Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4859–4866
Marseille, 11–16 May 2020

The MWN.PT WordNet for Portuguese: Projection, Validation, Cross-lingual Alignment and Distribution

António Branco, Sara Grilo, Márcia Bolrinha, Chakaveh Saedi, Ruben Branco, João Silva, Andreia Querido, Rita de Carvalho, Rosa del Gaudio, Mariana Avelãs, Clara Pinto
University of Lisbon
NLX—Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências
Campo Grande, 1749-016 Lisboa, Portugal
antonio.branco@di.fc.ul.pt

Abstract

The objective of the present paper is twofold, to present the MWN.PT WordNet and to report on its construction and on the lessons learned with it. The MWN.PT WordNet for Portuguese includes 41,000 concepts, expressed by 38,000 lexical units. Its synsets were manually validated and are linked to semantically equivalent synsets of the Princeton WordNet of English, and thus transitively to the many wordnets for other languages that are also linked to this English wordnet. To the best of our knowledge, it is the largest high quality, manually validated and cross-lingually integrated, wordnet of Portuguese distributed for reuse.

Its construction was initiated more than one decade ago and its description is published for the first time in the present paper. It follows a three step <projection, validation with alignment, completion> methodology consisting on the manual validation and expansion of the outcome of an automatic projection procedure of synsets and their hypernym relations, followed by another automatic procedure that transferred the relations of remaining semantic types across wordnets of different languages.

Keywords: Wordnet, ontology, Portuguese, lexical semantics, lexical semantic network

1. Introduction

Lexical semantic networks are pervasive in Natural Language Processing (NLP) as they play a key role in virtually all major applications, e.g. in information retrieval, information extraction, machine translation, summarization, and question answering, among others. In what concerns NLP specific tasks, ontologies have also been of paramount importance, for instance, in word sense disambiguation, anaphora resolution, and in semantic role assignment, to name just a few.

A great deal of research has been devoted to devising methods that support the development of ontologies in a way as much automatic as possible. Such methods tend to be of quite disparate nature from a technological point of view, and their chances of success rely very much on the quality of the linguistic resources from which the ontologies are to be extracted. These resources may range from highly structured sources of linguistic information (e.g. pre-existing ontologies or bilingual dictionaries) to totally unprocessed materials (e.g. collections of raw texts).

The construction of ontologies of significant size requires a considerable amount of human effort, and the methods aiming at supporting their (semi-)automatic extraction or development are usually said to be addressing the so-called knowledge acquisition bottleneck. However, regardless of how successful these methods may be in tackling this issue, it is worth noting that they will be finding another major bottleneck later on, which could be termed the knowledge validation bottleneck, so much so that many automatically built ontologies simply forgo manual validation of the extracted ontology, or only validate a small set of core concepts.

Against this background, the objective of the present paper is twofold, namely to present MWN.PT, a lexical ontology for Portuguese, and to report on the lessons learned with its construction.

The resulting MWN.PT has a unique set of features. It concerns the Portuguese language, covering both European and American variants; it is a wordnet with 41,000 concepts, expressed by 38,000 different words/expressions; its synsets were manually validated; its synsets are linked to semantically equivalent synsets of the Princeton WordNet of English, and thus transitively to the many wordnets for other languages that are also linked to Princeton WordNet; and it is distributed for free for reuse. To the best of our knowledge, it is the largest high quality, manually validated and cross-lingually integrated, wordnet of Portuguese available. Its construction was initiated in our group more than one decade ago and its description is being published for the first time in the present paper.

The three step <projection, validation with alignment, completion> construction methodology consisted on the manual validation and expansion of the result of an automatic projection procedure of synsets and hypernym relation among them, followed by another automatic procedure that transfered the semantic relations of remaining types across wordnets of different languages. This procedure relied on a triangulation technique that resorts to a pre-existing ontology for a language other than Portuguese — the Princeton WordNet for English — and to a machine-readable bilingual Portuguese-English dictionary.

This setup provided for a quite "controlled" ontology construction undertaking. First, the extraction procedure was applied to a highly structured source of linguistic information and thus offered a good prospect of getting at an

outcome of non-negligible quality. Second, the design of the output ontology aimed at (viz. wordnet) is well understood and the methodology for its verification relies on replicable and well documented lexicographic procedures (Vossen, 1996). Secondly, in view of obtaining the best level of performance in NLP applications and tasks where the extracted ontology is to be integrated, the extracted ontology of Portuguese was subsequently verified manually and corrected in an exhaustive fashion by experts on linguistics.

On the one hand, this exercise provided a good opportunity to assess the added value of the ontology extraction method adopted and to give it an objective evaluation. On the other hand, as the final Portuguese ontology evolved from a projection supported by a pre-existing ontology for English, this extraction exercise provided also a ground to gain insight on the similarity of two lexical semantics-based ontologies, for the same semantic domains, but with terms from two different languages, Portuguese and English. Finally, it provided also the chance to assess the quality of the pre-existing English ontology itself, as every wrong projection of concepts or semantic relations was virtually detected, including those resulting from a possible initial mistake in that pre-existing ontology of English.

