Seeing is Correcting: curating lexical resources using social interfaces

Livy Real
IBM Research
livyreal@gmail.com

Fabricio Chalub
IBM Research
fcbrbr@gmail.com

Valeria de Paiva
Nuance Communications
valeria.depaiva@gmail.com

Claudia Freitas
PUC-RJ
maclaudia.freitas@gmail.com

Alexandre Rademaker
IBM Research and FGV/EMAp
alexrad@br.ibm.com

Abstract

This note describes OpenWordnet-PT, an automatically created, manually curated wordnet for Portuguese and introduces the newly developed web interface we are using to speed up its manual curation. OpenWordNet-PT is part of a collection of wordnets for various languages, jointly described and distributed through the Open Multi-Lingual WordNet and the Global WordNet Association. OpenWordnet-PT has been primarily distributed, from the beginning, as RDF files along with its model description in OWL, and it is freely available for download. We contend the creation of such large, distributed and linkable lexical resources is on the cusp of revolutionizing multilingual language processing to the next truly semantic level. But to get there, there is a need for user interfaces that allow ordinary users and (not only computational) linguists to help in the checking and cleaning up of the quality of the resource. We present our suggestion of one such web interface and describe its features supporting the collaborative curation of the data. This showcases the use and importance of its linked data features, to keep track of information provenance during the whole life-cycle of the RDF resource.

1 Introduction

Lexical knowledge bases are organized repositories of information about words. These resources typically include information about the possible meanings of words, relations between these meanings, definitions and phrases that exemplify their use and maybe some numeric grades of confidence in the information provided. The Princeton wordnet model (Fellbaum, 1998), with English as its target language, is probably the most popular model of a lexical knowledge base. Our main goal is to provide good quality lexical resources for Portuguese, making use, as much as possible, of the effort already spent creating similar resources for English. Thus we are working towards a Portuguese wordnet, based on the Princeton model (de Paiva et al., 2012).

Linguistic resources are very easy to start working on, very hard to improve and extremely difficult to maintain, as the last two tasks do not get the recognition that the first one gets. Given this intrinsic barrier, many well-funded projects, with institutional or commercial backing cannot keep their momentum. Thus it is rather pleasing to see that a project like ours, without any kind of official infra-structure, has been able to continue development and improvement so far, re-inventing its tools and methods, to the extent that it has been chosen by Google Translate to be used as their source of lexical information for Portuguese[1].

This paper reports on a new web interface[2] for consulting, checking and collaborating on the improvement of OpenWordnet-PT. This is the automatically created, but manually verified wordnet for Portuguese, fully compatible and connected to Princeton's paradigmatic WordNet, that we are working on. It has been surprising how a simple interface can make content so much more perspicuous. Thus our title: if seeing is believing, new ways of seeing the data and of slicing it, according to our requirements, are necessary for curating, correcting and improving this data.

Correcting and improving linguistic data is a hard task, as the guidelines for what to aim for are not set in stone nor really known in advance. While the WordNet model has been paradigmatic in modern computational lexicography, this model is not without its failings and shortcomings, as far as specific tasks are concerned. Also it is easy and somewhat satisfying to provide copious quantitative descriptions of numbers of synsets, for different parts-of-speech, of triples associated to these synsets and of intersections with different subsets

[1] http://translate.google.com/about/intl/en_ALL/license.html

[2] http://wnpt.brlcloud.com/wn/

Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015), pages 20–29,
Beijing, China, July 31, 2015.

of Wordnet, etc. However, the whole community dedicated to creating wordnets in other languages, the Global WordNet Association[3], has not come up with criteria for semantic evaluation of these resources nor has it produced, so far, ways of comparing their relative quality or accuracy. Thus qualitative assessment of a new wordnet seems, presently, a matter of judgement and art, more than a commonly agreed practice. Believing that this qualitative assessment is important, and so far rather elusive, we propose in this note that having many eyes over the resource, with the ability to shape it in the directions wanted, is a main advantage. This notion of volunteer curated content, as first and foremost exemplified by Wikipedia, needs adaptation to work for lexical resources. This paper describes one such adaptation.

2 OpenWordnet-PT

The OpenWordnet-PT (Rademaker et al., 2014), abbreviated as OpenWN-PT, is a wordnet originally developed as a syntactic projection of the Universal WordNet (UWN) of de Melo and Weikum (de Melo and Weikum, 2009). Its long term goal is to serve as the main lexicon for a system of natural language processing focused on logical reasoning, based on representation of knowledge, using an ontology, such as SUMO (Pease and Fellbaum, 2010).

