Making Virtue of Necessity: a Verb Lexicon

Valeria de Paiva, Fabricio Chalub, Livy Real, and Alexandre Rademaker

[1] Nuance Communications, USA
[2] IBM Research, Brazil
[3] IBM Research, Brazil
[4] IBM Research and FGV/EMAp, Brazil
valeria.depaiva@nuance.com {fchalub,livym,alexrad}@br.ibm.com

Abstract. We describe the verb lexicon of OpenWordNet-PT, a wordnet-like resource for (mostly Brazilian) Portuguese and a series of experiments that we designed to extend its coverage. These experiments include checking online lists of most common verbs, checking corpora freely available such as the Bosque-UD (the Bosque corpus annotated with Universal Dependencies) and especially checking a dictionary of Brazilian politicians' biographies (the DHBB) that we consider an ideal corpus for the kind of information extraction we are after. We certainly succeeded into extending the coverage of the verb lexicon, however it remains to be seen whether this new coverage is enough for the original application.

1 Introduction

Verbs, together with nouns, are usually the main bearers of meaning in sentences. We could not agree more with [8] when they say

> Verbs are the primary vehicle for describing events and expressing relations between entities. Hence, verb semantics could help in many natural language processing (NLP) tasks that deal with events or relations between entities. For tasks which require canonicalization of natural language statements or derivation of (plausible) inferences from such statements, a particularly valuable resource is one which (i) relates verbs to one another and (ii) provides broad coverage of the verbs in the target language.

Portuguese is the 6th most spoken language in the world, according to Etnologue [11], but lexical resources for Portuguese are still not very well-developed. Despite some recent work on Portuguese verbs, such as VerbNet.BR [20, 19], Viper [3], and the catalog of Brazilian Portuguese Verbs [6], there are still no freely available, comprehensive resources that provide human users and automated programs with access to Portuguese verbs, their meanings and information about their subcategorization frames.

Given the essential role played by verbs in sentence understanding we decided to improve the state of the verb lexicon in the basic resource OpenWordNet-PT [14]. OpenWordnet-PT already provides some of the functionality desired, as it has 5902 verbal synsets in Portuguese, with as many as 4511 verbal lemmas. It also has 7865 synsets in English that are empty in Portuguese and for many of these we know there are Portuguese words that fit them perfectly, but they are not there, yet[5]. An example is the verb *popularize*: the verb *popularizar* exists in Portuguese with the same sense as popularize has in English. We only need to add it to the appropriate synset, but our problem is to find out within the 'soup' of these 7865 empty synsets, which ones are easy cases, where a corresponding verb exists with the same meaning in English, like this one. We also need to find out which ones are the hard verbs to translate or even the impossible ones. For an example of a Portuguese verb that is impossible to translate as a single word in English, we are using *apaulistar* (to make similar to what the natives of São Paulo, *paulistas*, do). We do not expect English to have such a word, nor many others like this that correspond to particular facets of Brazilian reality, but these should also be part of a truly useful Portuguese (verb) lexicon.

For the verbs OpenWordnet-PT already knows about, we can provide some indication of meaning, by giving other words that the given verb is related to, as well as its placing in both the OpenMultilingual Wordnet (OMW) [5] and in the SUMO ontology [16]. We can also provide some possible subcategorization frames, inherited from OMW, but not checked for Portuguese, so far. This, as well as making sure that all Portuguese synsets have verified glosses in Portuguese, is left to future work. For the moment we would like to acknowledge the helpful work of Alberto Simões [22] in producing automatically translated glosses, which are extremely helpful for the work described here.

2 OpenWordNet-PT

OpenWordnet-PT is a lexical-semantic resource describing (mostly Brazilian) Portuguese words and their relationships. It is modelled after and fully interoperable with the original Princeton WordNet [9] for English (henceforth PWN), relying on the same identifiers as Princeton WordNet 3.0. This means that one can easily find Portuguese equivalents for some specific English word senses and conversely. This also means that OpenWordnet-PT is part of a large ecosystem of compatible resources, including domain identifiers and mappings to Wikipedia, DBpedia and Wikidata, amongst others. OpenWordnet-PT is encoded and distributed in RDF/OWL [2].

