

Extending a Lexicon of Portuguese Nominalizations with Data from Corpora

Cláudia Freitas¹, Valeria de Paiva², Alexandre Rademaker³,
Gerard de Melo⁴, Livy Real⁵, and Anne Silva¹

¹ PUC-Rio, Brazil, {anne.acsilva,maclaudia.freitas}@gmail.com

² Nuance Communications, USA, valeria.depaiva@gmail.com

³ IBM Research and FGV/EMAp, Brazil, alexrad@br.ibm.com

⁴ Tsinghua University, China, gdm@demelo.org

⁵ UFPA, Brazil, livyreal@gmail.com

Abstract. We describe the extension of a lexicon of nominalizations in Portuguese, NomLex-PT, with nominals from a collection of corpora, the AC/DC corpora. The resulting lexicon of nominalizations is RDF-encoded and integrated with OpenWordNet-PT, a Portuguese WordNet freely available to download and consult. We discuss the reasons for this extension with corpus data, the methodology we followed, as well as our reasons for suggesting that the extended lexicon of nominalizations is a useful resource for researchers interested in Knowledge Representation of information extracted from Portuguese texts.

Keywords: nominalizations, Portuguese, lexical resources, corpora

1 Introduction

Lexical databases form an essential component of many modern Natural Language Processing (NLP) systems and frequently consist of large amounts of highly detailed and well-curated entries. Examples of such resources include the Princeton WordNet [1], a lexical-semantic network, NomLex [2], NomLexPlus, and NomBank [3], three resources that describe nouns, VerbNet [4], a verb lexicon, as well as FrameNet [5], a hierarchically structured database of prototypical situations, just to name a few. While each one of these resources is useful on its own, if we could unify them and relate them in a consistent manner, letting each resource complete the gaps in the others, we would have much more information than we have now.

In particular for languages where fewer resources exist, such as Portuguese, the ability to pool together the resources available and to complete them, using external resources and advanced natural language techniques to adapt these to the language at hand is a possible way forward.⁶ We have been engaged in a project for producing lexical resources for Portuguese for a few years. So far, we have been able to share two resources with the community. The

⁶ The complementarity of lexical resources for Portuguese NLP is investigated in [6].

first is OpenWordNet-PT [7], an open-source version of WordNet. The second, NOMLEX-PT, is the result of our first steps towards the goal of creating a comprehensive lexicon of Portuguese nominalizations.

Continuing the overarching project of creating lexical resources for Portuguese, we started a smaller project, aimed at producing a lexicon of Brazilian Portuguese nominalizations similar to the English NomLex. In addition to NomLex, we also made use of the resources like French Nomage [8] and the collaborative dictionary Wiktionary in creating our lexicon. NomLex-PT is now a medium-sized resource that has been integrated with OpenWordNet-PT and is in the process of helping to improve the bigger resource.

Despite this growth, one may raise concerns that the multilingual resources used in bootstrapping our NomLex resource might have limited its scope to the common ground between the involved languages, without doing proper justice to the characteristic features of Portuguese. To address these concerns, we decided to make use of the rich linguistic data available from the AC/DC project [9] to increase our stock of nominalizations. In particular, the project provides publicly available Portuguese language corpora with morpho-syntactic annotations. While our original version focused on Brazilian Portuguese and was initially called NomLex-BR, we call the new resulting lexicon NomLex-PT, as the corpus has Portuguese texts from Portugal and Mozambique as well as from Brazil.

In this paper, we describe the extraction of data from the corpus and its hand curation, as well as the kind of evaluation we were able to perform and the applications we envisage for our resources. We also take steps to ensure that the new entries are aligned with OpenWordNet-PT, which, given its use of the RDF standard, enables an easier interchange of data with the large body of linguistic resources that have already been published as Linked Data in the Semantic Web.

