NomLex-PT: A Lexicon of Portuguese Nominalizations

Valeria de Paiva[1], **Livy Real**[2], **Alexandre Rademaker**[3], **Gerard de Melo**[4]
1: Nuance Communications, Sunnyvale, CA, USA
2: Universidade Federal do Paraná, Curitiba, Brazil
3: IBM Research and FGV/EMAp, Rio de Janeiro, Brazil
4: Tsinghua University, Beijing, China
valeria.depaiva@gmail.com, livyreal@gmail.com, alexrad@br.ibm.com, gdm@demelo.org

Abstract

This paper presents NomLex-PT, a lexical resource describing Portuguese nominalizations. NomLex-PT connects verbs to their nominalizations, thereby enabling NLP systems to observe the potential semantic relationships between the two words when analysing a text. NomLex-PT is freely available and encoded in RDF for easy integration with other resources. Most notably, we have integrated NomLex-PT with OpenWordNet-PT, an open Portuguese Wordnet.

Keywords: NomLex, Portuguese, nominalizations

1. Introduction

Human language is marvellously flexible in providing numerous alternative ways to express an idea. Often, these alternatives transcend the more conventional associations between categories of form and meaning. While events (and some states) are typically expressed by verbs, there are also many nouns that can be used to refer to them. Unfortunately, this flexibility also leads to significant challenges for computational systems, which typically lack information about such connections between nouns and verbs. Thus, a system encountering the noun *proof* might have difficulty in recognizing its semantic relationship to the verb *to prove*.

In this paper, we describe a freely available computational lexicon for Portuguese that provides mappings between verbs and their nominalizations. While Portuguese is spoken by hundreds of millions of people, in some respects it is still a resource-poor language, especially with respect to freely available resources. The latter are particularly valuable because they enable people to build on each other's work and improve both our understanding of the language and the services that are built for the language.

Our approach is two-fold. Some of our work on building lexical resources is manual and requires detailed expert analyses of linguistic data, which is known to sometimes be boring and frustratingly time-consuming. At the same time, some of the resources that we would like to provide already exist in other languages, and so a certain amount of translation can get us quite far. In particular, we considered the English NOMLEX (Macleod et al., 1998), from which the name our resource, NomLex-PT (for Portuguese Nominalizations Lexicon) is derived, and the French resource NOMAGE (Balvet et al., 2009), a similar project to NOMLEX for French, more recent and based on corpus linguistics. Additionally, new cross-lingual induction algorithms are now mature enough to be applied to large crowdsourced data collections like Wiktionary, which has grown significantly in recent years. We thus set up to expand the fledgling NomLex-PT lexicon with pairs coming from Wiktionary as well as from FrameNet (Baker et al., 1998).

2. Creating the Lexicon

Our basic modus operandi has been to have two researchers independently translate and revise each other's work. We proceeded like that for the data from the NOMLEX project and the data from the NOMAGE lexicon.

2.1. The Initial NomLex-PT Core

To quickly bootstrap the process of creating NomLex-PT, the initial data was manually translated from the freely available English NOMLEX (Macleod et al., 1998), which contains 1,025 English nominalizations. We were pleasantly surprised by how straightforward the translations of the nominalizations in NOMLEX were and how frequently they seemed to correspond to nominalizations in Portuguese. The original NOMLEX has entries formed using the suffixes *-ion*, *-ment*, *-al*, *-er*, *-ee* and *-ing*. For these, we first tried to preserve to the extent possible a direct relation between the original and the translated nominalizations. Fortunately, in Portuguese, there are correspondent suffixes for most of these cases (for example, *-ion/ção*, *-ment/mento* and *-er/-or*). This enabled a straightforward translation for around 90% of the entries. For example, we found 506 entries in NOMLEX formed via the suffix *-ion* and 136 formed with *-ment*, while the Portuguese version Nomlex-PT contains 466 entries formed with *-ção* and 109 entries formed with *-mento*. Most of them also keep a strong relationship between the Portuguese and English verbal roots (e.g. *construction/construção*, *argument/argumento*).

Many nominalizations are erudite words (Su, 2011), especially those formed by suffixation processes. In Portuguese, the most frequent nominalizations are not formed by suffixation, but through zero derivation, as e.g. in the case of the words *compra* (buy) and *luta* (fight) (Rocha, 1998). As we wanted to keep our nominalizations as close as possible to the original entries, some erudite nominalizations that could be translated by a more common word in Portuguese were first translated by an erudite word, to keep the morphological pattern (for example, *arbitration* was translated to *arbitração*, despite the existence of the Portuguese form *arbitragem*, which seems a more frequent form). But in a

second stage we added as many nominalizations as we were able to elicit for each verb. This reflected a change of objectives: our goal changed from constructing a small lexicon aligned to NOMLEX to constructing a higher-coverage nominalization lexicon that could be embedded into our work on a Portuguese version of WordNet.