2.1 Related Work

As indicated above there are several works on Portuguese verbs, some more linguistic, some more computational. The more linguistic work seems not to be available online or tends not to have meanings associated to the verbs. The computational work on VerbNet.BR is very encompassing, but it has not been verified for consistency or accuracy. We discuss the golden subset of VerbNet.BR

[5] For up to date numbers check `http://wnpt.brlcloud.com/wn/stats`

in the following section. Moreover, we would not like to take the syntactic classes as so fundamental in our own work. The work on Viper is not open source, at the moment. The work on TeP [12] has unclear licensing status and its definitive version is, apparently, not available yet.

3 Extending the Verb Lexicon

It is always easier to check whether one has coverage of a lexical resource than accuracy of the same, so we decided to check the coverage of our verb lexicon, using resources available online. From this perspective this work is a continuation of [13]. A series of experiments, with different collections of verbs, from corpora and otherwise was devised. We describe some of these experiments here. Data sources used in our experiments can be found in our GitHub repository[6].

3.1 Golden VerbNetBR

Since there is an available VerbNet.BR 1.0, with a manually verified golden subset, we first decided to investigate whether we had all the verbs in this golden subset. Exactly 50 verbs were found to be missing from OpenWordNet-PT from the 604 verbs in the golden subset of VerbNet.BR. These verbs were added to OpenWordNet-PT, in their respective places, with the exception of two verbs *entreabrir, rebolar* (meaning, respectively 'to partially open' and 'to move your hips in a rolling way', both used literally and metaphorically in Brazilian Portuguese) that we did not find perfect placements for.

Adding these verbs was not difficult, but showed us some of the problems and issues we have to deal with. First there are typos and misspellings everywhere. Even the (short) list of golden verbs in VerbNet.BR[7] has a typo *captura*, instead of *capturar* (to capture). Then the different ways of writing in Portugal and Brazil sometimes duplicate entries. For instance the verb *adjectivar* (to add, perhaps too many, adjetives to your sentences), is not really different from *adjetivar*, the Brazilian spelling. This verb apparently does not exist in English (or at least in English as considered by PWN). This orthographic difference is well-known, but there are many entries like this. While there is an official agreement between Portuguese speaking countries that has 'settled' these orthographic differences, it seems absurd to ignore how the language is really written at the moment. Thirdly, as expected, many English verbs 'pack in' an adverb or two, when in Portuguese we only have the basic verb. For example the verb *to jog* is to run slowly or walk fast, hence between *correr* and *andar* in Portuguese, for the fun of it. In Portuguese we have no verb between running and walking, we need the adverbs *slowly, quickly* and we need to indicate that the purpose is fun. But of course this process also happens in the opposite direction and this is much harder for us to ascertain. We have a huge English lexicon

[6] `https://github.com/own-pt/cl-wnbrowser/tree/master/corpora`

[7] To be found at `http://www.nilc.icmc.usp.br/verbnetbr/`.

(117K synsets in PWN) and no guarantees that the humans trying to fit meanings into it, know the whole lexicon. Fourthly, the different kinds of affixes used both in English and Portuguese make some comparisons difficult. In particular a negating prefix, such as *mis-* does not exist in Portuguese, as such, while the Portuguese prefix *auto-* corresponding to doing something to yourself, *self-*, seems much more used in Portuguese than in English. The English PWN lists only one verb with prefix *self*, *self-destroy*, while the Bosque corpus (not a very large one) lists at least four verbs with the corresponding prefix *autodenominar/self-denominate, auto-excluir/self-exclude, autoparodiar/self-parody, autopunir/self-punish* in Portuguese. Finally, one of the main problems has to do with the frequency and popularity of lexical items. We have no reliable frequency data and particularly with two very different vocabularies, corresponding to Portugal and Brazil, for daily things and actions, it is hard to decide on the level of coverage that is required.

As said, we only had problems to fit in two verbs from the golden subset of VerbNet-BR: *entreabrir, rebolar*. The first one *entreabrir*, meaning 'open partially' shows the phenomenon described above: a kind of conceptualization that seems to be done via an adverb in English, as it is a *partial* opening. The second one *rebolar* we can find an approximated sense in the English verb 'to roll'. However, while in English this seems to indicate a particular gait, a way of walking, in Portuguese there is no need to cover any ground while you *ondulate your hips*. This small exercise made us wonder whether there were too many other verbs missing from our resource and we decided to investigate other resources, described below.

