Body-Part Nouns and Whole-Part Relations in Portuguese

Ilia Markov[1,3], Nuno Mamede[2,3], and Jorge Baptista[1,3]

[1] Universidade do Algarve/FCHS and CECL,
Campus de Gambelas, 8005-139 Faro, Portugal
jbaptis@ualg.pt

[2] Instituto Superior Técnico, Universidade de Lisboa,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
Nuno.Mamede@tecnico.ulisboa.pt

[3] INESC-ID Lisboa/L2F – Spoken Language Lab,
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
{Ilia.Markov,jbaptis}@l2f.inesc-id.pt

Abstract. In this paper, we target the extraction of *whole-part* relations involving human entities and *body-part nouns* occurrences in texts using STRING, a hybrid statistical and rule-based Natural Language Processing chain for Portuguese. Whole-part relation is a semantic relation between an entity that is perceived as a constituent part of another entity, or a member of a set.

Keywords: Whole-part relation, meronymy, body-part noun, disease noun, Portuguese.

1 Introduction

Automatic identification of semantic relations is an important step in extracting meaning out of texts, which may help several other Natural Language Processing (NLP) tasks, such as question answering, text summarization, machine translation, information extraction, information retrieval and others [13]. The whole-part relations acquired from a collection of documents are used in answering questions that normally cannot be handled based solely on keywords matching and proximity [14]. For automatic text summarization, where the most important information from a document or set of documents is extracted, semantic relations are useful for identifying related concepts and statements, so a document can be compressed [19].

The goal of this work is to improve the extraction of semantic relations between textual elements in STRING, a hybrid statistical and rule-based NLP chain for Portuguese [1][21]. This work will target whole-part relations (*meronymy*), that is, a semantic relation between an entity that is perceived as a constituent part of another entity, or a member of a set. In this case, we focus on the type of

[1] https://string.l2f.inesc-id.pt/ [last access: 04/05/2014].

J. Baptista et al. (Eds.): PROPOR 2014, LNAI 8775, pp. 125–136, 2014.

meronymy involving human entities and *body-part nouns* (henceforward, *Nbp*) when they co-occur in texts. Though STRING already extracts some types of semantic relations, meronymic relations are not yet being detected, in spite of the large set of *Nbp* that have already been semantically tagged in its lexicon. In other words, we expect to enhance the system's semantic relations extraction module by capturing meronymic relations. This paper is structured as follows: Section 2 briefly describes related work on whole-part dependencies extraction, while Section 3 explains with some detail how this task was implemented in STRING; Section 4 presents the evaluation procedure; and Section 5 draws the conclusions.

2 Related Work

Meronymy is a complex relation that "should be treated as a collection of relations, not as a single relation" [18]. In NLP, various information extraction techniques have been developed in order to capture whole-part relations from texts.

Hearst [16] tried to find lexical correlates to the *hyponymic* relations (type-of relations) by searching in unrestricted, domain-independent text for cases where known hyponyms appear in proximity. The author proposed six lexico-syntactic patterns; he then tested the patterns for validity, and used them to extract relations from a corpus. To validate his acquisition method, the author compared the results of the algorithm with information found in WordNet [9]. The author reports that when the set of 152 relations that fit the restrictions of the experiment (both the hyponyms and the hypernyms are unmodified) was looked up in WordNet:

> "180 out of the 226 unique words involved in the relations actually existed in the hierarchy, and 61 out of the 106 feasible relations (*i.e.*, relations in which both terms were already registered in WordNet) were found." [16, p. 544].

The author claims that he tried applying the same technique to meronymy, but without great success.

Berland and Charniak [5] addressed the acquisition of meronyms using manually-crafted patterns, similar to Hearst [16], in order to capture textual elements that denote whole objects (*e.g.*, *building*) and then to harvest possible part objects (*e.g.*, *room*). The authors used the North American News Corpus (NANC) - a compilation of the wire output of a certain number of newspapers; the corpus is about 1 million words. Their systems output was an ordered list of possible parts according to some statistical metrics. They report that their method finds parts with 55% accuracy for the top 50 words ranked by the system and a maximum accuracy of 70% over their top-20 results. The authors report that they came across various problems, such as tagger mistakes, idiomatic phrases, and sparse data – the source of most of the noise.

Girju *et al.* [13,14] present a supervised, domain independent approach for the automatic detection of whole-part relations in text. The algorithm identifies lexico-syntactic patterns that encode whole-part relations. The authors report an overall average precision of 80.95% and recall of 75.91%. The authors also state that they came across a large number of difficulties due to the highly ambiguous nature of syntactic constructions.