OpenWN-PT was built using machine learning techniques to create relations between graphs representing lexical information coming from versions (in multiple languages) of Wikipedia entries and open electronic dictionaries. For details, one can consult (de Melo and Weikum, 2009). Then a projection targeting only the synsets in Portuguese was produced. Despite starting out as a projection only, at the level of the lemmas in Portuguese and their relationships, the OpenWN-PT has been constantly improved through *linguistically motivated* additions and removals, either manually or by making use of large corpora.

The philosophy of OpenWN-PT is to maintain a close connection with Princeton's wordnet since this minimizes the impact of lexicographical decisions on the separation or grouping of senses in a given synset. Such disambiguation decisions are inherently arbitrary (Kilgarriff, 1997), thus the multilingual alignment gives us a pragmatic and practical solution. The solution of following the work in Princeton is practical, as WordNet remains the most used lexical resource in the world. It is also pragmatic, since those decisions will be more useful, if they are similar to what other wordnets say. Of course this does not mean that all decisions will be sorted out for us. As part of our processing is automated and error-prone, we strive to remove the biggest mistakes created by automation, using linguistic skills and tools. In this endeavour we are much helped by the linked data philosophy and implementation, as keeping the alignment between synsets is facilitated by looking at the synsets in several different languages in parallel. For this we make use of the Open Multilingual WordNet's interface (Bond and Foster, 2013).

This lexical enrichment process of OpenWN-PT employs three language strategies: (i) translation; (ii) corpus extraction; (iii) dictionaries. Regarding translations, glossaries and lists produced for other languages, such as English, French and Spanish are used, automatically translated and manually revised. The addition of corpora data contributes words or phrases in common use which may be specific to the Portuguese language (e.g. the verb *frevar*, which means to dance *frevo*, a typical Brazilian dance) or which do not appear via the automatic construction, for some reason. (One can conjecture that perhaps the word is too rare for the automatic methods to pick it up: an example would be the adjective *hidrogenada*, which is in use in every supermarket of Brasil. The verb *hydrogenate* is in the English wordnet, the verb exists exactly as expected in Portuguese *hidrogenar*, but the automatic methods did not find it nor the derived adjective.) The first corpora experiment in OpenWN-PT was the integration of the nominalizations lexicon, the NomLex-PT (Freitas et al., 2014). Use of a corpus, while helpful for specific conceptualizations in the language, brings additional challenges for mapping alignment, since it is expected that there will be expressions for which there is no synset in the English wordnet. Dictionaries were used both for the original creation of Portuguese synsets but also indirectly through the linguists' use of PAPEL (Gonçalo Oliveira et al., 2008) to construct extra pairs of words of the form (verb, nominalization).

3 Current status

The OpenWN-PT currently has 43,925 synsets, of which 32,696 correspond to nouns, 4,675 to verbs,

[3] http://globalwordnet.org/

5,575 to adjectives and 979 to adverbs. While it is not as comprehensive as Princeton's, the Finnish, or the Thai wordnets, it is still more than twice the size of the Russian wordnet, bigger than the Spanish and just a little smaller than the French wordnet. But as discussed in the introduction, the quality of these resources is much harder to compare.

Besides downloading it, the data on Portuguese can be retrieved via a SPARQL endpoint [4]. The multilingual base can be consulted and compared with other wordnets using the Open Multilingual Wordnet (OMWN) interface [5] and changing preferences to the desired languages, assuming the lexical item is found [6].

The ability of comparing senses in several languages was already useful when judging meanings in Portuguese. However, before the new interface was implemented, we did not the ability to compare a word with the collection of other words with the same meaning, or with different shades of meaning, appearing both in English and Portuguese. This all changed, since we started developing a new search and editing interface in September 2014.

4 Challenges of lexical enrichment

We set ourselves the task of building a wordnet for Portuguese, based on the Princeton wordnet model. This is not the same as building the Princeton wordnet in Portuguese. We do not propose to simply translate the original wordnet, but mean to create a wordnet for Portuguese based on Princeton's architecture and, as much as possible, linked to it at the level of the synsets.

The task of building a wordnet in Portuguese imposes many challenges and choices. Even the simple translation of a lexical resource, such as NomLex (Catherine Macleod, 1998) for comparison and further extension of our wordnet, requires different techniques and theoretical decisions. One example might help: the synsets automatically provided by OpenWN-PT tend to have relatively high register words, especially ones with Latin or Greek roots and present in several European languages. Thus we do not get many colloquialisms or everyday words from the translation dictionaries that are the sources for our wordnet. Worse, even when there is more than one possible translation of an English word, there is no way to make sure that the automatic process gets the most used variant in Portuguese. Thus we have to compensate and complete our synsets and many choices are necessary. These lexical choices have direct consequences on the type and quality of the resource been built.