2.2. Candidates from NOMAGE

NOMAGE is a French nominalization lexicon constructed from the French TreeBank, a one million word electronic corpus for French. From this corpus, possible candidates were automatically chosen based on their derivational suffixes and subsequently manually checked. In total, NOMAGE contains 736 entries, for which it encodes morphological, syntactic and semantic information. Its nominalizations are formed by the suffixes *-ade*, *-age*, *-ance/-ence*, *-ée*, *-ion*, *-ment* and *-ure* (Balvet et al., 2011). Most of them directly correspond to suffixes in Portuguese: *-ade/-ada* (*débandade/debandada*), *-age/-gem* (*décollage/decolagem*), *-ance/-ância* (*ignorance/ignorância*), *-ence/-ência* (*influence/influência*), *-tion/-ção* (*imagination/imaginação*), *-ment/-mento*(*affrontement/afrontamento*), *-ure/-ura* (*signature/assinatura*).

To translate NOMAGE, our first step was manually searching for new entries (comparing NOMAGE entries with our first version of NomLex-PT) and producing a list of 299 candidates. After checking possible translations and the more frequent ones in Portuguese , we added 275 new nominalizations to NomLex-PT.

When translating NOMAGE, we also maintained direct translations as much as possible (e.g., *amplification/amplificação*), but decided to add additional relevant entries when possible. When a word has two possible and frequently used correspondents in Portuguese, we decided to include both of them (*affrontement* brings us *afronta* and *afrontamento*). When we found an entry whose translation was already in NomLex-PT via NOMLEX, but another possible translation seemed to be more salient, we also added it. For example, from the NOMLEX entry *adjustment*, we got *ajustamento*, and based on the French entry *ajustement* we later added *ajuste*, the zero derivation in Portuguese. As a proxy for usage frequency we used simple Google searches in Brazil. For example, *ajuste* seems considerably more frequently used (52.700.000 occurrences in Google against only 1.580.000 for *ajustamento*, in a particular search). NOMAGE has a large variety of present-day words, such as *francisation* and *mondialisation*, since the base of the French TreeBank Corpus is the *Le Monde* newspaper. Their use of newspaper data helps to ensure we do not have only erudite words in our lexicon.

As far as the translation work on NOMLEX and NOMAGE is concerned, we did not consider a purely automatic translation followed by manual correction. While this would have been feasible, the small sizes of the lexica (1025 and 736 nouns, respectively) allowed us to opt for a manual translation process. Especially in this early phase of NomLex-PT, going over the lists gave us more insights about the kinds of nominalizations and phenomena that should be considered.

2.3. Candidates from Wiktionary

To further expand NomLex-PT, we then drew on Wiktionary, the collaboratively created online dictionary. This provided us with a different strand of candidate pairs for nominalizations and their verbs. These were manually checked following the same method as the translations.

As a first step, we extracted translations and other word relationships from both the Portuguese and the English versions of Wiktionary. This requires highly non-trival markup parsing code to turn the semi-structured format of Wiktionary into a machine-readable list of word relationships.

From this list we then first obtained all derivational relationships between a Portuguese word known to be a noun and a second Portuguese word that is known to be a verb (and not known to also be a noun). The part-of-speech information is often not provided together with the derivational relationships but obtained separately from article subsection headers. The result of this process became our first candidate list for manual checking.

However, since the Portuguese Wiktionary is rather small and the English Wiktionary does not cover enough derivational relationships between Portuguese words, we additionally adopted a second technique. We obtained all derivational relationships between an *English* word known to be a noun and another *English* word known to be a verb (and not known to also be a noun). We then obtained all Portuguese translation combinations for this word pair from the English and Portuguese Wiktionary. Any pair of Portuguese translations that shared a common prefix of length n was provided to the human annotators as a candidate nominalization entry. We experimented with different values for n and settled on $n = 4$, which seemed to yield a reasonable compromise between precision and recall.

2.4. Candidates from FrameNet

Finally, we used FrameNet (Baker et al., 1998), a frame-semantic resource to obtain possible candidates. In frame semantics, the same frame may be evoked by either a noun or a verb. With this in mind, we parsed FrameNet to find all English noun-verb pairs that evoke the same frame. For each pair, we obtained all possible Portuguese translations using the English and Portuguese editions of Wiktionary. Again, any pair of Portuguese translations that shared a common prefix of length n (=4 in our experiments) was provided to the human annotators as a candidate nominalization entry. This data was slightly noisier because we did not check the part-of-speech tags of the translations and because the English source pairs were not necessarily derivationally related.

