Extending NomLex-PT using AnCora-Nom

Livy Real[1], Valeria de Paiva[2], and Alexandre Rademaker[3]

[1] Universidade Federal do Paraná, BR
[2] Nuance Communications, USA
[3] IBM Research and FGV/EMAp, BR

Abstract. This work describes how we used AnCora-Nom, a Spanish nominalization lexicon, to extend NomLex-PT, a lexical resource for Portuguese, originally based on the English NomLex lexicon and fully integrated to OpenWordNet-PT, our freely available Portuguese WordNet. The complete Spanish lexicon, which contains 1,655 entries, was translated to Portuguese and then compared to our previous data. Further comparison between the different kinds of nominal classification in AnCora-Nom and NomLex-PT is underway.
Keywords: Nominalizations, Lexical Resource, Portuguese, Spanish

1 Introduction

This work is part of a larger project of creating a lexical resource called NomLex-PT [8], a dictionary of nominalizations in Portuguese, originally based on the English NomLex [7] lexicon and freely available at *github.com/arademaker/nomlex-pt*. Nominalizations are nouns derived from lexical items, usually verbs. These are also called deverbal nouns. In previous work, described in [8, 5], we explained the motivation and the method we followed to construct our lexicon of nominalizations. After completing what we thought was the last step in the construction of the entries of NomLex-PT, that is, after checking that the nominalizations appearing in the corpus AC/DC [13] were on our lexicon NomLex-PT, we realized that the Spanish group behind Ancora-ES [15] had a lexicon of nominalizations, similar in some ways, to our own. Thus we set out to translate this Spanish lexicon and to add the nominalizations not in our lexicon to it. This note describes how we did it and the lessons learned.

2 NomLex-PT

Freely inspired by NomLex [7], NomLex-PT comes from the necessity of having lexical resources in Portuguese, especially Brazilian Portuguese, for the goal of extracting information automatically from Portuguese texts. We started from a manual translation of NomLex to Brazilian Portuguese originally called NomLex-BR, with 1025 nominalizations. Since the English NomLex was composed by entries formed from nominalizers suffixes (*ion, -ment, -al, -er, -ee, -ing*), it had a straightforward translation to Portuguese for around 90% of the nominals.

The translation task was direct and quick, as we found many pairs, such as *construction/construção, argument/argumento* and *observer/observador* of parallel pairs in Portuguese and English. Also, many translated nominals did not keep the same root after the translation, but they kept the same suffix, as e.g. *eater/comedor* and *singer/cantor*. Since we opted to preserve a morphological relation between English/Portuguese pairs, a nominal such as *arbitration* was translated as *arbitração*, a possible and verified word in Portuguese, even if the most used word for it might be *arbitragem*. In this way, we kept NomLex-BR as close as possible to the original NomLex. The original NomLex contains many erudite words as nominalizations themselves tend to be erudite [14] and nominalizations formed by suffixation even more so [12]. This made our first version of NomLex-BR a very erudite database.

In order to increase our database and also to include more common usage nominalizations, especially formed by zero derivation, we invested in translating the French Nomage lexicon [3] and adding to it nominals derived from other databases such as Wiktionary, Wikcionario and Framenet. Nomage is a French nominalization lexicon which contains 736 entries collected from a *French Treebank* [1]. From Nomage, we added 275 new nominalizations to NomLex-PT, trying to keep a direct translation as much as possible. This was facilitated by the fact that the nominalizer suffixes that were chosen for the project Nomage also have direct translations to Portuguese. Candidate nominalizations from Framenet, Wiktionary and the corpora composing the AC/DC repository helped us to add more common nominalizations to NomLex-PT. Before considering AnCora-Nom, NomLex-PT had 4027 entries and now it consists of 4238 entries. An extended discussion of how and why we extended our lexicon can be found in [8].

Our NomLex-PT lexicon is also integrated with OpenWordnet-PT, an open-source licensed version of Princeton WordNet for Portuguese. A description of how we integrated both resources is detailed in [5]. This useful electronic lexicon can be found at github.com/arademaker/openWordnet-PT.

3 AnCora-Nom Nominals

To make sure that our lexicon is as good as possible in coverage, to extend our data and also be able to compare our work with other lexical resources, we looked at the Spanish nominals in the AnCora-Nom [9] lexicon. AnCora-Nom is a Spanish nominalizations database automatically extracted from the annotated AnCora-ES corpus [15]. It contains 1,655 entries and 3,094 senses. Each sense has a denotation type associated, the mapping of the nominals is complemented with arguments and the corresponding theta roles are also annotated.