In the next Section 2 we present related work, including a brief description of other wordnets for Portuguese. In Section 3 we introduce the methodology and the language resources used to extract the preliminary ontology for Portuguese. In Section 4, the subsequent manual verification phase is reported. In Sections 5 and 6, we report on the alignment and transfer of information between the ontology of Portuguese obtained and the pre-existing ontology for English used to project its first draft. This paper finishes with Sections 7 and 8 where information about distribution and concluding remarks are presented.

2. Related Work

There have been two major families of approaches available to develop a wordnet for a given language. In one of them, each synset and the semantic structure are defined manually from scratch, and related to other synsets, without any consideration to other wordnets. In the other approach, the Princeton WordNet has been used as one of the corners of the triangulation methodology by means of which a first draft of another ontology for another language is projected. Following the terminology in (Vossen, 1996), the first is known as the *merge model* and the latter is termed as the *expand model*.

The first approach was largely followed in the EuroWordNet project (Vossen, 1998). This was a project that aimed at the construction of a multilingual wordnet covering several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). Each language-specific wordnet is structured in the same way as the Princeton WordNet but was manually built, in a first step in isolation from the content of other wordnets, from available existing resources and lexical databases. In addition, the resulting wordnets were linked to an Inter-Lingual-Index (ILI) composed of non structured concepts. Via this ILI, the lexical semantics ontologies were then interconnected so that it is possible to go from the words in one language to semantically equivalent words in any other language.

The second approach, the *expand model*, was used in the MultiWordNet (MWN) project (Pianta et al., 2002). This project was initially aimed at building a wordnet for Italian, but it grew to include wordnets from several other languages, such as Spanish, Romanian, Hebrew, Latin a.o.

This approach allows for automatic procedures that support the speeding up of the definition of synsets and semantic relations between them, as well as the detection of divergences between Princeton WordNet and the wordnet being built. This approach makes use of the Princeton WordNet and of machine readable bilingual dictionaries between English and the target language in a triangulation scheme, by means of which a draft ontology for the target language is projected. A key feature of this approach is that it offers a development discipline that fosters the building of aligned wordnets for different languages.

This *expand model* has been said to present potential drawbacks, the main problem being the risk of losing the lexical and structural specificities of a language when starting from a first draft that is close to the Princeton WordNet. The contrasting assumption underlying this model, however, is that two different languages share a similar conceptual structure, that is the correspondence between two concepts is the rule and not the exception. Furthermore, eventual biasing can be repaired in the subsequent phase of manual validation of the wordnet projected.

On the other hand, the alternative *merge model* is also not without potential problems. First, the process of building a wordnet manually starting from scratch requires a huge amount of effort, which only in rare institutional circumstances can be fully supported. Second, the construction of a lexical database implies numerous subjective decisions that create differences between wordnets that may not reflect the real structure of the languages addressed. Finally, connected to the previous point, the more two wordnets differ in their structure the greater the work overhead needed to subsequently connect them.

2.1. Wordnets for Portuguese

There have been a few efforts to create wordnet-style ontologies for Portuguese reported in the literature.

OpenWordNet-PT (De Paiva et al., 2012) is a wordnet for American Portuguese (Brazil), built through a cross-lingual projection using translation dictionaries from the English Princeton WordNet. It contains close to 118,000 synsets, 82,000 of which contain nouns. However, apart from some top-level base concepts, the projected synsets have not been manually validated.

Onto.PT (Gonçalo Oliveira and Gomes, 2014) was automatically created from dictionaries, thesauri and corpora through the use of several information extraction techniques, such as pattern matching to get relations between words and clustering to group words into synsets. It contains about 109,000 synsets, but it has no cross-lingual linking to wordnets of other languages and validation was performed only for a small sample of the relations.

Both these wordnets are very large, by virtue of the automatic processes used to gather their contents, but on the

flip side they lack the thorough validation only granted by a manual process.
Another wordnet for Portuguese, WordNet.PT (Marrafa et al., 2006), presents opposite characteristics. Developed at CLUL, the Center for Linguistics of the University of Lisbon, it has been manually built, but is reported to contain only 19,000 expressions. Besides this, this wordnet is not distributed for reuse.
To the best of our knowledge (Branco et al., 2019; De Paiva et al., 2016), MWN.PT differs from other wordnets for Portuguese in that it is the only high quality, manually validated and cross-lingually integrated, large-scale wordnet available for this language.