2.5. Candidates from Corpora

A clear next step in this work would be to verify a list of nominalization-forming suffixes in a traditional corpus of Portuguese such as the AC/DC corpus (Santos and Bick, 2000). AC/DC stands for *Acesso a corpora/Disponibilização de corpora* ("acess and availability of corpora"), and was created as one of the activities of the Portuguese Linguateca. The AC/DC cluster contains more than 1 billion words, distributed over the following

genres, ordered by their proportion in the distribution: general newspaper text; narrative fiction; specialized newspaper text; other or non-classified (which includes at least e-mail spam, EU calls, business letters, legal documents and web texts, especially blogs); informative, technical; oral. With regards to language varieties, it includes material from Portugal, Brazil, and Mozambique. It seems reasonable to create lists of nouns with specific suffixes and see which kinds of nominalizations they correspond to.

An approach to construction of a nominalization lexicon, via extraction from corpora was also taken by the authors of Ancora-ES (Peris and Taulé, 2011), a lexicon of Spanish nominalizations, which we only learned about after submitting the first version of this work. Extending our lexicon with nominalizations from corpora has the good effect of making sure that we list roots and lemmas that are typical for Portuguese. As a byproduct, if performed by different linguists, this also enables us to re-check the nominalizations already in place. We are currently pursuing this line of work in a new collaboration with Claudia Freitas, but the results are not yet covered in this paper.

2.6. Additional Extensions

Lastly, since one of us (Real) had written a monograph (Real, 2008) on nominalizations in Portuguese ending in the suffix *-ura*, we created a new list of OpenWordNet-PT nouns finishing in *-ura* and hand-checked which of those were nominalizations.

3. Nominalization Type Annotation and Extension

To make the data more useful for NLP tools, we additionally started marking nominalization types. The need for devising a classification that we could all accept as natural and feasible was keenly felt. There are too many classifications of nominalizations in the literature and consensus does not emerge easily on what to mark and how to mark it. Annotation guidelines were decided on recently, and examples of boundary, difficult cases are still being collected and analysed.

3.1. Agentives

The need for adding additional nominalization types was first felt when, already having a sizeable collection of word pairs, we started measuring the coverage of our lexicon on Portuguese corpora. Our initial attempts revealed that we were far from having the coverage that we had imagined. The experiment that showed this was simply listing the most frequent nouns in the text of our corpus of biographies of Brazilian historical figures (DHBB) (de Abreu et al., 2010) and marking them as nominalizations or not. This experiment, described in more detail later on in Section 4.1., showed that the number of agentive nominalizations in our lexicon was rather low compared to the more event-like nominalizations.

While annotating the automatically created lists above for agent-like denoting nominalizations, we came across many candidate nominalizations that do not denote the agent but for which another agent-denoting noun was known. In these cases, we added additional entries to NomLex-PT so that it covers both the agentive and non-agentive nouns.

To decide which nominalizations are in fact agentives, we have been using a syntactic test. Only nominalizations that refer to the subject position of the verb base were considered agentive nominalizations. So if the subject + verb entails the nominalization, it is marked as agentive, as *ele usa – usuário* (someone who uses – user), *ele canta – cantor* (someone who sings – singer). Agentive nominalizations seem to match this test relatively well, despite its simplicity.

Given that agentives are (grammatically at least) the 'doers' of the verbs, a new annotator asked, should they not simply be all the agents in our data? If so, *instituição* (institution) would be an agentive of the verb *instituir*, however. The agentive form of the verb would have to be *instituidor*, considering the grammatical test. Still, the grammatical test is only a weak marker, as by and large agentive-like nouns can be created on the fly by native speakers for many verbs and we do not necessarily want to add to our lexicon these on the fly constructions. In fact, *instituidor* is an example of a nominalization that one of us thought did not exist, but could be created and it turns out that it is very much used in the specific realm of finance and pensions in Brazil.

3.2. Lexicalized Meanings

Even with this simple test in place, examining the data can provide questions and dilemmas: When has the meaning of a nominalization changed enough that we should call it *lexicalized*? Sometimes this is easy to decide: while a *procuração* (power of attorney) is morphologically related to the verb *procurar*, its meaning is clearly lexicalized. Many of the agentives *aspirador* (vacuum cleaner), *apontador* (sharpener), *condicionador* (conditioner), *elevador* (elevator), while passing the syntactic test, have lost some of the meaning of 'doer-of-the-verb' or taken only a very particular meaning of the verb. Thus while the verb *aspirar* in Portuguese means mostly *to aim to*, the agentive *aspirador* is simply the tool that draws dust from furniture, definitely a lexicalized nominalization. Other examples, however, are less clear: by itself an *abridor* normally is a can opener, but one can use the word with a complement *abridor de caminhos* (opener of paths), *abridor de mares* (opener of seas) as a non-lexicalized deverbal.