This section discusses some of the problems and issues we have when trying to deal with the mistakes we perceive in OpenWordnet-PT and principally how to deal with senses in wordnet that do not have a clearly corresponding sense in Portuguese.

Our most important decisions so far were related to (i) which variants of Portuguese to treat, (ii) how to deal with mistakes, ungrammaticalities and other problems in our entries, (iii) how to deal with senses in wordnet that apparently do not have a straightforward corresponding sense in Portuguese and (iv) how to add senses in Portuguese that do not seem to exist in English (or at least in the Princeton's version).

We have decided that OpenWN-PT should, in principle, include all variants of Portuguese. First because European Portuguese and Brazilian Portuguese are not that different, then because there is a single Wikipedia/Wiktionary in Portuguese but mostly because it is more complicated to decide which words are used where, than to be inclusive and have all variants. Thus senses that can be expressed through words that have different spellings on different Portuguese dialects (e.g. *gênio, génio*) should include all these variants.

First, to clean up our knowledge base, we still have to remove some Spanish, Galician or Catalan words that are easily misjudged as Portuguese by the automatic processing. We also have to make sure that the part of speech (POS) classification is preserved: many times the popularity driven automatic process prefers the noun meaning of a verb that can be both, or conversely. For example in the noun synset `06688274-n` the automatic processing chose the verb *creditar* 'to credit' instead of the related noun *crédito* 'credit'. We also have several problems with the lemmatization and capitalization of entries, as criteria for the use of capitals are different in English and Portuguese and our entries were not lemmatized beforehand. We follow

[4] `http://logics.emap.fgv.br:10035/repositories/wn30`

[5] `http://compling.hss.ntu.edu.sg/omw`

[6] Since we do not have control over the update processes of OMWN the reader should be aware that OMWN might be not up-to-date with the current OpenWN-PT version.

the Portuguese dictionary traditions and mostly only list the masculine singular form of nouns and adjectives.

Much more complicated than the cleaning up task is the issue (iii) of Princeton wordnet's concepts that do not have a exact, single word correspondent in Portuguese. Several, related problems can be seen here. The original WordNet has many multi-word expressions as part of its synsets. The proverbial *kick the bucket* (and one of its corresponding idiomatic Portuguese expressions *bater as botas* – literally *click the boots*) comes to mind. Thus we do not have a problem with the idea of using expressions, but we do have a problem in deciding which kinds of expressions should be considered in Portuguese. For one example, we do not have a verb corresponding to *to jog* in Portuguese. People use *correr* which means *to run*, but this is not quite right. For one, running is faster than jogging, so jogging is a slow running, and then jogging is for fun or exercise, as WordNet tells us. The Anglicism *fazer jogging* is also very used in Brazil and forces us to think about how 'truly Portuguese' should our entries be.

Another kind of example of English synsets that have no exact Portuguese words at the moment, but it is easier to deal with, is a synset like `08139795-n`, which corresponds to the United States Department of the Treasury. This is a named organization. Should it be in a Portuguese wordnet? Which instances of entities should a wordnet carry? All the names of countries and cities and regions of the world? There are specialized resources like GeoNames that might be much better equipped for that. Which famous people and companies and organizations should a dictionary or a thesaurus have? Again encyclopedic resources like Wikipedia, DBpedia or Wikidata seem much more appropriate. This is what we call amongst ourselves the *A-Box problem*, in a reference to the way Description Logics classify statements. Having translations for all these A-box instances causes us no problem, but not having them is not a big issue either, if we have other sources of information to rely on.

For a third, perhaps harder example, consider the synset `13390244-n`, which uses a specific word (a 'quarter') for the concept of "a United States or Canadian coin worth one fourth of a dollar". We have no reason to have this concept in a Portuguese wordnet and we have no word for it in Portuguese. But we can use a commom expression, such as *moeda de 25 centavos [de dolar]*, for it. Although '25 cents coin', strictly speaking, might not be the same concept as 'quarter'. This will depend on which notion of equality of concepts you are willing to use, a much harder discussion.

For now, for the general problem of what to do with synsets that have no exact corresponding word or synset, we have no clear theoretical decisions or guidelines in place, yet. These problems are still being decided, via a lazy strategy of cleaning up what is clearly wrong first, and collecting subsidies for the more intricate lexicographic decisions later on. Some of these discussions and decisions were described in (de Paiva et al., 2014).