3. Projection

The wordnet being presented in this paper has been built according to the *expand model* along two major devlopment campaigns.
A first construction phase took place in the context of the MultiWordNet project (Pianta et al., 2002), leading to an initial, small sized version that is part of the collection of wordnets that resulted from that project. That inaugural version contained only a few thousand validated synsets under a relatively narrow sub-ontology meant to support a Question Answering application.
In a second phase, we undertook a large effort of validating most of the remaining ontology, which resulted in the wordnet being presented here, now featuring over 41,000 verified synsets of nouns.
In the remainder of this section, the construction of the wordnet is presented in detail.

3.1. Extraction by triangulation

The methodology to extract ontologies via triangulation resorts to two basic, automatic procedures (Pianta et al., 2002). The first procedure is called the ASSIGN-procedure. It uses as primary resources the Princeton WordNet of English and a machine readable bilingual dictionary between English and the target language. On the basis of the translational equivalent provided in the bilingual dictionary, for each Portuguese word sense in this dictionary, this procedure selects a weighted list of the most likely corresponding Princeton WordNet synsets. The algorithm performs this by searching for all the synsets containing the translation proposed in the dictionary. In this way word forms that may correspond to the same English synset are tentatively grouped in a same Portuguese synset, and globally a set of Portuguese candidate synsets are then gathered.
In a second phase the algorithm ranks the elements of this set using a number of heuristics, such as generic probability, back translation, etc.[1] A threshold is defined in order to include in the final list only those projected synsets that score above a minimum level of confidence, and the remaining synsets are partitioned into three subsets and marked accordingly if the confidence score is at a high, medium or low level.
The weighted list obtained is presented to lexicographers. This list is usually composed by a restricted number of synsets that include the correct one and the task of the lexicographer is to select the correct synset. In a few cases, the lexicographer may have to add extra ones or amend (remove or add word forms) those proposed by this procedure.
The second procedure is called the LG-procedure, and is aimed at detecting *lexical gaps*. Lexical gaps arise when a lexicalized concept of a language has no correspondence to any lexicalized concept in the other language. In such case, the target language wordnet is allowed to diverge from the Princeton WordNet. For this procedure, and besides the bilingual dictionary and the English wordnet, other lexical resources are used, such as monolingual dictionaries that explicitly list idioms and restricted collocations.
The main goal here is to provide lexicographers with a list of likely lexical gaps, so that they prioritize their handling.

3.2. Linguistic resources

The method of extraction by triangulation requires a reference wordnet and a bilingual dictionary. As for the bilingual dictionary, we used a machine readable version of the third edition of the Collins Portuguese Dictionary, licensed from HarperCollins Publishers Ltd. This is a medium size lexicon with 23,257 headwords in the Portuguese section, and 25,746 in the English one. Overall, the dictionary contained the definition of 45,950 different senses in the Portuguese section, and 48,777 senses in the English one.
The dictionary was parsed to isolate each sense. In order to take advantage of further information in the dictionary, a mapping between glosses that indicate domain labels and the WordNet domains (Magnini and Cavaglia, 2000) was carried out. Additionally, we established a mapping between the morphosyntactic category tags of the dictionary and the corresponding tags used in the Princeton WordNet. We ended up with a table for each section of the dictionary, where each row describes one of the possible senses of a word and how this sense is translated into English. These tables were then used as input for the ASSIGN-procedure.
As a first result, it turned out that 32,083 word senses happened to have at least one translation equivalent in the English wordnet. This represents the upper bound, determined by the size of the input bilingual dictionary for the coverage of the ASSIGN-procedure algorithm.

3.3. Assessing the automatic extraction

In this section we provide quantitative data about the input and output data of the ASSIGN-procedure.

3.3.1. Statistics about dictionary entries

As mentioned in Section 3.2, the Portuguese section of the Collins dictionary includes 23,257 headwords, and 45,014 word senses. The number of different words senses actually given as input to the ASSIGN-procedure is slightly larger (45,950) as some of the headwords, containing packed representations of headword variants, have been unpacked. For instance, the headword "a cavalo/pé" was unfolded in two entries "a cavalo" (on horseback) and "a pé" (on foot). Around 40% of the available word senses included some gloss (e.g. "macerar": *(fig mortificar)* to mortify). The number of glosses is relevant for the success of the ASSIGN-procedure as glosses are exploited by various word-to-synset assignment rules. A class of glosses which turns

[1] For a complete description see (Pianta et al., 2002).

out to be particularly useful for good assignments are subject domain glosses (e.g. cabecear: *(football)* to head). However only 4.1% of all word senses included this specific kind of gloss.

3.3.2. Statistics about assignment

The number of word senses in the Portuguese section having at least one translation equivalent included in some English synset is 70.1% of the total. This represents the upper bound for the coverage of the ASSIGN-procedure.

If one considers only word senses with a translation equivalent in WordNet, only 66.4% of them got assigned to some WordNet synset, which gives the actual coverage of the ASSIGN-procedure with reference to the upper bound mentioned above. Note that the same word sense can be assigned to more than one synset, although we expect that only one of such assignments be right, that is we expect a one-to-one correspondence between dictionary word senses and WordNet synsets in the vast majority of cases.