As in English, Portuguese nominalizations can be lexicalized to various degrees. For example for the verb *cobrir* (to cover) the natural agentive nominalization would be *cobertor*, but this has been lexicalized as the noun for a bed cover or blanket, so when one wants to talk about the agentive for the more abstract sense of *cover*, as in *a cover-up*, stopping people from knowing something, one instead uses *acobertador*, which has its own corresponding verb *acobertar*. Thus we mark *cobertor* as agentive and lexicalized, while marking *acobertador* as agentive.

3.3. Other Types

While agentive nominalizations are the easiest ones to mark, we are also in the process of marking further classes of nominalizations, such as events states, results. Marking these other kinds of nominalizations is much more difficult than marking agentive nominalizations and maybe having aspec-

tual markings on the verbs related to the nominalizations will help in this task. For instance, for a word like *building*, one can say 'the building of the company took him ten years', and you may mean the physical building (the building that houses the company took him ten years to erect) or the abstract building of the company (growing the customer base, distribution channels, etc.). This example shows again that nominalizations are often subject to polysemy, and that the specific sense may only become clear in context.

The example works similarly in Portuguese, except that while the abstract meaning of 'building of a company' ('a construção da empresa levou dez anos') works as in English, the lexicalized nominalization (*construção*) means not the finished building (in Portuguese *edifício, prédio*) but the working site.

Agentive nominalizations are easier to recognize as they tend not to have too many independent meanings: a *pintor* (painter) is someone who paints, whether we are thinking of high art or of painting houses. By way of contrast, a suffix like *-mento/ção* always has the basic meaning of an event, but it usually has more than one meaning and the extra meanings can be hard to classify. Real's recent work (Real, 2014) suggests eight possible classes: action of (which we also call 'event'), result of, physical result of, iteration of the act of, resulting state from, abstract result of, locative, collectivization of.

For instance, *parada* (stop) can be used as an event, a resultative state and a locative. The word *oposição* (opposition) can assume six of these eight meanings (it does not take the iteration and locative senses). *Desenvolvimento* (development) is used in the sense of event, result of, abstract result and resultative state, while *construção* can mean event, result of, physical result and locative. This proliferation of meanings can be daunting for computationally motivated lexica. In her more theoretical work, Real (op. cit.) proposes a lexical system where part of the lexical meaning is only defined in specific contexts, being underspecified in general. Deciding how to translate the more theoretical insights into computational data annotation is work in progress.

4. Results

We now have an automatically seeded, but manually checked lexicon of over 3,000 nominalizations and corresponding verbs. Statistics are provided in Tables 1 and 2. Given that the original NOMLEX covers many of the most important nominalizations in English, it is not unnatural for our manual translations of NOMLEX to play an important role. The table also reveals significant numbers of manual additions as well as pairs coming from other sources. Note that the counts for these later sources exclude any duplicates already covered by the original NOMLEX translations. The DHBB-motivated manual additions are explained below, as they were derived from our coverage experiment.

The additions from the lists produced via Wiktionary and FrameNet, as well as the manual ones, made us realize the existence of many nominalization forming suffixes in Portuguese that do not seem to exist in English or French, such as the suffixes *-ida*, *-iz* or *-ouro*. Collecting nominalizations with these suffixes is mostly future work, hence the small numbers in Table 2.

Table 1: Coverage of NomLex-PT

Approach	Number of Entries
NOMLEX Translations	1031
Linguateca Additions	860
Manual Additions	709
NOMAGE Translations	262
DHBB-Motivated Manual Additions	158
Wiktionary	152
FrameNet	142
OpenWordNet-PT	82
Total	3,396

Currently, around 600 of the nouns are marked as agentive, while the majority of the rest prototypically denote eventualities, except for those that are marked as having a lexicalized meaning referring to something else (so far around 100 have been marked).