To find where to fit in the PWN network the 'missing' Portuguese verbs from the golden VerbNet.BR we established a modus operandi: we translate the desired Portuguese verbs using machine translation and then we manually verify the translation. A list of words in Portuguese and corresponding words in English is then fed to an algorithm that looks for strict matches both of Portuguese and English words, in synsets and in glosses and then suggests these synsets to the human annotators. Finally at least two human annotators have to agree on the appropriateness of the word sense and its placement into the network to make it part of the official resource. The suggesting and voting processes of OpenWordNet-PT are described in [17].

3.2 Basic Coverage

First we used a list of the thousand most common Portuguese verbs as collected by the 'Corpus do Português'[8]. The list in that website actually has 999 verbs instead of a thousand ones and we have all of them in OpenWordNet-PT.

Then we investigated a Swadesh list of the most important Portuguese words [9]. American linguist Morris Swadesh used vocabulary lists to try to understand not only change of languages over time but also the relationships between extant

[8] http://www.corpusdoportugues.org/

[9] `https://en.wiktionary.org/wiki/Appendix:Portuguese_Swadesh_list`

languages. He based his lists on meanings he presumed would be available in as many cultures as possible, so we are using his list here as a basic sanity check. There are several variations of Swadesh lists, and we used the one coming from the archives of the Open Language Archives Community (OLAC) of the University of Pennsylvania. The file in their link to the Project Rosetta[10] was easy to deal with, but seems more about European Portuguese than the list at Wiktionary itself. The whole Swadesh list has 298 items, but many are pronouns and demonstratives that are not part of a traditional wordnet. From this list we found two verbs that we did not have (*fender/'to split', desamolar/'blunt'*), which we added in, but that are not that common in Brazilian Portuguese.

3.3 VerbOcean Translated

A different source of verbs to extend our lexicon was VerbOcean[11]. Work on textual entailment of the traditional kind, using logical forms, could be helped considerably if the algorithms doing the matching of assumptions and conclusions had access to relations of entailment and causation between verbs. One of Princeton's Wordnet's weaknesses is that not many of these relations are recorded in the database. Chklovski and Pantel's work [8] in VerbOcean was meant to address this problem. Given our avowed disposition to do logical reasoning with our representations, as soon as possible, it made sense for us to discuss the collection of verbs in VerbOcean.

Previous work in [13] describes a first attempt to clean up and improve the extant verb lexicon of OWN-PT, using a constructed manual translation of the verbs in VerbOcean. We have checked that all these translated VerbOcean verbs are included in OpenWordnet-PT. Out of the original 2119 verbs in VerbOcean, we already had in OWN-PT 1182 verbs. Now we also have in suggestions 930 verbs. Altogether there were only six verbs still missing: *escantear*, *gazetear*, *prototipar*, *reconfigurar*, *subempregar*, and *desinstalar*. These show that, even if morphologically related, sometimes words can have very different meanings, the so-called *semantic drifting*. While the verb *gazette* in English means *to publish in a gazette*, in Portuguese the verb *gazetear* means *to play truant*. The verbs *prototipar,'to prototype'*, *desinstalar,'to uninstall'*, and *reconfigurar, 'to reconfigure'* seem to arise from technology and hence should exist in English, but in PWN one does not have these verbs. Maybe one should. The verb *subempregar* shows a different social reality. In English one says *underpay* for the practice of paying less than customary to workers, but in Portuguese we prefer to say *subempregar*, or 'under-employ'. Finally the verb *escantear* shows the issues with different national sports as being represented in lexicons. There is a whole collection of verbs in PWN having to do with baseball, American football, golf and basketball that have no direct correspondents in Portuguese (e.g. *to tee* in golf). By contrast in Brazilian Portuguese we have many verbs and especially verbal expressions derived from soccer, the national sport, as *escantear*.

[10] http://dla.library.upenn.edu/dla/olac/record.html?id=rosettaproject_org_rosettaproject_por_swadesh-1

[11] http://demo.patrickpantel.com/demos/verbocean/

3.4 'Bosque' Universal Dependencies

The corpus Bosque [1] is a paradigmatic corpus of Portuguese. It has been used in the CoNLL-X Shared Task in dependency parsing (2006); and very recently it has been converted to Universal Dependencies [21]. The corpus consists of texts in Portuguese (both from Brazil and Portugal) annotated (and analyzed) automatically by the syntactic parser PALAVRAS [4] and reviewed by trained, native speaker linguists. The data comes from news sources. For many reasons, it would be reasonable to expect to have all verbs in this corpus already in OpenWordNet-PT. Nonetheless we found out that a massive number of verbs were not available in OpenWordNet-PT, in any of their senses.