In practice, lexicographers will check only assignments with a confidence score higher than a minimal threshold. Such threshold is fixed at 40% of the confidence score which we consider acceptable for a totally automatic assignment (e.g. in case manual check was not an available option). The minimum threshold is fixed so to balance two conflicting needs, which are on the one hand the need to maximize the probability that one of the assignments is right, and on the other hand to reduce as much as possible the number of assignments the lexicographer needs to check.

The number of assignments with confidence higher than the minimum threshold turns out to be 44,618, which corresponds to the entries that lexicographers actually need to manually check. The number of Portuguese lemmas (word/pos) with at least one minimal assignment is 21,319, and the number of WordNet synsets to which at least one word sense has been assigned with minimal confidence is 24,085.

The last two numbers give the size of the draft Portuguese WordNet produced by the ASSIGN-procedure, before the start of manual checking.

4. Validation

After the extraction procedure described above was run, we get a draft Portuguese wordnet projected from Princeton WordNet. In order to meet quality standards that render it a reliable resource, manual verification and completion of the data projected by experts with a degree in Linguistics was subsequently undertaken.

4.1. Resources for manual validation

Each synset contains all synonyms (word forms) conveying a specific concept. In order to proceed with the manual validation, every word has to be checked.

The optimal situation is the one in which the ASSIGN-procedure has proposed one or more Portuguese word forms for each synset, possibly with varying degrees of confidence. If a given word form is misplaced in a certain sysnset, it is rejected. New words may be introduced, either to complete an existing projected synset or to give a certain synset at least one word form.

Hence, synsets which have not been provided by the automatic procedure with any Portuguese word form undergo a similar procedure, and the lexicographer is expected to add an entry for each appropriate synonym. At every synset, the goal is to come up with as many synonyms as possible.

Besides direct validation (by accepting or discarding word forms suggested for a given sysnset), the manual validation task involves also a considerable amount of translation for those synsets in the Princeton WordNet that happened to receive no projection into the draft Portuguese wordnet. Here, the typical procedure involves consulting bilingual dictionaries and then re-checking the translation on reference Portuguese dictionaries, such as Houaiss (Houaiss and Villar, 2001) and Dicionário da Academia (Academia das Ciências de Lisboa, 2001).

For word forms belonging to some terminological field, the Eur-Lex[2] has proven rather useful. This is a freely accessible online database containing documents on European Law, such as legislation, treaties, case-law and legislative proposals, which can be displayed in parallel, with a Portuguese document in one side and the corresponding English translation in the other side.

When it becomes convenient to assess the frequency of a given word form, search engines are used, mainly Google, and special attention is always paid to the credibility of the source documents.

4.2. Validation methodology

In order to proceed with the manual validation, a strategy was followed where different areas of the ontology were delimited to be worked on.

In a first stage, we started by revising the synsets that correspond to the top levels as they appear in the Princeton WordNet. This permitted to address the more generic concepts and arguably the most universal ones, thus ensuring a good basis for the validation of the remainder of the ontology. Also at this stage we targeted the Base Concepts as defined by the Global Wordnet Association,[3] which are "supposed to be the concepts that play the most important role in the various wordnets of different languages" on the basis of their position in the semantic hierarchy and a high number of relations to other concepts; and we also validated the synsets that were selected as the Core Concepts by the EuroWordnet, given that they are considered to be shared and of paramount importance in at least three different languages.

After checking the top ontology formed by all the concepts just mentioned, in the first construction phase, we moved on to validating sub-ontologies that were motivated by work on a specific application, namely a Question Answering system that, as part of its functioning, needs to determine the expected semantic type of the answer in order to select possible answers using a named entity recognizer.

For instance, the analysis of a question like *Which painter...?* should be able to anticipate that possible answers are of the semantic type PERSON by taking the synset

[2] https://eur-lex.europa.eu

[3] http://globalwordnet.org

Table 1: Lexical items per synset

items	#	%
1	42,712	84.69
2	5,344	10.60
3	1,479	2.93
4	504	1.00
5 or more	395	0.78

for the word "painter" and traversing through its hypernyms and reaching the synset for the "person" concept.

Accordingly, the sub-ontologies that were addressed at this stage of validation were those below the concepts of PERSON, ORGANIZATION, LOCATION, EVENT and WORK, which correspond to the classes that can be assigned by the named entity recognizer. This application-focused validation allowed us to quickly ramp-up a validated wordnet for supporting this QA system.

The bulk of the remaining validation work was undertaken in a second construction phase. For this, we adopted an exhaustive validation strategy, whereby the lexicographers go through the list of unchecked synsets, checking them one by one.