We are not yet working on the Portuguese senses that do not seem to have a corresponding synset in Princeton's wordnet (issue (iv) above). An example might be *jogo de cintura*, which means a property of someone who can easily adapt his/her aims and feelings to a certain situation (the literal meaning is more like "[have] hip moves"). We will add in these new synsets, once we finish the first version of a cleaned up OpenWN-PT that we are completing at the moment. For the time being, we are simply collecting interesting examples of Portuguese words that do not seem to have a direct translation, such as *avacalhar, encafifar*.

But apart from phenomena that have to be dealt with in a uniform way, we have also one-off disambiguation problems, like the verb *to date=datar*, that in Portuguese is only used to put a date (on to a document, a monument or a rock), when in English it also means *to go out with*. Thus the automatic processing ended up with a synset meaning both "finding the age of" and "going out with", `00619183-v`, which is a bad mistake. To see and check this kind of situation, it was decided that the interface would allow linguists to accept or remove a word, a gloss and examples of the use of the synset.

5 The New Interface

The need for an online and searchable version of OpenWN-PT arises for two reasons: (i) to have an accessible tool for end users, (ii) to improve our strategy to correct and improve the resource. As far as being accessible to end users the open source interface, available from GitHub [7] seems

[7] `https://github.com/fcbr/cl-wnbrowser`

a success: after a couple of months online, we have gathered over 4000 suggestions from the web interface, incorporated over 125000 suggestions from automatic processes that are being evaluated, and over 7000 votes have been cast. As for the social/discussion aspect, over 2600 comments have been made on the system and we have registered some sort of conflict in over a hundred suggestions (where we have votes agreeing and disagreeing over the same suggestion). All this being done by a team of five people, where usually two and three are mostly active. More usage statistics are being collected, but it seems clear that it is useful. Considering (ii), our main purpose with the new interface is to edit the entries of OpenWN-PT as they exist. The first design decision was that before adding new synsets corresponding to the Portuguese reality, we should clean up the network from its most egregious mistakes, caused by the automatic processing of the entries.

Figure 1 shows how a synset appears in the new interface. Note the *voting mechanism*, vaguely inspired by Reddit. Trusted users vote for their desired modifications. Here the expression *sair com* has been voted to be removed, three times. There are also links to the same synset in SUMO and OMW.

We encourage the collaborative revision of OpenWN-PT and have been working on guidelines to foster consistency of suggestions. These describe the desired format of examples, glosses and variations of the words in synsets. The preliminary and evolving guidelines for annotators are now available online[8] and we also started documenting the features of the system for end users[9].

But the new interface was much more useful than simply offering the possibility of local rewrites, as it has allowed us to make faceted search for different classes of synsets and of words, both in English and in Portuguese.

Figure 2 shows the synsets that have no words in Portuguese (via facets on the number of words in English and Portuguese), which allows us to target these synsets and to decide whether they are simply missing a not very popular word (e.g. `00117230-v` is missing the not terribly interesting verb *opalizar*, an exact correspondent to *opalize*) or they correspond to a sense that does not work exactly the same way in English and Portuguese. For example, back to the verb *to date* as in *romantically going out with someone*, English seems to leave underspecified whether it is a habitual event or a single one, while in Portuguese we use different verbs, *namorar* or *sair*, but if we want to not commit ourselves to either kind of engagement, we use the verbal expression *sair com*.

Regarding the technologies adopted for development, the interface runs on the IBM BlueMix(blu,) cloud platform implemented in three layers. A Cloudant(clo,) database service for data storage is queried and updated via an API written in NodeJS(nod,). The user interface is coded in Common Lisp using a collection of packages for web development, such as `hunchentoot`, `closure-template`, and `yason`. We have plans to increase the use of Javascript libraries to make the interface more usable, responsive, and mobile-friendly.

Our principal goal in developing the web interface is to provide an application that supports the achievement of consensus in the manual revisions. For this, we follow certain aspects commonly used in social networking websites, such as votes and comments. Contributors can submit suggestions and vote on already submitted suggestions. While anyone can submit any suggestion, in this initial phase only selected users can vote. We currently specify that we need at least three positive votes to accept a suggestion, but two negative votes are enough to reject it. A batch process counts the votes every night and accepts/rejects the suggestions. Finally, another batch process commits the accepted suggestions in the data, removing or adding new information. This modular architecture provides good performance and maintainability. We never delete suggestions, even the rejected ones. This way we keep track of the provenance of all changes in the data.