Despite being initially surprised by this finding, we believe that this shows the beginnings of the maturity of OpenWordnet-PT. While subscribing to the view that meaning can be translated from language to language, it seems also clear that different languages will conceptualize different realities, so while an English speaker may not need verbs such as *abrasileirar, aportuguesar, apaulistar, argentinizar, africanizar* (to make or to make more Brazilian, Portuguese, native of São Paulo, Argentinian or African), a Brazilian speaker does need them. These are very easy to explain. Then there are misspellings: despite the fact that the corpus was hand-checked, apparently there was a theoretical decision not to touch the contents of the texts themselves, only hand-correct the processing. Hence we have 'verbs' that are instead typos such as *abanadonar, apretechar, assessoriar, assitir, atinjir*. These are all clearly typos from *abandonar, apetrechar, assessorar, assistir, atingir* (*'to abandon', 'to equip', 'to be a consultant', 'to assist', 'to reach'*), which are all present in the lexicon.

Before removing typos and deciding on what to do about prefixes we have 1981 verbs in Bosque-UD. We had already in OWN-PT 1043 of these. We have managed to add suggestions to 831 synsets. But there are still some 107 missing, which are mostly cases of prefixes and typos. There are six verbs where the prefix *recém* (meaning *recently*) was added to an existing verb, 13 with the prefix *des-*, and 20 typos. A few true Portuguese verbs, arising from nouns (e.g. *vampirizar, 'vampirize'*) and adjectives (*minorar, 'to make minor'*) which do not seem to be conceptualized as verbs in English were added to the list of candidate Portuguese-only synsets, to be dealt with later on.

However, all in all, easy typos and clear-cut cases of a different social reality are rare. Most of the cases of the verbs missing from OpenWordNet-PT seem to be either differences in prefixes used and cases of adjectives and nouns that are made into verbs in Portuguese, but not in English. The prefixes *des-, di-, re-, in-* are used extensively in both Portuguese and English, but they apply to different verbs. For instance, in Portuguese we have *independer, indeterminar* for *to be independent*, and for *not determining* something. These are not treated as verbs in English, or so it seems to us. In Portuguese we use the suffix *-ar* to make verbs out of nouns and adjectives and many in our list of candidates to truly missing Portuguese synsets correspond to these, e.g. *bacharelar, biografar, conveniar, desertificar*, respectively, 'to obtain a Bachelor's degree', 'to write a biography', 'to get a *convênio* (a contract for health insurance or such like) set up' or 'to make a place a desert'.

Of course, deciding that there is no verb in English that expresses exactly the same idea of a Portuguese word is a much harder task then deciding which words in Portuguese fit a given synset. Given this state-of-affairs and the difficult task of deciding which new Portuguese synsets we need to create, we have decided to collect these "possibly Portuguese-only" candidate synsets into a spreadsheet to see if others would be able to find appropriate PWN synsets for these meanings. A comma separated file[12] is available in GitHub with these proposed new synsets. The number of proposed extensions coming from Bosque is not so big, around a hundred, but these still need another checking and devising of principles to add them in.

3.5 Diário Gaúcho

The structured corpus of the Diário Gaúcho (here called DG) is one of the products of PorPopular project[13] that aims to describe and study patterns of written popular Portuguese. Diário Gaúcho is a popular newspaper from the south of Brazil and we have chosen to work with this corpus hoping to find colloquial verbs that were not in OpenWordnet-PT, yet. The DG corpus has approximately 5 millions of tokens and the news were extracted from newspaper issues from 2008. Since OpenWordnet-PT comes from bilingual dictionaries and Wikipedia links, as well as some translated lists, we were worried that we might lack some popular or colloquial verbs.

But our worries were mostly unfounded. Many of the colloquialisms brought in by the DG corpus had synsets that was were good fits. Actually out of all the 2042 verbs in the corpus, 1044 were in OWN-PT and 937 were already in suggestions. Most of the missing 61 verbs are actually typos and processing errors. However, from this work the question of how to deal with different kinds of orthography resurfaced. Differently from some other languages, Portuguese has an official formally approved lexicon that dictates which words are sanctioned as Portuguese words, which ones are not.