To avoid a recursive chaining of checks that would be easy to lose track of, this validation is carefully controlled. Given a yet unchecked synset, the lexicographer validates only (i) the senses in that synset, (ii) the chain of hypernym relations to the first already checked synset, (iii) the relations to all its immediate hyponyms, and (iv) its linking to the corresponding synset in Princeton WordNet.

This keeps validation focused on the synset to be checked instead of having the lexicographer follow recursive chains of checks (e.g. checking the hyponym of the hyponym), but does not preclude the thorough checking of the ontology.

This final validation work was greatly helped through the use of the WordNetLoom (Piasecki et al., 2013) tool, which provides a graphical interface, shown in Figure 1, with which the ontology can be edited.

The final validated wordnet comprises 41,240 synsets. The majority of these contain a single lexical item (cf. Table 1), for an overall number of 1.23 lexical items per synset.

5. Cross-lingual Alignment

5.1. Cross-lingual gaps

When building wordnets for different languages that are to be aligned among them, differences between languages are expected to emerge. Among the possible type of discrepancies, the so-called *lexical gap*s is a most significant one (Bentivogli and Pianta, 2000; Bentivogli and Pianta, 2000): a lexical gap exists in a certain language when another language expresses a concept with a lexical unit whereas the language at stake can only express the same concept with a compositional combination of words.

In our validation work on the Portuguese wordnet, a ENgap relates to the absence of a lexical unit to express in English a certain concept (that can be expressed in Portuguese with a lexical unit). A PTgap, in turn, is a non-lexicalized concept in Portuguese (that happens to be lexicalized in English). For instance, the English word *advisee* expresses a concept that is not lexicalized in Portuguese and a PTgap is to be included in the Portuguese wordnet. In turn, the Portuguese *maternidade* expresses the concept of a hospital specific for birthing for which we found no lexicalization in English. ENgaps and PTgaps are thus nodes at which no connection is established among the Portuguese and the English wordnets.

Naturally, the emergence of these concept gaps are more likely to emerge not at the top-levels of the ontology but at the levels concerned with more specific concepts. Also, different sub-ontologies of interest may display different rates of concept gaps, either in the target or in the source language.

The validated wordnet contains 9,194 gap synsets. Note that this is in addition to the 41,240 non-gap synsets, yielding a total of 50,434 non-gap and gap synsets.

5.2. Divergences

Because the validation work requires looking carefully also at the source English data, that is the Princeton WordNet, we have come across an additional lexicographic challenge, namely concerning what shall be done when the English synsets do not meet the established lexicographic criteria for their inclusion in the wordnet.

Instead of ignoring such cases or merely translating every synset into Portuguese, we have created a set of *Lexical Divergences*, which we use in order to signal those synsets where we think the English data could be improved. Also in these cases, no connection is established among the Portuguese and the English wordnets. Below we describe the types of Lexical Divergences we have came across.

Incorrect lexicalization. Legitimate linguistic doubts often arise as to whether a given noun included in the Princeton WordNet can be considered a proper lexicalization in English or not. This applies mostly to multi-word expressions, which tend to be semantically compositional in many instances. Nevertheless, if there is an equivalent compositional expression in Portuguese with significant frequency, we have chosen to translate it, regardless of its probable lexical inadequacy. Examples include {bottle collection}, {large person} and {small person}.

Incorrect lexical category. Dictionaries tend to be inaccurate or to embrace conflicting linguistic approaches regarding the criteria to distinguish nouns and adjectives. In this respect, we opted for a conservative approach, since a wordnet is meant to be a lexical resource and not to represent productive syntactic behaviour. In order to establish reliable criteria to classify an adjective word also as a noun a sequence of three tests is applied:

1. Aristotelic definition: x is a noun if it is semantically acceptable in definitory formulas like "x is a type of y which z", in which x is the term in question, y is an hyperonym of that term, and z is the specific characteristic that distinguishes x from its co-hyponyms. For instance: "a tram (x) is a vehicle (y) which travels on rails (z)".