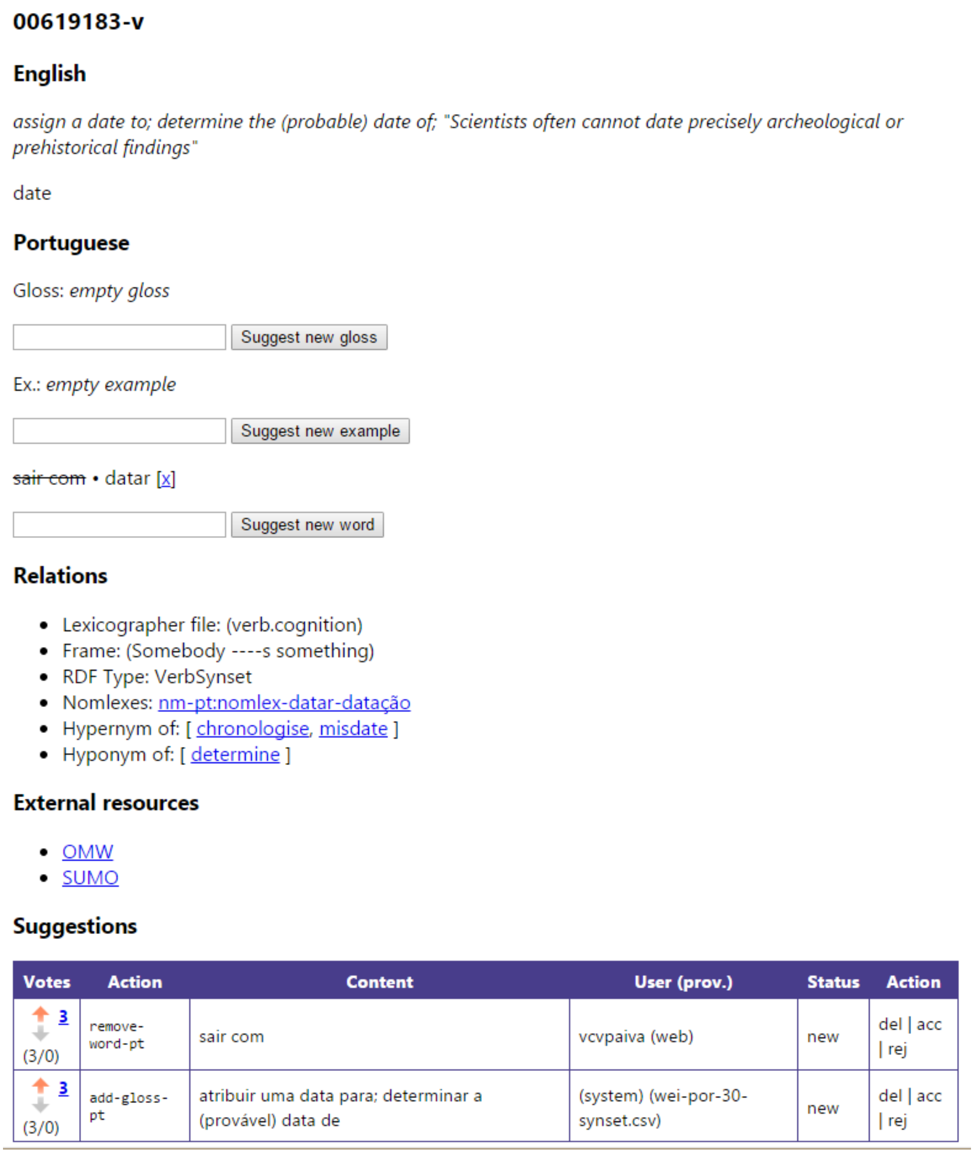
We encourage patterns of communication between users frequently associated with social networks such as Twitter and Reddit where users can mention other users in comments (thus asking for attention on that particular topic). Comments may also contain ‘hash tags’ that are used, for instance, to tag particular synsets for later consideration by other users.

6 Linked Data Rationale

As it is well-known linked data, as proposed by (Berners-Lee, 2011), has four main principles for

[8] `https://goo.gl/tIROu0`

[9] `https://goo.gl/yzXVR9`



00619183-v

English

assign a date to; determine the (probable) date of; "Scientists often cannot date precisely archeological or prehistorical findings"

date

Portuguese

Gloss: *empty gloss*

Ex.: *empty example*

~~sair com~~ • datar [x]

Relations

- Lexicographer file: (verb.cognition)
- Frame: (Somebody ----s something)
- RDF Type: VerbSynset
- Nomlexes: nm-pt:nomlex-datar-datação
- Hypernym of: [chronologise, misdate]
- Hyponym of: [determine]