There is also the 'new' Portuguese Language Orthographic Agreement, an agreement from 1990, revised several times since then, signed by nine Portuguese speaking countries, which is trying to achieve an unified way to spell Portuguese. The Brazilian Government has announced that the agreement will be fully implemented in 2016.

Mostly the lists of verbs we have been using so far already follow the proposed new Orthographic Agreement. However in the DG corpus, we have found many words that do not follow the new rules, since the corpus was extracted from 2008's news; for example *argüir, 'to argue'*, officially spelled nowadays as *arguir* and *sub-alugar*, officially spelled as *subalugar, 'sublet'.* Since our resource also works as a dictionary for human users, these cases of old spelling rules made us

[12] `http://wnpt.brlcloud.com/wn/prototypes/corpora#candidates`

[13] `http://www.ufrgs.br/textecc/porlexbras/porpopular/`

worry that populating the base with all the ways of spelling the same word is not the best course of action. Those old ways are now considered wrong and might lead users astray. On the other hand, to be used as a tool for the analysis of texts written before the agreement (the largest part of documents in Portuguese), we should, perhaps, have those old forms in the lexicon. We decided to include in our base mainly the forms that follow the agreement, but also to report some old spellings that appear in the chosen corpora. Thus we include e.g. *argüir* and *sub-alugar* in the same synsets that have *arguir* and *subalugar* but do not insist on making it consistently for all variants in the base.

3.6 Verbs from other Sources

One of the applications envisaged for our lexical resources is their use on Digital Humanities studies, in particular to help with information extraction from unstructured text. In our case we work with a small, but very interesting corpus of biographies of historical figures in Brazil, from the 1930's onwards, which we abbreviate as DHBB ('Dicionário Histórico Biográfico Brasileiro, in Portuguese, or "Dictionary of Brazilian Historical Biographies"). What makes this corpus particularly nice for information extraction is that the writers of the entries were asked to follow a set of guidelines with respect to the information that these entries about the historical figures should contain. Hence there is a sense in which the corpus is 'semantically contained'.

For the purposes of our work here it means that we expect to find a relatively limited collection of verbs that would appear, about people being born, marrying, campaigning, being elected, writing laws, approving them and such like. The first attempt at analyzing the results from a shallow processing of this corpus have been described in [15]. We quickly realized that for processing this corpus we needed to deal with *named entities* (NE) and their recognition. Our processing of NEs is not up to the levels expected, yet. However, even processing that is not as precise and accurate as one would like it to be, can be enough to provide useful pointers when it comes to verifying the coverage, precision and recall of your lexical resource. Thus we use the list of verbs of the DHBB, as processed by Freeling [7], as an evaluation measure for the coverage of our system.

From the whole DHBB corpus we choose to check verbs with more than ten occurrences. We still have 51 such verbs missing. Amongst these, there are no typos, but there are some specific items from the politics domain (e.g. the verb *subsecretariar, 'to act as a subsecretary'*) and some oddities that need investigation (e.g verbs *pedrar*, *extremar* and *bondar*). If we add to these, the list of candidate synsets already extracted from Bosque-UD and the other corpora we have some a hundred and fifty verbs that we think deserve new Portuguese synsets. Providing synsets for these these, we do have all verbs in both Bosque-UD and the DHBB corpora covered. It is interesting to note that indeed the DHBB corpus shows some interesting social differences: we have several different verbs in Portuguese for graduating from college *bacharelar, graduar, formar, doutorar, mestrar*, while there is simply *graduate* in PWN. We also have a collection of verbs related to *enrol*, both in schools and in political parties, such as *ingressar, afiliar, matricular, juntar-se*. And we have at least three different ways of expressing the meaning of *separate from your spouse* in Portuguese, with different legal status, *descasar, desquitar, divorciar*, of which only the last one exists as such in PWN.

The work in PropBank-BR [18] says it has identified, automatically, 5688 verbs as candidate members of VerbNet.Br, distributed within 257 classes, inherited from VerbNet. They suggest that a human verification of their results would be highly valuable. We have not been able to verify all of those verbs, which means that there is no resource with semantics for them, at the moment. Hopefully our methods will produce similar results in terms of subcategorization frames, when we get to these and to semantic role labelling.