2 Nominalizations in NomLex-PT and OpenWordNet-PT

Deverbal nouns, or nominalizations, can pose serious challenges for knowledge representation systems. A sentence like *Alexander destroyed the city in 332 BC* is easily parsed and its semantic arguments, such as the agent (Alexander), the thing destroyed (the city) and the time (332 BC), are readily obtained for a proposed logical representation of the sentence. By contrast, a sentence like *Alexander's destruction of the city took place in 332 BC* is much harder to interpret for many computational systems. It describes the same event of destruction, with the same semantic arguments, but these arguments are harder to obtain automatically from a syntactic parsing of the sentence. Hence it is advantageous to have a lexicon that marks such relatedness of meaning between the verbs and their nominalizations.

Deverbal nouns are well-studied in English, with the NomLex project [2] providing a well-established, open access baseline for corresponding results in other languages. Previous work has shown that NomLex is useful for natural language and knowledge representation tasks [10].

To construct a Portuguese lexicon of nominalizations, we initially started with a manual translation of the entries in the original NomLex. This gave us many of the most straightforward nominals in Portuguese, e.g. *construction/construção* and *writer/escritor*.

We then used cross-lingual projection methods to suggest additional candidates using resources like the French lexicon NOMAGE [8], the collaborative dictionary Wiktionary and its Portuguese version Wikcionário, as well as the FrameNet lexical resource [5]. These candidates were manually evaluated for correctness by two native speakers of Portuguese. Table 1 recalls the provenance of the nominals in our lexicon.

Table 1. Provenance of entries in NomLex-PT

| Source | Number of Entries |
|---------------------------------|-------------------|
| NomLex Translations | 1017 |
| AC/DC Additions | 1261 |
| PAPEL Additions | 458 |
| Manual Additions | 709 |
| NOMAGE Translations | 262 |
| DHBB-Motivated Manual Additions | 158 |
| Wiktionary | 152 |
| FrameNet | 142 |
| OpenWordNet-PT | 82 |
| Total | 4,241 |

Finally, as reported in our previous work [11] and [12], we integrated all of this data with our Portuguese wordnet. That resource, OpenWordNet-PT, started as a “projection” of the much more ambitious Universal WordNet (UWN) project, developed by Gerard de Melo and Gerhard Weikum [13]. The universal wordnet UWN extends the original Princeton WordNet with around 1.5 million meaning links for 800,000 words in over 200 languages, using data from Wiktionary, Wikipedia, and various other sources. OpenWordNet-PT is based on the Portuguese part of UWN but adds numerous manual revisions and new Portuguese-language glosses.

3 Corpus-Based Nominalization Detection

In this paper, we use a different technique to increase the stock of nominalizations in our database. We rely on text corpora, as explained below, to identify and compile lists of likely nominalization/verb pairs. These lists are then manually checked before being added to the inventory.

3.1 Gathering Data: The AC/DC Project

The AC/DC project was launched in 1998, with the goal of providing a place where it is possible to freely access Portuguese corpora. The acronym AC/DC

stands for *Acesso a Corpora/Disponibilização de Corpora* (loosely translated as ‘access and availability of corpora’), and its main purpose is ‘to make a large number of corpus resources in Portuguese available on the web with a unified and simple interface that allows people to interact with the corpora without requiring physical access to institutions or software’, as [14] reminds us.

The AC/DC cluster contains more than 1 billion words, distributed over the following genre parameters, listed in the order of their frequency 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, and oral. In terms of language varieties, it contains Portuguese spoken in Portugal, Brazil, and Mozambique. For a detailed description of the Linguateca repository and the AC/DC project, the reader is referred to [15]. All AC/DC material was automatically parsed using the PALAVRAS parser [16]. The data thus comes with a range of linguistic information, including lemma, part-of-speech, morphology (gender, number, tense form, etc), and syntactical function.

3.2 Approach

Portuguese has specific suffixes that indicate the morphological relationship between nouns and verbs, such as *-ção*, *-agem* or *-or*⁷. Of these, some also convey, quite regularly, the semantic notion of agent or instrument, like the suffix *-or* or *-ura*. In light of these considerations, we query the corpora that make up the AC/DC cluster to find lists with instances of nominalization candidates that end with the chosen suffixes.

However, these suffix-based clues, despite being highly regular, are not infallible. Considering the suffix *-ção*, for example, while there is *relação-relacionar / relation-relate* or *produção-produzir / production-produce*, the pair *ignição-*ignar* does not exist, and many other nouns ending in *-ção* have no corresponding verb. Therefore a careful review of the lists generated via suffixes is imperative.