To produce this lexicon, their authors first mark if an entry is lexicalized or not – using the attribute 'lexicalized' and the value 'yes' – then the attribute 'denotationtype' is assigned to the deverbal noun together with the attribute 'originlexicalid', whose value is the base verb; this links an entry with the corresponding verbal lexical entry in AnCora-VerbNet-Es [2]. In AnCora-Nom, there

are three different denotation types: event, result, and underspecified. Those types were decided upon considering several criteria discussed in [10], which follow the traditional linguistic literature on nominalizations, as [6]. Those criteria include the presence of an incorporated (internal) argument, plurality, determiners, complementation and Vendler's classification of verbal classes. Following [6], for example, only resultative nominals can pluralize and they do not accept internal arguments.

AnCora-Nom uses the AnCora-Verb lexicon [2] to obtain the semantic verbal information related to the argument structure. So AnCora-Nom has two different kinds of semantic information encoded, denotation type and argument structure, and the last one is automatically extracted from the verb corpora. NomLex-PT does not provide information about argument structure since we take the view that the verbal structure is not always inherited by the nominals [11, 4] and does not necessarily determine the final meaning of a nominalization.

4 Method and Results

We first translated all the nominalizations from Ancora-Nom to Portuguese via Google Translator and manually checked all the results. We only use the nominal/verb pair information from Ancora-Nom. When it is possible to keep a direct translation from a Spanish nominal into Portuguese, we prefer it, as in for example, *acontecimiento/acontecimento* and *agrupación/agrupação*, even if those words could be translated as *evento* and *grupo*, and this might be more colloquial. A few nominals were translated as two different Portuguese entries, trying to keep the direct morphological translation but also the most used one, for example *outorgamento*(*grant*) was translated as *outorgamento* and *outorga*.

In a second step we verify the presence (or not) of both elements of the pair nominal/verb in both NomLex-PT and OpenWordNet-PT. Finally we compare our semantic classification of nominals with the AnCora-Nom classification. From AnCora-Nom we needed to add 211 nominals to NomLex-PT and 136 new verbs to OpenWordNet-PT. Most of these new verbs and nominals are derived forms from other entries that were already included on our database, as *pré-inscrever* (*pre-enroll*) and *pré-matrícula* (*pre-registration*), derived forms from *inscrever* and *matrícula*, that were already included. The other new forms included are, in general, either very informal, as in *dopar* (*to dope*) and *azucrinar* (*to pester*), or very erudite, as *escrutinizar* (*to scrutinize*) and *vaticinar* (*to predict*).

AnCora-Nom also includes 'cousin' nominals, which are nouns that are semantically related to another part of speech lexical item. Examples are nouns derived from adjectives such as *complacência* (*complacency*) and *brancura* (*whiteness*), derived from the adjectives *complacente* (*complacent*) and *branco* (*white*). Since NomLex-PT only includes deverbal nominalizations, some of the cousin nominals present on AnCora-Nom are considered out of scope for NomLex-PT. But these are few, only 24 items.

Given that our lexicon NomLex-PT is RDF-encoded, we hope to perform further experiments, measuring how well the verbal argument structures predicted by AnCora-Nom correspond to the nominal structures in the corpora that we have been studying.

5 A semantic classification of nominalizations

To increase the quality of the information we have on our lexical resource, we also started a semantic classification of those nominals. Since the first author has recently finished a doctoral thesis on the semantics of eventive nominalizations [11] in Portuguese, it was a natural move for us to simplify Real's classification of nominalizations and seek a simpler formulation of that theoretical work, more adapted to the computational work in NomLex-PT.

Recalling briefly Real's work [11], nominalizations that keep as the main meaning the same event as the base verb, are called **eventive nominals** (e.g. *construção/construction*) and they can have seven possible semantic types. They are: event, physical result, abstract result, resultative state, collectivization, locative, and instrument. Some examples are in the appendix.

For Nomlex-PT we are interested in all kinds of deverbal nominals, not only on the eventive ones. Thus our classification is different. Most of those nominals that carry the eventive reading are vague and can mean more than simply eventive. *Construção* (*construction*) besides being the process of constructing something, is also the result (abstract or not) of a process and *parada* (*stop*) can be the event of stopping or the location where the stopping happens.

However, agentive nominals seem to be more semantically stable. *Escritor* (*writer*) is someone who writes, whether or not professionally. *Pintor* (*painter*) is someone who paints (great pictures or building walls). Also agentive nominals have a simpler morphology: in general, they are formed by the *-or* morpheme, which has *-tor* and *-dor* as allophones. Another agentive suffix in Portuguese is *-nte*, as in *estudante/student*. Because of this regularity, we decided to start our classification with agentive nominals. Agentive nominals are marked 'agentive' on the lexicon and, in general, they have this only single meaning.