4.1. Coverage Experiment

One of our concerns was that the nominalizations in NomLex-PT might not be representative of the ones used in mainstream written Portuguese. Since nominalizations tend to be more frequent in erudite texts, we chose to evaluate our lexicon on the 'Historical Dictionary of Brazilian Biographies' (the acronym in Portuguese is DHBB) (de Abreu et al., 2010), a somewhat more academic corpus of short biographies or "snippets" for historical figures, taken from a project that some of us had started working on with colleagues at the Getulio Vargas Foundation (FGV). Our goal was to measure NomLex-PT's coverage with respect to the DHBB corpus. While still somewhat erudite, the authors of this data were asked to conform to a specific kind of style and vocabulary. This is by no means a sort of "controlled Portuguese" – authors of entries in the DHBB were only given general guidelines, requesting them to be accessible in tone and content to undergraduate students of humanities. We processed this corpus of short biographical and historical entries with with Freeling (Padró and Stanilovsky, 2012) to obtain part-of-speech tags and lemmatized forms. From the set of all noun lemmas with more than 750 occurrences in the corpus, a total of 454 distinct nouns, we had 165 nominalizations already in NomLex-PT, 169 nouns that we deemed not nominalizations, and 111 missing nominalizations, as well as 9 mistakes in processing. Thus our error rate is 24.9% (111/445), and the missing nominalizations have subsequently been added to the lexicon (these are most of the DHBB-motivated additions listed in Table 1).

4.2. RDF Encoding and Embedding into OpenWN-PT

We integrated NomLex-PT into OpenWordNet-PT, a version of WordNet for Brazilian Portuguese. OpenWordNet-PT's main characteristics are its open-source license, its direct correspondence with Princeton WordNet, and, given its origins in UWN, the Universal Wordnet (de Melo and Weikum, 2009), both a high recall and a high precision for the more salient words in the language.

Our choice of encoding both OpenWordNet-PT and NomLex-PT in RDF makes the merging of these resources straightforward. Traditionally, lexical resources are distributed in a number of different formats, which make their re-use and inter-connection with other resources laborious and error-prone. The Ontology-Lexica Community Group is one of these groups that aims to demonstrate the added value of representing lexica using Semantic Web and Linked Data standards to improve the re-usability of existing linguistic information. In particular, RDF and OWL are promising as technologies to share data in a particularly flexible way. For the most part, this flexibility comes from their ability to describe not only the data instances but also the data model used in the data, in the same language. The annotations over classes and properties make it easy to create and document new models and embed them in the resource. Another important aspect of the use of these standards is the use of URIs to name entities (Bizer et al., 2009). This allows different groups and institutions to easily exchange globally named entities in their resources without name conflicts and ambiguity.

To represent OpenWordNet-PT and NomLex-PT as RDF resources, we started from an already developed WordNet RDF Vocabulary (van Assem et al., 2006) and extended it as needed. In order to incorporate NomLex-PT into this encoding, we extended the RDF-based vocabulary to additionally describe relevant parts of the NOMLEX syntax. Figure 1 presents a subgraph for the nominalization entry *promover/promoção* and its connection to OpenWordNet-PT. Note that the link between NomLex-PT and OpenWordNet-PT is achieved through the properties *noun* and *verb*. Both properties have as domain an instance of the class *Nominalization* and as co-domain an instance of *WordSense* or *Word* (from the OpenWordNet-PT vocabulary). The details of this encoding are described elsewhere in detail (Rademaker et al., 2014). Our wordnet and NomLex vocabularies are freely available and downloadable from `https://github.com/arademaker/openWordnet-PT`.

This embedding of NomLex-PT into the open version of a Portuguese wordnet was helpful in multiple respects. First, it solved some minor problems with handling diacriticals, as OpenWordNet-PT has a consistent treatment of these. Secondly, by checking how the nominalizations from NomLex-PT were related to the corresponding verbs in the wordnet version, we realized that some synsets were missing in OpenWordNet-PT. These have been added manually. Finally, by various extensions of the original English NOMLEX leading up to NomBank (Meyers et al., 2004), we hope to spot-check the consistency of OpenWordNet-PT with respect to other specific linguistic phenomena, e.g., the phenomenon of diminutivization of nominalizations. Portuguese uses many diminutives and there seems to be a scale of importance between the suffix used for a nominalizations. For example the verb *ler* (to read) produces nominalizations *leitura*, *lida*, *leitor*, where the last is the agentive nominalization (reader) and the two first ones can be seen as the event/result of reading, but the second is less important, more superficial.

There is some interesting data (de Medeiros, 2008) on other suffixes for forming nominalizations in Portuguese, where you have a definitive semantic effect of making the nominalization less important, less serious, some 'diminutivization' that does not seem to occur in English. Examples are *vassourada*, *olhada*, *lida*. A *vassoura* is a broom in Portuguese and the verb *vassourar* is not very frequently used but it would correspond to 'to clean with a broom'. Its associated nominalization *vassourada*, corresponding to 'a superficial cleaning', is very frequently used indeed. Similarly, an *olhada* is a superficial look (the verb in Portuguese is *olhar*) or glance, and *lida* is a superficial, less important kind of *reading*.