At this stage of NomLex-PT, we considered the following five suffixes, which we describe with examples both in Portuguese and English: *-ção* (and its allomorph *-ização*), for example *padronizar-padronização / standard-standardization*; *-ncia* (as in *dominar-dominância / dominate-dominance*); *-agem* an example that does not change form in English (*reciclar-reciclagem / recycle-recycle*); and the suffixes for agentives *-or* (as in *trabalhar-trabalhador / work-worker*); and *-nte* as in (*estudar-estudante / study-student*).

Through simple morphological rules, we produced the candidate verbs and then the noun-verb pairs were manually evaluated by two annotators. The annotators not only checked the pairs for correctness, but also added semantic markers to the nominalizations, a process we describe in great detail below in Section 4.

⁷ The suffix *-or*, the most common agentive suffix, has two allomorphs: *-tor* (*construtor/constructor*) and *-dor* (*esfaqueador/*knifer, someone who attacks with a knife*).

3.3 Extension for Regressive and Other Nominalizations

In addition to the suffix-based search, we also considered regressive nominalizations, also known as zero-derivation nominalizations. It is well-known that these make up a great number of Portuguese nominals (for example *atacar-ataque/attack-attack*; *lutar-luta/fight-fight*), see for instance [17]. However, this morphological process is not accompanied by an easily discernible morphological trace, hence some linguists calling it a zero derivation. This hinders any corpus search, as obviously the clue “ends with a vowel” is a very weak one and other tools are required to bear on the problem.

Thus we turned our attention to PAPEL [18], a lexical resource for NLP of Portuguese, based on the (semi-)automatic extraction of semantic relations between the words from a general language dictionary. PAPEL makes available lists of word and tuples grouped by semantic relation. Since PAPEL was created semi-automatically, it is reasonable to assume that the lists contain mistakes. However, there are no errors such as non-words. Obtained from a dictionary, all listed words do exist in Portuguese. Given that we had already captured a considerable amount of nominal-verb pairs from the corpus, we used the following strategy: Only the verb-noun pairs that did not contain the already analyzed suffixes went under revision by two linguists. The results were then reviewed, just as the previous ones, by two linguists.

4 Linguistic Annotation

While links between nominalizations and the verbs they are derived from are useful, natural language processing systems often require additional information to interpret them correctly. While reviewing the noun-verb pairs, our annotators thus also added additional annotations to assess the nominalizations in terms of agentivity, animacy, and the possibility of lexicalization.

4.1 Agentives

Agentive nominals were defined following a grammatical criterion. Agentives are nouns for which the paraphrase *one that Xs/ o que Xs* (where *X* is the verb) is possible. So, *pintor (painter)* is *one that paints/o que pinta*; *estudante (student)*, is *one who studies/o que estuda*; *computador (computer)* is *one that computes/o que computa*⁸. The values of the feature ‘agentive’ are simply Y (yes) or N (no). The criterion for identification is only formal, so *democratização (democratisation)* (or *globalização-globalisation*), for example, in *A democratização da mídia permitiu a emergência do fenômeno (the democratisation of the media allowed*

⁸ Note that in Portuguese, as in most Romance languages, unlike English, the paraphrase does not have to decide a priori on the animacy (or not) of the subject: *computers/computadores* and *fans/ventiladores* are treated as ‘he’, the same pronoun used for humans and we do not have to choose between ‘who’ and ‘which’, ‘that’ works for all.

the emergence of the phenomenon), although it may, from one point of view, be compatible with the idea of an agent, it does not match the paraphrase “*o que democratiza*” (*one that democratizes*), therefore it is not considered agentive. For these cases, the agentive nominal would be *democratizador* (or *globalizador*). We should keep in mind that most of the nominalization suffixes are polysemous, so a nominal such as *democratização* (*democratisation*) has many meanings. Even if we can use it somehow as an agent nominal, its main use is still connected to the eventive reading, the process of *democratizar* (*democratize*).