Van Hage *et al.* [15] developed a method for learning whole-part relations from vocabularies and text sources. The authors' method learns whole-part relations by

> "first learning phrase patterns that connect parts to wholes from a training set of known part-whole pairs using a search engine, and then applying the patterns to find new part-whole relations, again using a search engine." [15, p. 30].

The authors reported that they were able to acquire 503 whole-part pairs from the AGROVOC Thesaurus[2] to learn 91 reliable whole-part patterns. They changed the patterns' part arguments with known entities to introduce web-search queries. Corresponding whole entities were then extracted from documents in the query results, with a precision of 74%.

The Espresso algorithm [27] was developed in order to harvest semantic relations in a text. Espresso is based on the framework adopted in Hearst [16]:

> "It is a minimally supervised bootstrapping algorithm that takes as input a few seed instances of a particular relation and iteratively learns surface patterns to extract more instances." [27, § 3].

Thus, the algorithm extracts surface patterns by connecting the seeds (tuples) in a given corpus. The algorithm obtains a precision of 80% in learning whole-part relations from the Acquaint (TREC-9) newswire text collection, with almost 6 million words.

Thereby, for the English language, it appears that the acquisition of whole-part relation pairs by way of machine-learning techniques achieves fairly good results.

According to the very recent review of the literature on semantic relations extraction [1], no works on whole-part relations extraction for Portuguese have been identified. The current work also aims at extracting a specific type of whole-part relations, involving *Nbp*, but we adopt a rule-based approach, using the tools and resources available in STRING.

Next, in this work, we focus on state-of-the-art relations extraction in Portuguese, in the scope of ontology building. Some work has already been done on building *knowledge bases* for Portuguese, most of which include the concept of whole-part relations. These knowledge bases are often referred to as *lexical ontologies*, because they have properties of a lexicon as well as properties of an ontology [17,31]. Well-known, existing lexical ontologies for Portuguese are Portuguese WordNet.PT [22,23], later extended to WordNet.PT Global (Rede Léxico-Conceptual das Variedades do Português) [24]; MWN.PT-MultiWordNet

[2] `http://www.fao.org/agrovoc` [last access: 19.02.2014].

of Portuguese [30]; PAPEL (Palavras Associadas Porto Editora Linguateca) [26]; and Onto.PT [25]. Some of these ontologies are not freely available for the general public, while others just provide the definitions associated to each lexical entry without the information on whole-part relations. Furthermore, the type of whole-part relation targeted in this work, involving any human entity and its related *Nbp*, can not be adequately captured using those resources (or, at least, only those resources).

Attention was also paid to two well-known parsers of Portuguese, in order to discern how do they handle the whole-part relations extraction: the PALAVRAS parser [6], consulted using the Visual Interactive Syntax Learning (*VISL*) environment, and LX Semantic Role Labeller [7]. Judging from the available online versions/demos of these systems, apparently, none of these parsers extracts whole-part relations, at least explicitly.

In conclusion, the available resources identify some whole-part relations between lexical items in Portuguese, but they are not sufficient for the task of automatic extraction of whole-part relations as they occur between texts' instances of human entities and body-part nouns, as we here have targeted. Furthermore, we adopt a rule-based approach in order to extract this kind of relations from texts, considering the NLP system in which we intend to implement this module.

3 Whole-Part Dependency Extraction Module in STRING

3.1 STRING Overview

STRING performs all the basic steps of natural language processing (tokenization, sentence splitting, POS-tagging, POS-disambiguation and parsing) for Portuguese texts. The architecture of STRING is given in Fig. 1.

STRING has a modular, pipe-line structure, where: (i) the preprocessing stage (tokenization, sentence splitting, text normalization) and lexical analysis are performed by LexMan; (ii) followed by RuDriCo, which applies disambiguation rules, handles contractions and several special types of compound words; (iii) the

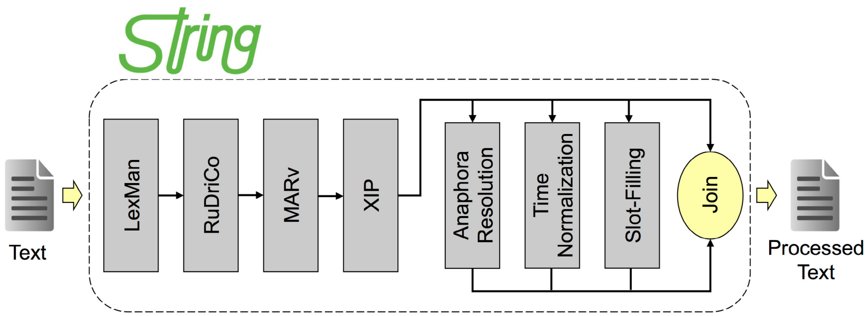

Fig. 1. STRING Architecture

MARv module then performs POS-disambiguation, using HMM and the Viterbi algorithm; and, finally, (iv) the XIP rule-based parser (Xerox Incremental Parser) [2] segments sentences into chunks and extracts dependency relations among chunks' heads. XIP also performs named entities recognition (NER). A set of post-parser modules have also been developed to handle certain NLP tasks such as anaphora resolution, temporal expressions' normalization, and slot-filling.