Table 2: Nominalizations in NomLex-PT by Suffix

Nominalization Suffix	Total Number
-ção	969
-mento	329
-ida	9
-ura	96
-or	891
-nte	111
-ada	97
other	894

5. Discussion

As an organic project, this work has seen some changes of direction. We started simply trying to reproduce the manually curated NOMLEX and NOMAGE projects in Portuguese, in order to establish a baseline. Then we decided to enhance this first version of the lexicon, using Wiktionary and FrameNet as sources.

Meanwhile, we are experimenting with growing our lexicon using Portuguese corpora, more specifically using the large AC/DC corpus. Extending NomLex-PT with curated lists of nominalizations extracted from the AC/DC corpus, we should be relatively confident of our coverage. Our next step is to complete the list of nominalizations with further suffixes such as *-ada*, *-iz* (*debandada*, *chamariz*) that we have not investigated. As mentioned, we are obtaining lists of candidate nominalizations from the AC/DC corpus for all those suffixes. We also plan on going through our list of frequent nouns in the DHBB to check for possible nominalizations with less than 750 occurrences, to complete a second pass in the DHBB project.

For the most part, however, we believe we have a reasonable coverage of Portuguese nominalizations in use and that we should investigate ways of making these nominalizations more *informative*, via the connection to the corresponding verbs. While adding new pairs of verbs and their nominalizations turned out to be easy, classifying the nominalizations to, eventually, extract more computationally relevant information from the lexicon has proved challenging. So far the classification has been only of agentive and/or lexicalized nominalizations. We are in the process of deciding whether a small number of additional nominalization types can be annotated, as described earlier in Section 3., in order to make NomLex-PT effective in the process of detecting events via

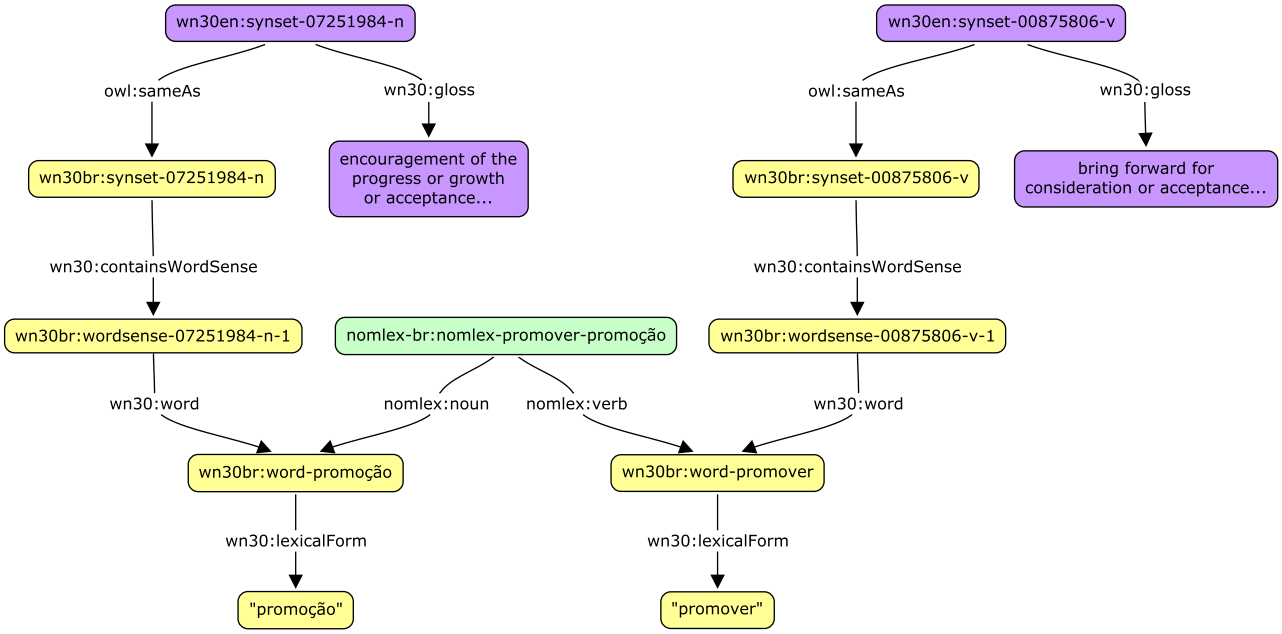

Figure 1: Entry *promover/promoção*

language processing. The literature has several markings such as property, location, etc., which we do not capture at the moment. While increasing our stock of nominalizations via Portuguese corpora, we are also investigating the kinds of classification of nominalizations that are easy to annotate and devising ways of incorporating this information into our RDF lexicon. One idea is to bootstrap our decisions from the ones recommended by the lists of Port4Nooj (Barreiro, 2010) and Cartão (Oliveira et al., 2009), this last one following a suggestion of Hugo Gonçalo Oliveira, gratefully acknowledged.