4.2 Animacy

The values for the feature animacy are: A (animate), I (inanimate), or Both (underspecified). For instance, *pintor/painter* is animate, *utilização/utilization* is inanimate, and the label ‘Both’ applies to nouns that can be used as either, see *recolhedor/collector* below. Animacy is known to be not as straightforward as it might appear, see the discussion in [19].

4.3 Lexicalization

Lexicalized nominals are those in which the actual meaning of the word no longer corresponds to the meaning resulting from the morphological process: *procuração* (the noun for ‘power of attorney’ in English) does not mean the act of *procurar/to search*, so we would say it is a lexicalized nominal; *coordenação* (*coordination*) corresponds both (a) to the act of *coordenar/to coordinate* as well as as to (b) the entity/group of people in charge of the coordination, the latter being a lexicalized noun. Examples:

- a) A coordenação da pesquisa foi feita pelo departamento. The coordination of the research was done by the department.
- b) A coordenação decidiu não aceitar a proposta. The coordination decided not to accept the proposal.

It is not always easy to decide whether we are facing a lexicalized nominal or not. The lexicalization process would be better described as a continuum, and some words can be more lexicalized than others. As an example, it can be difficult to decide if *educação/education* is lexicalized in *educação de jovens e adultos/youth and adult education*, since *jovens e adultos/youth and adults* can be either interpreted as arguments of the verb *educar/to educate* or as noun modifiers. On the other hand, *procuração/‘power of attorney’* is clearly a noun that does not mean the act of *procurar/to search*. Besides, words can often have a lexicalized equivalent in a specific domain: *ocorrência/occurrence*, the event form of the verb *ocorrer/to occur*, is a deverbal nominal in *ocorrência de chuvas/occurrence of rain* but a lexicalized nominal in *ocorrência policial/police incident*. The values we assigned to the lexicalization feature are 0 (non-lexicalized), 1 (lexicalized), and B (both).

We assigned *procuração* 1; for *ocultação/concealment* we assigned 0; and for *declaração/declaration* we assigned B, since most speakers of Portuguese will know *declaração* both as an official document, as well as the action of the verb *declarar*. For nouns such as *civilização/civilization*, *organização/organization*, *coordenação/coordination*, and *vigilância/surveillance*, where the lexicalized version is considered equivalent to the group of people associated, we will consider the nominal *animate*, but they will not be agentive. The sentences below illustrate some cases as well as the importance of considering the sentence context:

- (c) Até os profissionais da área de *vigilância* temem uma nova escalada de violência. Even the *surveillance* professionals fear a new climbing of the violence. (The nominal *vigilância/surveillance* is considered lexicalized and inanimate.)
- (d) Em outro ponto do galpão, a *vigilância* encontrou alimentos como molho de tomate, arroz, bolacha, margarina e macarrão com prazo de validade vencido. In another area of the deposit, the *surveillance* (team) found provisions such as tomato sauce, rice, crackers, margarine and pasta with past sellby dates. (The same nominal is lexicalized and animate.)

4.4 Issues and Choices

By far, lexicalization was the most challenging point during our review process, but the *-or* suffix added some extra challenges. First of all, in order to generate the list of words ending by the suffix *-or*, we needed the syntactic annotation provided by the parser PALAVRAS applied to the AC/DC corpora. Since many *-or* nouns can also be used as adjectives, we refined the corpus search to search only for words that were classified as nouns. The search results still needed some revision, but the strategy saved us some time.

It was also crucial to check our *-or* candidates against their occurrences in corpus, in order to

- (i) determine their use as animate/inanimate, and
- (ii) verify any special context/domain in which the word might occur as a result of a lexicalization process.

Concerning the issue of animacy (i), we were informed by the corpora about the double nature (animate and inanimate) of nominals such as *recolhedor* and *desviador*, for example:

- (e) A limpeza do córrego está sendo realizada com barreiras para evitar a disseminação do óleo, com equipamentos de sucção e com *recolhedores* de óleo. The cleaning of the creek is done with barriers to avoid dissemination of the oil, with suction equipment and with *collectors* of oil. (inanimate)
- (f) Primeiro, trabalhou como *recolhedor* das apostas e há três anos teria ganho alguns pontos de jogo na zona norte. First (he) worked as a bet *collector* and three years ago (he) was supposed to have won some game spots in the North Zone. (animate)