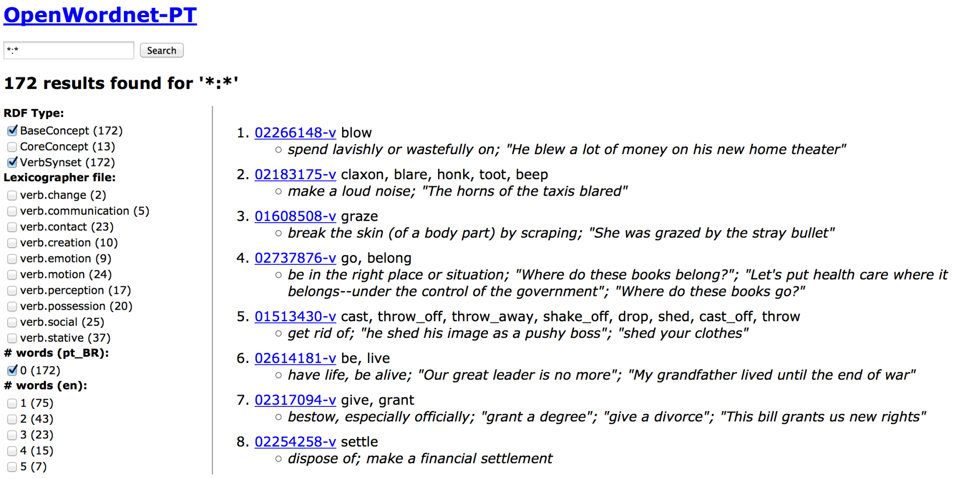
External resources

- OMW
- SUMO

Suggestions

Votes	Action	Content	User (prov.)	Status	Action
3 (3/0)	remove-word-pt	sair com	vcvpaiva (web)	new	del | acc | rej
3 (3/0)	add-gloss-pt	atribuir uma data para; determinar a (provável) data de	(system) (wei-por-30-synset.csv)	new	del | acc | rej

Figure 1: Synset *00619183-v* while voting



OpenWordnet-PT

172 results found for '*:*'

RDF Type:
☑ BaseConcept (172)
□ CoreConcept (13)
☑ VerbSynset (172)
Lexicographer file:
□ verb.change (2)
□ verb.communication (5)
□ verb.contact (23)
□ verb.creation (10)
□ verb.emotion (9)
□ verb.motion (24)
□ verb.perception (17)
□ verb.possession (20)
□ verb.social (25)
□ verb.stative (37)
# words (pt_BR):
☑ 0 (172)
# words (en):
□ 1 (75)
□ 2 (43)
□ 3 (23)
□ 4 (15)
□ 5 (7)

1. 02266148-v blow
   - *spend lavishly or wastefully on; "He blew a lot of money on his new home theater"*
2. 02183175-v claxon, blare, honk, toot, beep
   - *make a loud noise; "The horns of the taxis blared"*
3. 01608508-v graze
   - *break the skin (of a body part) by scraping; "She was grazed by the stray bullet"*
4. 02737876-v go, belong
   - *be in the right place or situation; "Where do these books belong?"; "Let's put health care where it belongs--under the control of the government"; "Where do these books go?"*
5. 01513430-v cast, throw_off, throw_away, shake_off, drop, shed, cast_off, throw
   - *get rid of; "he shed his image as a pushy boss"; "shed your clothes"*
6. 02614181-v be, live
   - *have life, be alive; "Our great leader is no more"; "My grandfather lived until the end of war"*
7. 02317094-v give, grant
   - *bestow, especially officially; "grant a degree"; "give a divorce"; "This bill grants us new rights"*
8. 02254258-v settle
   - *dispose of; make a financial settlement*

Figure 2: Search for ‘All’ (*:*) occurrences of BaseConcepts & VerbSynsets

publishing data: (1) data should rely on URIs to identify objects; (2) URIs should be resolvable; (3) semantic information must be returned, using standards such as RDF; and (4) resources in different datasets should be reused through links between those. Linked Data principles and technologies promote the publication of data on the Web and through this effort guides the emergence of the so-called Linguistic Linked Open Data (LLOD) (Chiarcos et al., 2011b) in which resources and datasets are represented in RDF format and linked to each other. Like many other lexical resources, e.g. Onto.PT (Gonçalo Oliveira and Gomes, 2010) and, more recently, Princeton Wordnet (McCrae et al., 2014), OpenWN-PT is primarily distributed as RDF files, following and expanding when necessary, the original mappings proposed by (van Assem et al., 2006). Both the data for the OpenWN-PT and the vocabulary or RDF model (classes and properties) are freely available for download as RDF and OWL files.