2. Absence of anaphoric antecedent: Typically adjectives are acceptable in structures such as "this is x" in

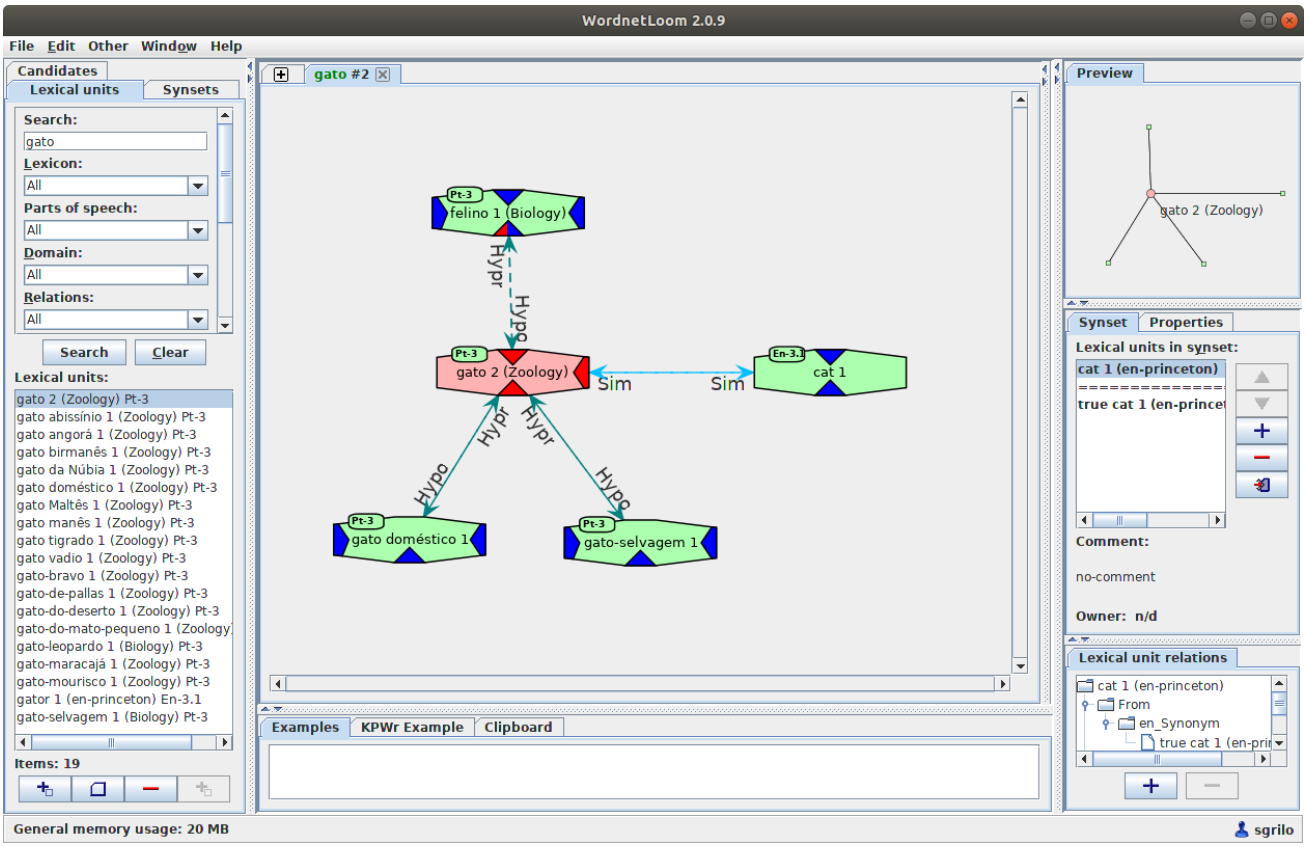

Figure 1: The WordNetLoom graphical interface used in the construction of MWN.PT. The currently selected synset, {gato} (in red), is shown with its immediate hypernym (Hypr) and hyponyms (Hypo) synsets (the chain could be further opened by clicking on the blue triangles), and with the similarity (Sim) relation used to link it to the corresponding synset in Princeton WordNet, {cat}.

instances of anaphoric reference—that is, when they modify an elliptical noun mentioned previously. If x is acceptable in such template with no anaphoric antecedent available, it is considered to be a noun.

3. Adverbial modification: Adjectives can be modified by adverbials but nouns cannot.

If after these three tests a Portuguese lexical unit is classified as a noun and the translation equivalent which it is aligned to occurs in Princeton WordNet is classified as an adjective, we mark the corresponding synset with this type of Lexical Divergence.
Some examples we found are {baffled}, {brave}, {free, free people}, {retreated} and {immune}.

Slashed collocations. Synsets that fall under this case express in their definitory gloss a meaning that holds of a collocation although the nouns in that synset contain only part of the collocation. For instance, the collocation (i) may require a preposition which is missing (e.g. {footing; terms} standing for "on a friendly footing" and "on good terms", {ready} for "at the ready" as in "their guns were at the ready", or {play} for "in play" as in "the ball was still in play"); (ii) may refer to an incomplete idiom (e.g. {pedestal} standing for the whole meaning of "putting someone on a pedestal" or {force} for "to join forces with someone/something"}); or (iii) may simply be an NP with one or more constituents missing (e.g. {emergency} for "state of emergency", {shebang} for "the whole shebang", or {blood} for "young blood"). Sometimes the actual collocation is the only example provided in the gloss, as in the {blood} example: "we need more young blood in this organization".

Incorrect top concept. Synsets under this type of divergence express hyperonymy/hyponymy relations incorrectly defined as far as the top level of the relevant sub-ontology is concerned. For instance, {cable railway, funicular, funicular railway} being under {group, grouping}.