- (g) A armadura de protensão não aderente, disposta em elementos fictícios, paralelos aos elementos reais da estrutura, tendo em comum apenas os nós de extremidade da peça estrutural e os nós localizados nos *desviadores*. The shell of adhering protension, disposed in fictitious elements, parallel to the real elements of the structure, had in common only the nodes of the extremities of the structural piece and the nodes located on the *routers*. (inanimate)
- (h) Na presidência do Senado talvez o maior bandido do Brasil e o maior *desviador* de dinheiro público. In the Senate presidency, [we had] perhaps the biggest thief of Brazil and the biggest *swindler* of public funds. (animate)

Concerning the process of lexicalization (ii), the corpora showed us specialized uses we were not aware of initially. The observation of domain-specific uses was facilitated by the very nature of the AC/DC cluster. On the one hand, the Corpus Brasileiro (Brazilian Corpus) contains technical material; on the other hand, the presence of texts from both Brazilian and Portuguese varieties helped us with lexical differences.

- (j) Valores médios de pH e acidez dos pães de forma produzidos com diferentes concentrações de *melhoradores* naturais (In, Mi e ML) e do melhorador comercial (MC). Average values of pH and acidity of sliced breads produced from different concentrations of natural and commercial *improvers*.
- (k) Outro aliado de Gullit, Rijkaard, o *centralizador* dos lances de meio-campo. Another ally, Gullit, Rijkaard, the centralizer of midfield bids.

5 Results and Discussion

Table 2 shows the number of nominalizations (lemmas) after revision and their distribution by suffixes.

Table 2. Nominalizations in NomLex-PT by suffix

| Suffix | Nominalizations |
|--------|-----------------|
| -ção | 1345 |
| -mento | 334 |
| -ida | 24 |
| -ura | 107 |
| -or | 896 |
| -nte | 115 |
| -ada | 115 |
| other | 1302 |

There is no other large lexicon of Portuguese nominalizations, freely available on the web, that we can compare our results with. As previously described, we

started our project by first translating nominalizations from the English NomLex lexicon, followed by translating the French Nomage ones and gradually adding additional sources of data. Concerned that the translations might be missing the most used nominalizations in Portuguese, we had previously conducted a small coverage-focused experiment on a corpus of biographical stories. Subsequently, as reported in this paper, we decided to go for a full experiment on extracting nominalizations from the very large AC/DC corpora.

Lists of nominalization-verb candidate pairs were automatically produced and then manually checked by two linguists, who discussed difficult cases (some described in the paper) and reached a consensus. Since the process of reviewing and classifying the nominalizations was ultimately completely manual, traditional measures of precision and recall do not seem appropriate. We are considering other kinds of measures to test for coverage and accuracy of the classifications obtained.

For instance, we only became aware of the Spanish ANCORA lexicon recently and we intend to check the intersection of its coverage with ours. Another future work we are considering is extracting verbal subcategorization frames from the corpus, as also suggested by one of our reviewers. We are keenly aware that a sophisticated verb lexicon is a necessity for work on linguistic inference but this is a sizeable project, which would benefit substantially if other verbal lexica of Portuguese were freely available.

6 Conclusions

Using a combination of manual and automated methods and drawing on existing resources, we have created a representative lexicon of nominalizations in Portuguese that we hope is useful for linguists as well as for computational applications. 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 [20]. The 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 [21] could accomplish for English. Many other components remain to be designed and implemented.

The approach taken here, populating a Portuguese nominalization lexicon with data extracted from a collection of corpora, adds a bottom-up dimension to a process that, so far, had been done in a top-down fashion. We do not start here from the (English or French-based) dictionaries, but from everyday Portuguese words. This has the good effect of making sure that we list roots and lemmas that are typically Portuguese (e.g. *esfaqueador*, *sambista*) and as a by-product, since it is performed by different linguists, re-checks the nominalizations already in place. We also retrieve words that are hardly present in dictionaries, such as slang and new constructions, which seem to have emerged in the language to describe specific new situations. Having both the top-down and the bottom-up approaches seems an appropriate way to obtain a balanced lexicon.