Finally to see how much work we would still need to do to claim comprehensive coverage, we compared our number of verbs with the ones in the Portal da Língua Portuguesa[14]. Of course we are missing even more verbs, but the Portal is an inclusive resource, covering many variations of Portuguese and not concentrating in high frequency items, like us. José Pedro Ferreira, from the Portal da Língua Portuguesa, was kind enough to extract from their database only verbs mainly used in Brazil and of higher frequency. The construction of the lexical resources in the Portal website are described in [10]. This list has 3918 verbs, of which we have in OWN-PT 1822 verbs, and in suggestions 1290. We still miss 806 verbs, at the moment.

4 What's Next?

Like VerbNet-BR [20], our goal is a domain independent lexicon that provides semantic and syntactic information about Portuguese verbs. Like VerbNet-BR, we would like to have a Levin-style classification of Portuguese verbs, together with a comprehensive listing of the subcategorization frames for these verbs in our lexicon. This is because subcategorization frames provide information about the syntactic realization of verbs as well as diathesis alternations, which can be used as features for machine learning of semantic roles, our eventual goal.

Thus far we are completing our translations from English to Portuguese, using as criteria of evaluation of coverage open source corpora such as Bosque-UD and PropBank-BR, which is also the same corpus Bosque, but under different processing. We would like to bootstrap a comprehensive lexicon of subcategorization frames from both the minimal frames already present in Princeton WordNet and the annotated corpora available.

Princeton WordNet has 13767 verbal synsets. More than half of these synsets have no words in Portuguese. How many of these really constitute synsets that should not exist in a Portuguese wordnet? And how many new synsets do we need to add to have a resource as useful for Portuguese as PWN is for English? Surely we must have coverage to deal with basic news text, such as the ones in Bosque, Diário Gaúcho or the DHBB. But we do not have, as yet, an worked-out measure for accuracy or adequacy of our resource.

[14] `http://www.portaldalinguaportuguesa.org/`

5 Conclusions

This work describes some of our effort to complete and to extend the verb lexicon of OpenWordnet-PT, our Portuguese wordnet. As OpenWordnet-PT is a semi-automatic construction from the translation of Princeton Wordnet 3.0, we do need some strategies to complete the synsets where the automated process was not refined enough. Thus we have used a few different verb lists complete the English synsets with no Portuguese words. Since our work was also based in (recent) corpora, we have found some verbs that should, perhaps, be in Princeton Wordnet and are not, and many truly 'Brazilian verbs', that should be in a Portuguese wordnet, but have no reason to appear in an English one.

There is still plenty of work to do, both in terms of coverage and accuracy of our verb lexicon. But we reckon we now have a much more substantial verb lexicon. We verified that we do have all the verbs in the golden standard subset of VerbNet.BR, except two. We verified that we have all the verbs in the translation of VerbOcean, with the exception of six verbs, that we need to add to OpenWordNet-PT. We verified the list of the thousand most frequent verbs in the Corpus do Português and only four verbs needed to be added. We verified all the verbs in the Bosque-UD corpus and DG corpus, adding the ones that we need to create to a jointly curated list, of approximately a hundred 'missing' items. Similarly we verified all the verbs that occur in the DHBB ten or more times and all of these have been either added to OpenWordNet-PT or added to the list of new candidate synsets. With all of these in place, we believe that the verb lexicon is comprehensive enough for our immediate goals.

We need new ways of making sure that the empty synsets in English are there because there are no good translations. Most importantly, we need to come up with principled ways of extending OpenWordNet-PT in the directions that we are clear that it needs to be extended. Frequency in corpora that we deem relevant is one of the tools to be used. Another idea we are pursuing is using the data from the common orthographic Vocabulary of the Portuguese Language (VOC) [10] to improve OpenWordnet-PT, given that we have very little in terms of morphology and frequency of use information, which they do have. On a different direction, we would like to find ways of verifying the Portuguese glosses. We would like to tackle also the issue of the nominalizations, that VOC has many more than we do. Continuing the theme of incorporating morphology together with semantics, we would like to finish the inclusion of morpho-semantic links in OpenWordnet-PT.