Finally, another aspect of our project has been looking into embedding NomLex-PT into our version of an open Portuguese wordnet, OpenWordNet-PT. There is some hope that this integration will help us tackle some of the polysemy-related issues related to nominalization types. However, OpenWordNet-PT being fully aligned with Princeton WordNet also brings in all the issues with the fine-grained senses of WordNet. Putting it bluntly, choosing consistent synsets for both the verb and the noun senses of our NomLex-PT pairs is complicated. We are still working on a mechanism for defining consistent choices. More importantly, the process of extending the originally translated lexicon of Portuguese nominalizations has helped us devise several ways of improving the information contained in the bigger resource OpenWordNet-PT. While we always thought we should build OpenWordNet-PT by automatic harvesting of freely available dictionary and Wikipedia data, followed by manual curation, we mistakenly thought the curation could have been done on the basis of curating individual items of OpenWN-PT. After several weeks of trying this strategy with the help of a junior intern, we realized that curation is better done along the lines of linguistic phenomena, such as nominalization and its classification, correlation between adjectives and adverbs, nominalization of adjectives, detection and classification of support verbs, classes of multi-word expressions, etc.

Most of the work we had done so far on curating OpenWordNet-PT had shown the need for removing excessive synsets, where the English WordNet itself makes (perhaps too) fine-grained distinctions that we could not see in Portuguese. While it seems true that there are more verbs that indicate walking in a specified manner in English than in Portuguese, (for example, to say *jogging* in Portuguese we need to say almost-running or walking-very-fast, to say *strutting* we need to say walking-with-pride, etc) it is also the case that for the verb *walk* itself WordNet has ten synsets, of which only two correspond to clearly identifiable senses in Portuguese. Thus we had started our curation process by removing synsets that seemed excessive. But the work on nominalizations made us realize that more important than removing these fine-grained senses was to make sure that we keep verbal and noun senses connected, when they exist in both languages. And for that, some re-structuring of the verb hierarchy, preferably paying attention to WordNet semantic classes would be a plus. This work is only beginning now.

We still need an open and free version of a lexicon of verbs in the mould of Palmer and Kipper's VerbNet for Portuguese. By projecting the verbs in OpenWordNet-PT and trying to complete the lexicon thus obtained with verb subcategorization information, we should have a generic improvement of OpenWordNet-PT. Several initiatives (Scarton and Aluısio, 2012; Cançado, to appear) for designing such a resource, a VerbNet-PT, are already in course, but, to the best of our knowledge, none is really available for use and downloading.

Nevertheless, the current form of the extended NomLex-PT lexicon already seems to be useful for some applications. First of all, there is an obvious potential for relation extraction. In the small historical biographies DHBB project described in (Rademaker et al., 2013), we wish to recognize events like "casou-se em 1905 com.../was married in 1905

to..." as being the same semantically as "seu casamento em 1905 com.../his marriage in 1905 to...". The specific nominalizations required for this project seem to be limited in number (due to the strict guidelines imposed to the writers of the short biographical texts) and hence to complete the project we need to deal with other temporal and named entity detection and recognition problems.

6. Conclusion

Using a combination of manual and automated methods drawing on existing resources, we have created a representative lexicon of nominalizations in Portuguese that is useful for linguists as well as in digital humanities settings, e.g. for the exploration of corpora related to historical biographies. Still, much work remains to be done to obtain freely available, high quality, semantics-orientated lexical resources and tools for Portuguese. Our previous work on OpenWordNet-PT (de Paiva et al., 2012) also combined automatic and manual methods to create a large Portuguese wordnet. We have taken first steps to combine the two resources by connecting NomLex-PT to it using an RDF representation. Our original motivation to produce a lexicon of Portuguese nominalizations was to help with the knowledge representation of events in unstructured text, in a manner similar to the one described in (Gurevich et al., 2006). This idea of a lexicon of nominalizations in Portuguese is part of a larger project of providing systems and resources to produce semantic analyses of Portuguese text, in a manner similar to what the Bridge Project at PARC (Bobrow et al., 2007) could do for English. Many other components have still to be designed and implemented. We thus hope that NomLex-PT will be an important building block towards this larger endeavor.