Given that the lists of nominals we discussed have a wealth of information that we have not been able to fully operationalize, we are considering making these lists publicly downloadable as well as the lexicon. Our NomLex-PT data is available from <https://github.com/arademaker/nomlex-pt>, and we look forward to discussions on how to improve this resource as well as OpenWordNet-PT.

Much work remains to be done to increase the supply of freely available and robust semantic oriented lexical resources and tools for Portuguese. There is, however, a growing amount of available resources of which we have made heavy use, and to which we can now add NomLex-PT.

References

- [1] Fellbaum, C.: WordNet: An electronic lexical database. The MIT press (1998)
- [2] Macleod, C., Grishman, R., Meyers, A., Barrett, L., Reeves, R.: Nomlex: A lexicon of nominalizations. In: In Proceedings of Euralex98. (1998) 187–193
- [3] Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., Grishman, R.: The nombank project: An interim report. In Meyers, A., ed.: HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation, Boston, Massachusetts, USA, Association for Computational Linguistics (May 2 - May 7 2004) 24–31
- [4] Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: Extending verbnet with novel verb classes. In: Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy (June 2006)
- [5] Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proc. COLING-ACL 1998. (1998) 86–90
- [6] Santos, D., Barreiro, A., Freitas, C., Gonçalo Oliveira, H., Medeiros, J.C., Costa, L., Gomes, P., Silva, R.: Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In: Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística. APL 2009. APL (2010) 681–700
- [7] de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: An open Brazilian wordnet for reasoning. (2012)
- [8] Balvet, A., Barque, L., Condette, M.H., Haas, P., Huyghe, R., Marn, R., Merlo, A.: La ressource Nomage: Confronter les attentes thoriques aux observations du comportement linguistique des nominalisations en corpus. TAL **52**(3) (2011) 124
- [9] Costa, L., Santos, D., Rocha, P.A.: Estudando o portugus tal como usado: o servio AC/DC. In: The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009). (8-11 de Setembro 2009)
- [10] Gurevich, O., Crouch, R., King, T.H., de Paiva, V.: Deverbal nouns in knowledge representation. Journal of Logic and Computation (2008)
- [11] Valeria de Paiva, Livy Real, A.R., de Melo, G.: NomLex-BR: A lexicon of portuguese nominalizations. In: Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014), Paris, France, ELRA (2014)
- [12] 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, Tartu, Estonia (jan 2014) 378–382
- [13] de Melo, G., Weikum, G.: Towards a universal wordnet by learning from combined evidence. In: Proceedings of the 18th ACM Conference on Information

- and Knowledge Management (CIKM 2009), New York, NY, USA, ACM (2009) 513–522
- [14] Costa, L., Santos, D., Rocha, P.A.: Estudando o português tal como é usado: o serviço AC/DC. In: 7th Brazilian Symposium in Information and Human Language Technology – STIL 2009, São Carlos, Brazil (2009)
- [15] Santos, D., Barreiro, A., Freitas, C., Oliveira, H.G., Medeiros, J.C., Costa, L., Gomes, P., Silva, R.: Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In: XXV Encontro Nacional da Associação Portuguesa de Linguística, Porto, Portugal, APL (2009) 681–700
- [16] Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Aarhus University (2000)
- [17] Basílio, M.: Estruturas lexicais do português; uma abordagem gerativa. Ática (1980)
- [18] Gonçalo Oliveira, H., Santos, D., Gomes, P., Seco, N.: 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., Aveiro, Portugal, Springer (September 2008) 31–40
- [19] Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M., Wasow, T.: Animacy encoding in english: why and how. In: Proc. of the Association for Computational Linguistics Workshop on Discourse Annotation. (2004)
- [20] Gurevich, O., Crouch, D., King, T.H., de Paiva, V.: Deverbal nouns in knowledge representation. FLAIRS: The Florida Artificial Intelligence Research Society (May 2006)
- [21] Bobrow, D.G., Cheslow, B., Condoravdi, C., Karttunen, L., King, T.H., Nairn, R., de Paiva, V., Price, C., Zaenen, A.: PARC's bridge and question answering system. In: Proceedings of Grammar Engineering Across Frameworks. (2007) 26–45