Eventive nominals are more difficult to classify, but the eventive reading is still the basic meaning of nominalizations, thus we decide to mark them as the default, that is simply 0 (none). The nominals which have lost this basic eventive meaning, we mark 'lex', which stands for lexicalized. So nominals that mean simply the event do not receive any special mark, but nominals that have lost this meaning and started to refer to other relations, as for example *cruzador* (*cruiser*) and *abotoadura* (*cufflink*), are marked as 'lex'. The ones that have lexicalized meanings but also keep the eventive one, we mark "b", which stands for 'both meanings', as e.g. *declaração*, which can mean *declaration* but it is also a special document with juridical value.

6 Conclusion

We described comparing and adding the nominalizations from the Spanish AnCora-Nom to the Portuguese nominalizations in NomLex-PT. From AnCora-Nom, we added to NomLex-PT 211 nominals and now our lexicon consists of 4238 entries. NomLex-PT has a wider scope of nominalizations, including agentive ones, which are not considered by AnCora-Nom, as well as eventive and lexicalized ones. Nevertheless we found a relevant number of entries which were not in the NomLex-PT already, most of them derived forms by prefixation. We believe this reflects a characteristic of the AnCora-Nom database more than a special difference between Spanish and Portuguese nominals, as almost all of them have a straightforward translation into Portuguese, as for example *recubrimiento/recobrimento* (*covering*) and *reincorporación/reincorporação* (*reincorporation*).

Since Portuguese and Spanish are such closely related languages, we expected a close relationship between their collections of nominalizations, which seems to exist. Classification of those nominals of course is a different problem and more differences were expected, and this turnout to be the case too. It remains future work to work out means of measuring the usefulness of these parallel classifications. Comparing how those different lexical resources work might shed some light to purely linguistic discussions, as different linguistic theories tell very different stories about nominalizations and more practical work would be useful to check the correctness and usability of those resources.

References

1. Abeillé, A., Clément, L., Toussenel, F.: Building a treebank for french. In: Abeillé, A. (ed.) Treebanks. Kluwer (2003)
2. Aparicio, J., Taulé, M., Martí, M.A.: Ancora-verb: A lexical resource for the semantic annotation of corpora. In: Proceedings of Language, Resources and Evaluation (2008)
3. Balvet, A., Barque, L., Condette, M.H., Haas, P., Huyghe, R., Marín, R., Merlo, A.: La ressource nomage. Traitement Automatique des Langues 52(3), 1–24 (2011)
4. Brandtner, R.: Deverbal nominals in context: Meaning variation and copredication. SinSpeC 8 (2011)
5. Coelho, L.M.R., Rademaker, A., Paiva, V.D., de Melo, G.: Embedding nomlex-br nominalizations into openwordnet-pt. In: Orav, H., Fellbaum, C., Vossen, P. (eds.) Proceedings of the 7th Global WordNet Conference. pp. 378–382. Tartu, Estonia (jan 2014)
6. Grimshaw, J.: Argument Structure. MIT Press (1990)
7. Macleod, C., Grishman, R., Meyers, A., Barret, L., Reeves, R.: Nomlex: a lexicon of nominalizations. Proceedings of Euralex (1998)
8. de Paiva, V.D., Real, L., Rademaker, A., Melo, G.D.: Nomlex-pt: A lexicon of portuguese nominalizations. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland (may 2014)
9. Peris, A., Taulé, M.: Ancora-nom: A spanish lexicon of deverbal nominalizations. Procesamiento del Lenguaje Natural 46, 11–18 (2011)

10. Peris, A., Taulé, M., Rodríguez, H.: Semantic annotation of deverbal nominalizations in the spanish ancora corpus. In: Dickinson, M., Müürisep, K., Passarotti, M. (eds.) Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (2010)
11. Real, L.: Nominalizações. Ph.D. thesis, Universidade Federal do Paraná (2014)
12. Rocha, L.C.A.: Estruturas morfológicas do português. Editora UFMG, Belo Horizonte (1998)
13. Santos, D., Bick, E.: Providing internet access to portuguese corpora: the AC/DC project. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000). pp. 205–210. Athens, Greece (June 2000), http://www.linguateca.pt/documentos/SantosBickLREC2000.pdf
14. Su, L.I.w.: Nominalization as a rethorical device of academic discourse. YZU Workshop on Language Structure and Language Learning (2011)
15. Taulé, M., Martí, M.A., Recasens, M.: Ancora: Multilevel annotated corpora for catalan and spanish (2008)

A Examples of nominalizations and their Real classification

(1) *A assinatura dura três meses.* (resultative state)
The subscription lasts three months.

(2) *A assinatura está torta.* (physical result)
The signature is crooked.

(3) *A assinatura custou caro.* (abstract result)
The signing was expensive.

(4) *A assinatura do contrato levou três horas.* (event)
The signing of the contract lasted three hours.

(5) *A administração está louca.* (collectivization).
The administration is crazy.

(6) *A saída é aqui.* (locative)
The way out is here.

(7) *A obturação está quebrada.* (instrument).
The filling is broken.