Possibly the main motivation for lexical resources to adopt RDF is the first of the Linked Data principles. The use of URIs allows the easy reuse of entities produced by different researchers and groups. When we started the OpenWN-PT project, there was no official RDF distribution of Princeton Wordnet 3.0 available. We developed our mappings to RDF starting from the original data files and proposed patterns for the URIs to name the original Princeton synsets and our own OpenWN-PT synsets. Following the general convention, to avoid conflict of names, we used a domain name that we have control of. The recently created official RDF distribution of Princeton Wordnet [10] could now serve us better without causing any huge impact on our data. That is, without much effort we can start using the new RDF provided by WordNet Princeton linking it to our RDF files, fulfilling some of the general promise of the semantic web. For instance, looking at the first noun synset of Princeton Wordnet, `00001740-n`: regardless of the different URIs people assign to it, one can readily say that all of them represent the same resource. The fragment of Figure 3 shows the declaration of two statements using the `sameAs` property of OWL ontology.

But there are other advantages of the use of RDF, besides providing a universal way to identify entities. RDF allows us to formally specify the properties and classes that we use to model our data. In our case, we had to suggest properties and classes to represent the extensions to the original WordNet data model that allowed us to embed the lexicon of nominalizations NomLex-PT(de Paiva; Livy Real; Alexandre Rademaker; Gerard de Melo, 2014) into OpenWN-PT. The complete specification of our vocabulary is available at the project repository.[11]

We need to improve our new web interface further as, strictly speaking, the interface does not follow the Linked Data principles two and three: although we do provide the RDF data and an SPARQL endpoint for queries, the URLs of the synsets in the interface are not the same nor are they redirected from the URI of our RDF data. Still, while we intend to conform to the principles in the long run, in the mean time we already harvest some of linked data affordances in terms of provenance capture and use.

Provenance can be used for many purposes, including understanding how data was collected, so that it can be meaningfully used; determining ownership; making judgements about information to determine whether to trust it; verifying that the process and steps used to obtain a result comply with given requirements and reproducing how something was generated (Gil and Miles, 2013). We choose to keep track of the evolution of OpenWN-PT using the provenance PROV (Gil and Miles, 2013) data model and make it available in RDF together with the openWN-PT RDF itself. Figure 4 shows our encoding in PROV data model format of a subset of the current possible suggestions that contributors can make to openWN-PT. The contributors are the actors and they are modeled as `foaf:Person` instances. The `prov:Actitivites` are the possible suggestions of modifications in the data and the `prov:Entity` are the items that can be modified in openWN-PT. Although not present in the figure, the PROV data model allows us to also represent the set of suggestions made by one single automated process.

6.1 Testing and Verifying

To investigate how well the *voting mechanism* is coping with the main issues of end users collaborative work, we have tested it for two weeks

[10] `http://wordnet-rdf.princeton.edu`

[11] `http://github.com/arademaker/openWordnet-PT/`

```
prefix wn30pt: <http://arademaker.github.com/wn30-br/instances/>
prefix wn30en: <http://arademaker.github.com/wn30/instances/>
prefix wn30pr: <http://wordnet-rdf.princeton.edu/wn30/>
prefix owl: <http://www.w3.org/2002/07/owl#>

wn30pt:synset-00001740-n owl:sameAs wn30en:synset-00001740-n .
wn30en:synset-00001740-n owl:sameAs wn30pr:00001740-a .
```

Figure 3: Linking resources using RDF

```
wnlog:AddSense    rdf:type prov:Activity .
wnlog:RemoveSense rdf:type prov:Activity .
wn30:WordSense    rdf:type prov:Entity .
wn30:Synset       rdf:type prov:Entity .

:aword rdf:type wn:Word .
:aword wn30:lexicalForm "ente"@pt .
:asense rdf:type wn:WordSense .
:asense wn30:word :aword .

:s1 prov:used wn30pt:synset-00001740-n .
:s1 prov:used :asense .
:s1 rdf:type wnlog:AddSense .
:s1 prov:atTime "2015-04-15"^^xsd:dateTime .
:s1 prov:wasAssociatedWith :a1 .

:s2 prov:used wn30pt:synset-00001740-n .
:s2 prov:used :anothersense .
:s1 prov:atTime "2015-04-15"^^xsd:dateTime .
:s2 rdf:type wnlog:RemoveSense .
:s2 prov:wasAssociatedWith :a1 .

:a1 rdf:type prov:Agent, foaf:Person .
:a1 foaf:name "Alexandre Rademaker" .
:a2 rdf:type prov:Agent, foaf:Person .
:a2 foaf:name "Livy Real" .
```

Figure 4: Preserving provenance information in the RDF

with three Portuguese native speakers who are researchers interested in language. After two weeks of part-time work, over 2400 votes were cast, 2240 suggestions and 110 comments were made, and we identified over 80 new requirements both for functionality and usability—a testimony, we reckon, of the potential of the tool.

