In *Actas de las V Jornadas de la Red en Tratamiento de la Información Multilingüe y Multimodal*, volume 1199, pages 29–32. CEUR Workshop Proceedings (CEUR-WS.org), 2014.
- 42 Piek Vossen, editor. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, 1998.