Incorrect immediate hypernym. This Lexical Divergence holds when hyperonymy/hyponymy relations are incorrectly set as far as the immediate level of the sub-ontology is concerned. Instances cover a whole range of different situations, but usually the divergence occurs because a given synset contains lexical units which are in fact synonyms of what is being proposed as its hyperonym (e.g. {order} as hyponym of {command; bid; bidding; dictation} or {disease} as hyponym of {illness, malady, sickness}). Sometimes the divergence is somehow acknowledged in the gloss. For instance, {building society} is defined as the *"British equivalent of US savings and loan association"*, and has as direct hyperonym {savings and loan, savings and loan association}; {rest, eternal rest, sleep, eternal sleep, quietus} are rightly considered *"euphemisms of death"*, though {death} is its immediate hyperonym; and

{rail} is defined as *"a short for railway"*, while being set as an hyponym of {railway, railroad, railroad line, railway line, railway system}.

Unwarranted semantic restriction on arguments. The synsets that fall in this type of divergence are only acceptable as hyponyms of the given hyperonym if semantic restrictions are applied to one of their arguments. Such is the case of {return, paying back, getting even}, which can only be considered a type of {group action} if one of its arguments, namely the agent, is plural and therefore denotes a group instead of an individual.

Prolific use of plurals. Synsets in this type contain nouns which are only acceptable in their current place in the ontology if inflected in the plural. Notwithstanding the surface forms of those word being the same for plural and singular in English — unlike in Portuguese, where lemmatization would highlight this anomaly — the gloss clearly points out to a plural sense. These are mostly hyponyms of the synset {people} under the top synset {group, grouping} which refer to singular nouns that are not collective and therefore should not be considered a type of {group, grouping}. Instances include {blind}, {timid, cautious}, {damned}, {dead}, {living}, {deaf}, {disabled}, {initiate, enlightened}, {uninitiate}, {mentally retarded}, {sick}, {wounded, maimed}.

The nodes we marked as lexical divergences are not included in the version of MWN.PT distributed. Nevertheless, as they may be interesting for further research, information on them is available on demand from our group.

6. Transfer

With the projection of a draft wordnet followed by manual validation, a backbone was obtained that had been reliably verified as for the correctness and completeness of its synsets, the hypernym relations among them, and their alignment with semantically equivalent synsets in the English wordnet.

With this backbone in place, the construction of the wordnet was continued with the transfer of relations among synsets of other semantic types from the English to the Portuguese wordnet.

For every two nodes EN_i and EN_j in the English wordnet such that they are linked to at least one synset in the Portuguese wordnet each, say to PT_i and PT_j respectively, every relation instance of any semantic type between EN_i and EN_j is transferred to PT_i and PT_j, preserving the direction of the relation where relevant.

7. Distribution

The MWN.PT WordNet of Portuguese is distributed for free by PORTULAN CLARIN, the Research Infrastructure for the Science and Technology of Language,[4] in RDF and Princeton formats, under the license CC-BY-ND.

Additionally, the content of MWN.PT can be perused with a wordnet browser presented in (Branco et al., 2018) that can be found also at PORTULAN CLARIN[5] via the GUI shown in Figure 2.

[4] `https://portulanclarin.net`

[5] https://portulanclarin.net/workbench/lx/wnbrowser/

8. Conclusions and Future Work

In this paper we have presented the MWN.PT WordNet for Portuguese and discussed the methodology used for and the lessons learned with its construction.

To the best of our knowledge, MWN.PT is the largest high quality, manually validated and cross-lingually integrated, wordnet that is available for Portuguese, featuring 41,000 concepts expressed by a vocabulary of 38,000 expressions. Its construction was initiated in our group more than one decade ago and its description was published for the first time in the present paper.

The validation work so far has focused on nouns. Besides keeping enlarging the volume of this language resource, future work will broaden its scope to also include the validation of other categories.

Acknowledgements

The research reported here was partially supported by PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT - Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

We thanks Maciej Piasecki for WordNetLoom, and Tomasz Naskret for his help running this tool.

Bibliographical References

Academia das Ciências de Lisboa. (2001). *Dicionário da Língua Portuguesa Contemporânea*. Academia de Ciências de Lisboa e Editorial Verbo.

Bentivogli, L. and Pianta, E. (2000). Looking for lexical gaps. In *Proceedings of the 9th EURALEX International Congress*, pages 8–12.

Branco, A., Branco, R., Saedi, C., and Silva, J. (2018). Coping with lexical gaps when building aligned multilingual wordnets. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, pages 4562–4569.

Branco, R., Rodrigues, J. A., Saedi, C., and Branco, A. (2019). Assessing wordnets with wordnet embeddings. In *Proceedings of the 10th Global Wordnet Conference (GWC)*.

De Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An open Brazilian WordNet for reasoning. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING): Demonstration Papers*, pages 353–360.