7. References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proc. COLING-ACL 1998*, pages 86–90.

Balvet, A., Haas, P., Huyghe, R., Jugnet, A., and Marín, R. (2009). The NOMAGE project: Annotating the semantic features of french nominalizations. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 264–267, Tilburg, The Netherlands, January. Association for Computational Linguistics.

Balvet, A., Barque, L., Condette, M.-H., Haas, P., Huyghe, R., Marín, R., and Merlo, A. (2011). La ressource Nomage: Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus. *TAL*, 52(3):1–24.

Barreiro, A. (2010). Port4nooj: an open source, ontology-driven portuguese linguistic system with applications in machine translation. In *Proceedings of the 2008 International NooJ Conference (NooJ'08)*, Budapest, Hungary, June. Cambridge Scholars Publishing.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Bobrow, D. G., Cheslow, B., Condoravdi, C., Karttunen, L., King, T. H., Nairn, R., de Paiva, V., Price, C., and Zaenen, A. (2007). PARC's bridge and question answering system. In *Proceedings of Grammar Engineering Across Frameworks*, pages 26–45.

Cançado, M., G. L. A. L. (to appear). *Catálogo de Verbos do Português Brasileiro: classificação verbal segundo a decomposição de predicados. Parte I: verbos de mudança*. Editora UFMG.

de Abreu, A. A., Lattman-Weltman, F., and de Paula, C. J., editors. (2010). *Dicionário Histórico-Biográfico Brasileiro pos-1930*. CPDOC/FGV, Rio de Janeiro, 3 edition. available at http://cpdoc.fgv.br/acervo/dhbb.

de Medeiros, A. B. (2008). *Tracos Morfossintáticos e Subespecificacão Morfológica na Gramática do Português: Um Estudo das Formas Participiais*. Ph.D. thesis, UFRJ.

de Melo, G. and Weikum, G. (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.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An open Brazilian wordnet for reasoning.

Gurevich, O., Crouch, D., King, T. H., and de Paiva, V. (2006). Deverbal nouns in knowledge representation. *FLAIRS: The Florida Artificial Intelligence Research Society*, May.

Macleod, C., Grishman, R., Meyers, A., Barret, L., and Reeves, R. (1998). Nomlex: A lexicon of nominalizations. In *Proceedings of Euralex 1998*, Liege, Belgium.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The nombank project: An interim report. In Meyers, A., editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Oliveira, H. G., Santos, D., and Gomes, P. (2009). Relations extracted from a portuguese dictionary: results and first evaluation. In *Local Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA 2009)*, Aveiro, Portugal, October.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *Proc. of the 8th Intern. Conf. on Language Resources and Evaluation (LREC'12)*, pages 23–25, Istanbul, Turkey, may.

Peris, A. and Taulé, M. (2011). Ancora-nom: A spanish lexicon of deverbal nominalizations.

Rademaker, A., Higuchi, S., and Oliveira, D. A. B. (2013). A linked open data architecture for contemporary historical archives. In Predoiu, L., Mitschick, A., Nurnberger, A., Risse, T., and Ross, S., editors, *Proceedings of 3rd edition of the Semantic Digital Archives Workshop*, Valetta, Malta. to be published. Workshop website at `http://mt.inf.tu-dresden.de/sda2013/`. Proceedings at `http://ceur-ws.org/Vol-1091/`.

Real, L. (2008). Uma análise do sufixo -ura com base na morfologia categorial. *Revista InterteXto*, 1.

Real, L. (2014). *Nominalizações*. Ph.D. thesis, Federal University of Paraná (Submitted).

Rocha, L. C. A. (1998). *Estruturas morfológicas do português*. Editora UFMG, Belo Horizonte, Minas Gerais, Brazil.

Santos, D. and Bick, E. (2000). Providing internet access to portuguese corpora: the AC/DC project. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pages 205–210, Athens, Greece, June.

Scarton, C. and Aluısio, S. (2012). Towards a cross-linguistic VerbNet-style lexicon for Brazilian Portuguese. In *Proceedings of the LREC 2012 Workshop on Creating Cross-language Resources for Disconnected Languages and Styles*.

Su, L. I.-w. (2011). Nominalization as a rethorical device of academic discourse. In *YZU Workshop on Language Structure and Language Learning*.

van Assem, M., Gangemi, A., and Schreiber, G. (2006). RDF/OWL representation of WordNet. Technical Report W3C Working Draft 19 June 2006, W3C. `http://www.w3.org/TR/wordnet-rdf/`.