References

1. Afonso, S., Bick, E., Haber, R., Santos, D.: Floresta sintá (c) tica: A treebank for portuguese. In: Proceedings of LREC 2002 (2002)
2. van Assem, M., Gangemi, A., Schreiber, G.: RDF/OWL representation of WordNet. W3c working draft, World Wide Web Consortium (June 2006), http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/

3. Baptista, J.: Viper: A lexicon-grammar of european portuguese verbs. In: 31 e Colloque International sur le Lexique et la Grammaire (2012)
4. Bick, E.: The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework. Ph.D. thesis, Aarhus University (2000)
5. Bond, F., Paik, K.: A survey of wordnets and their licenses. In: Proceedings of the 6th Global WordNet Conference (GWC 2012). Matsue (2012), 64–71
6. Cançado, M., Godoy, L., Amaral, L.: The construction of a catalog of Brazilian Portuguese verbs. In: Empirical Methods in Natural Language Processing: Proceedings of the Conference on Natural Language Processing, ed. Jancsary, J. (2012)
7. Carreras, X., Chao, I., Padró, L., Padró, M.: Freeling: An open-source suite of language analyzers. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04) (2004)
8. Chklovski, T., Pantel, P.: VerbOcean: Mining the web for fine-grained semantic verb relations. In: Proceedings of EMNLP (2004)
9. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
10. Ferreira, J.P., Janssen, M., Correia, M., De Oliveira, G.M.: The common orthographic vocabulary of the portuguese language: a set of open lexical resources for a pluricentric language. In: Proceedings of LREC (2012)
11. Lewis, M. Paul, G.F.S., (eds.), C.D.F.: Ethnologue: Languages of the World, Eighteenth edition. SIL International, Dallas, Texas (2015), `http://www.ethnologue.com`
12. da-Silva; Helio Roberto de Moraes, B.C.D.: A construção de um thesaurus para o português do brasil. Alfa 47(2), 101–115 (2003)
13. de Paiva, V., Freitas, C., Real, L., Rademaker, A.: Improving the verb lexicon of openwordnet-pt. In: Alemany, L.A., Padró, M., Rademaker, A., Villavicencio, A. (eds.) Proceedings of Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish (ToRPorEsp). Biblioteca Digital Brasileira de Computação, UFMG, Brazil, São Carlos, Brazil (oct 2014), `http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Evento?id=755`
14. de Paiva, V., Rademaker, A., de Melo, G.: Openwordnet-pt: An open Brazilian Wordnet for reasoning. In: Proceedings of COLING 2012: Demonstration Papers. pp. 353–360. The COLING 2012 Organizing Committee, Mumbai, India (dec 2012), `http://www.aclweb.org/anthology/C12-3044`, published also as Techreport http://hdl.handle.net/10438/10274
15. Paiva, V.D., Oliveira, D., Higuchi, S., Rademaker, A., Melo, G.D.: Exploratory information extraction from a historical dictionary. In: IEEE 10th International Conference on e-Science (e-Science). vol. 2, pp. 11–18. IEEE (oct 2014)
16. Pease, A.: Ontology: a practical guide. Articulate Software Press (2011)
17. Real, L., Chalub, F., de Paiva, V., Freitas, C., Rademaker, A.: Seeing is correcting: curating lexical resources using social interfaces. In: Proceedings of 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference on Natural Language Processing of Asian Federation of Natural Language Processing - Fourth Workshop on Linked Data in Linguistics: Resources and Applications (LDL 2015). Beijing, China (jul 2015)
18. Scarton, C., Alusio, S.: Towards a cross-linguistic verbnet-style lexicon for brazilian portuguese. In: Workshop on Creating Cross-language Resources for Disconnected Languages and Styles Workshop Programme. p. 11 (2012)

19. Scarton, C., Sun, L., Kipper-Schuler, K., Duran, M.S., Palmer, M., Korhonen, A.: Verb clustering for brazilian portuguese. In: Computational Linguistics and Intelligent Text Processing, pp. 25–39. Springer Berlin Heidelberg (2014)
20. Scarton, C.E.: Verbnet. br: construção semiautomática de um léxico computacional de verbos para o português do brasil. In: 8th Brazilian Symposium in Information and Human Language Technology. pp. 20–29 (2011)
21. Silveira, N., Dozat, T., de Marneffe, M.C., Bowman, S., Connor, M., Bauer, J., Manning, C.D.: A gold standard dependency corpus for English. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014) (2014)
22. Simões, A., Guinovart, X.G.: Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets. In: Advances in Speech and Language Technologies for Iberian Languages, Proceedings of 2nd International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain. LNCS, vol. 8854, pp. 239–248. Springer (2014), http://dx.doi.org/10.1007/978-3-319-13623-3_25