De Paiva, V., Real, L., Gonçalo Oliveira, H., Rademaker, A., Freitas, C., and Simões, A. (2016). An overview of Portuguese wordnets. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 27–30.

Gonçalo Oliveira, H. and Gomes, P. (2014). ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation Journal*, 48(2):373–393.

Houaiss, A. and Villar, M. d. S. (2001). *Dicionário Houaiss da Língua Portuguesa*. Temas e Debates.

Magnini, B. and Cavaglia, G. (2000). Integrating subject field codes into WordNet. In *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC)*, pages 1413–1418.

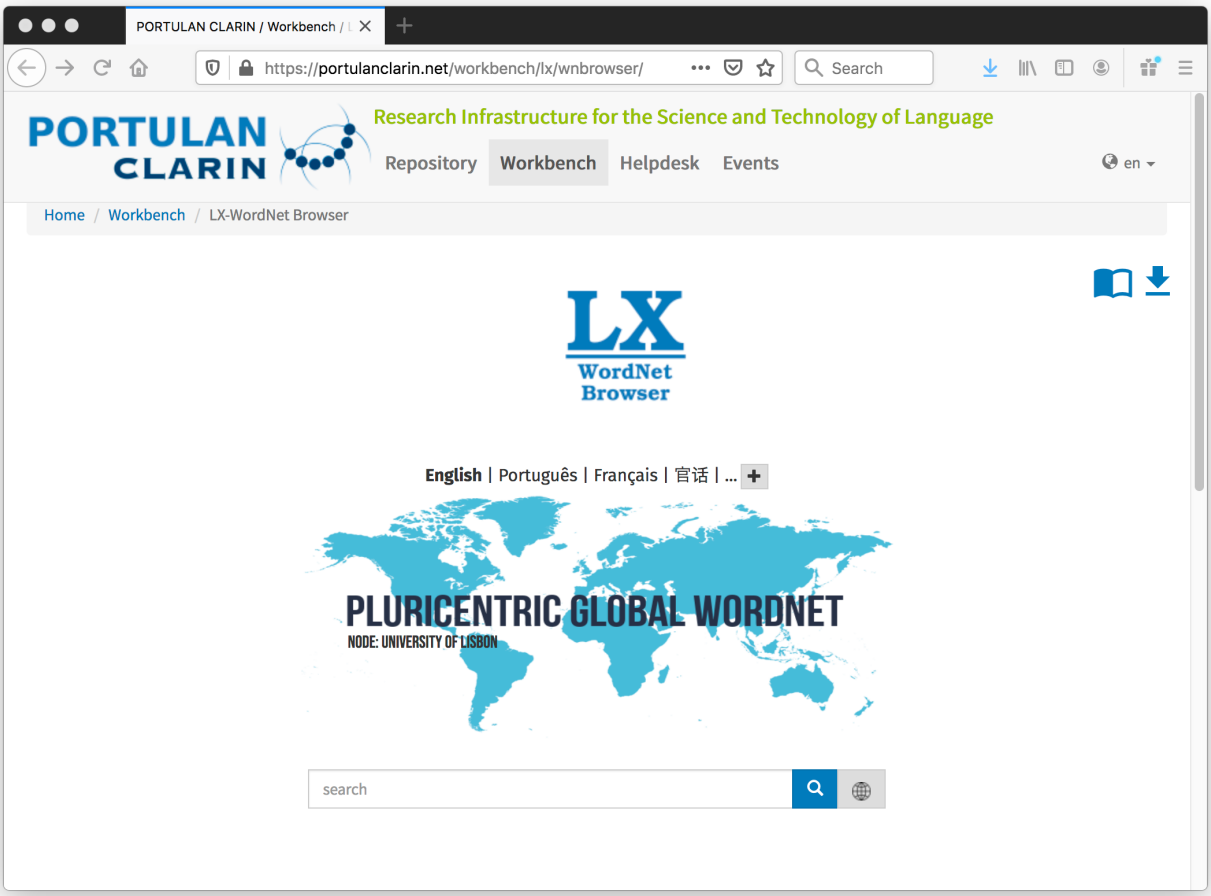

Figure 2: Screenshot of the online browser GUI that allows to peruse MWN.PT

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C., and Mendes, S. (2006). WordNet.PT new directions. In *Proceedings of the 3rd Global Wordnet Conference*, pages 319–320.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the 1st International WordNet Conference*, pages 293–302.

Piasecki, M., Marcińczuk, M., Ramocki, R., and Maziarz, M. (2013). WordNetLoom: a WordNet development system integrating form-based and graph-based perspectives. *International Journal of Data Mining, Modelling and Management*, 5(3):210–232.

Vossen, P. (1996). Right or wrong, combining lexical resources in the EuroWordNet project. In *Proceedings of the Euralex-96 International Congress*.

Vossen, P. (1998). EuroWordNet: Linguistic ontologies in a multilingual database. *Communication and Cognition for Artificial Intelligence*.