These numbers, although preliminary, show how much the new interface helped us to quickly edit and correct existing synsets. Also, we were pleasantly surprised to realize that, during these two weeks, we had two uknown users, not from the team, collaborating with us, by suggesting entries on the new interface. Since we have not announced the suggestions facility at all, so far, this seems to indicate the easiness of use of the tool. Hence we would like to conclude that there is a need for interfaces that allow ordinary users, not only computational linguists to help on the construction, checking, cleaning up and verification of the quality of (lexical) resources. Just like Wikpedia, we hope to tap into this potential good will.

7 Future Work

We still need to complete our main task, the checking of words, glosses and examples from many English synsets and this is our most pressing work. The theoretical and practical decisions on how to integrate the Portuguese senses that are missing from English are major tasks that will require careful thinking, as these choices will have a huge impact not only on the eventual shaping of OpenWN-PT, but also on our other work with Portuguese NLP.

It seems to us clear that the main design choice of creating a lexical resource for Portuguese by automated methods, complemented by manual curation, following Princeton's model, was the right decision. The curation process is not trivial, but it would not be facilitated by starting manually. Neither do we believe that more could be achieved using only automated methods. Keeping the close alignment with Princeton's wordnet is beneficial in many ways, not least of them, because it allows us to connect to the linked open data community and the ontologies it supports. We are still investigating the benefits of using a lexical model such as lemon (Chiarcos et al., 2011a) and of a possible alignment with it.

References

Tim Berners-Lee. 2011. Linked data-design issues. Technical report.

Bluemix. `http://www.bluemix.net`.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August. Association for Computational Linguistics.

Adam Meyers Leslie Barrett Ruth Reeves Catherine Macleod, Ralph Grishman. 1998. Nomlex: A lexicon of nominalizations. In *Proceedings of EURALEX'98*, Liege, Belgium.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011a. S.: Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, pages 245–275.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011b. Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, 52(3):245–275.

Cloudant. `http://www.cloudant.com`.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open brazilian wordnet for reasoning. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

Valeria de Paiva, Cláudia Freitas, Livy Real, and Alexandre Rademaker. 2014. Improving the verb lexicon of openwordnet-pt. In Laura Alonso Alemany, Muntsa Padró, Alexandre Rademaker, and Aline Villavicencio, editors, *Proceedings of Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish (ToRPorEsp)*, São Carlos, Brazil, oct. Biblioteca Digital Brasileira de Computao, UFMG, Brazil.

Valeria de Paiva; Livy Real; Alexandre Rademaker; Gerard de Melo. 2014. Nomlex-pt: A lexicon of portuguese nominalizations. *Proceedings of LREC 2014*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real, and Anne de Araujo Correia da Silva. 2014. Extending a lexicon of portuguese nominalizations with data from corpora. In Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A. S. Pardo, and Maria das Graças Volpe Nunes, editors, *Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014*, São Carlos, Brazil, oct. Springer.

Yolanda Gil and Simon Miles. 2013. Prov model primer. Technical report. `http://www.w3.org/TR/prov-primer/`.

Hugo Gonçalo Oliveira and Paulo Gomes. 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*, volume 222 of *Frontiers in Artificial Intelligence and Applications*, pages 199–211. IOS Press.

Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes, and Nuno Seco. 2008. PAPEL: A dictionary-based lexical ontology for Portuguese. In *Proceedings of Computational Processing of the Portuguese Language - 8th International Conference (PROPOR 2008)*, volume 5190 of *LNCS/LNAI*, pages 31–40, Aveiro, Portugal, September. Springer.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, (31):91–113.

John McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

NodeJS. `https://nodejs.org`.

Adam Pease and Christiane Fellbaum. 2010. Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet linking project. In *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 2, pages 25–35. Cambridge University Press.

Alexandre Rademaker, Valeria De Paiva, Gerard de Melo, Livy Maria Real Coelho, and Maira Gatti. 2014. Openwordnet-pt: A project report. In *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, jan.

Mark van Assem, Aldo Gangemi, and Guus Schreiber. 2006. RDF/OWL representation of WordNet. W3c working draft, World Wide Web Consortium, June.