As part of the parsing process, XIP extracts *dependencies*. These dependencies correspond to different syntactic-semantic relations between the nodes of the sentence chunking tree, namely, the chunks' heads. Dependencies can thus be viewed as equivalent to (or representing) the syntactic relations holding between different elements in a sentence. Some of the dependencies extracted by XIP represent rather complex relations, such as the notion of *subject* (`SUBJ`) or *direct object* (`CDIR`), which imply a higher level of analysis of a given sentence. Other dependencies are much simpler and sometimes quite straightforward, like the determinative dependency `DETD`, holding between an article and the noun it determines, *e.g.*, *o livro* 'the book' – `DETD(livro,o)`. Some dependencies can also be seen as auxiliary dependencies, and are required to build the more complex ones.

3.2 A Whole-Part Extraction Module in STRING

Next, we describe the way some of whole-part dependencies involving *Nbp* are extracted in the Portuguese grammar for XIP. To this end, a new module of the rule-based grammar was built, which is the first step towards a meronymy extraction module for Portuguese, and it contains most of the rules required for this work. In order to better present the different syntactic-semantic situations that the meronymy extraction module targets, some of the more simple cases are illustrated first and then some of the more complex situations follow. Example (1) is a simple case where there is a determinative `PP`, complement *de* 'of' N of the *Nbp*, so that the meronymy is overtly expressed in the text:

(1) *O Pedro partiu o braço do João* 'Pedro broke the arm of João'

The next rule captures the meronymy relation between *João* and *braço* 'arm':

```
IF( MOD[POST](#2[UMB-Anatomical-human],#1[human]) & PREPD(#1,?[lemma:de]) &
    CDIR[POST](#3,#2) & ~WHOLE-PART(#1,#2) )
    WHOLE-PART(#1,#2)
```

This rule is built using the XIP dependency rules' syntax, and it reads as follows: first, the parser determines the existence of a `[MOD]`ifier dependency, already calculated, between an *Nbp* (variable `#2`) and a human noun (variable `#1`); notice that, according to XIP conventions, the governor of the dependency is its first argument, hence *João* is said to be a modifier of *braço* 'arm'; this modifier must also be introduced by preposition *de* 'of', which is expressed by the dependency `PREPD`; then, a constraint is defined that the *Nbp* must be a direct object (`CDIR`) of a given verb (variable `#3`); and, finally, that there is still no previously calculated `WHOLE-PART` dependency between the *Nbp* and the human noun (variable `#1`); this last constraint is meant to ensure that there is only one meronymy relation between each *Nbp* and a given noun; if all these conditions are met, then, the parser builds the `WHOLE-PART` relation between the human determinative complement and the *Nbp*.

Next, in example (2), we present the (apparently) more simple case of a sentence with just a human subject and an *Nbp* direct object:

(2) *O Pedro partiu um braço* 'Pedro broke an arm'

In Portuguese, in the absence of a determinative complement, a possessive determiner or a dative complement (eventually reduced to a clitic dative pronoun), sentences like (2) are preferably interpreted as holding a whole-part relation between the human subject and the object *Nbp*. Thus, if there is a subject and a direct complement dependency holding between a verb and a human, on one side, and the verb and an *Nbp*, respectively; and if no `WHOLE-PART` dependency has yet been extracted for that *Nbp*, either for that human subject or another element in the same sentence, then the `WHOLE-PART` dependency is extracted.

There may be a relation within the same sentence between different *Nbp*, like in example (3). In this case, the `WHOLE-PART` relation should be established not only between the subject of the sentence and the *Nbp*, but also between *Nbp* in the sentence.

(3) *A Ana pinta as unhas dos pés* (lit: Ana paints the nails of the feet)
'Ana paints the toenails'

In example (3), there is a meronymic relation between *Ana* and *unhas* 'nails', but also between *pés* 'feet' and *unhas* 'nails', so that two `WHOLE-PART` relations should be extracted.

There may be also a relation within the same sentence between an *Nbp* and a noun that designates a part of that same *Nbp*, and which we will call *npart* (*ponta da língua* 'tip of the tongue', *palma da mão* 'palm', *etc.*). This case differs from the previous one because, on the one hand, the whole-part relation should be established between the human noun and the *Nbp* and **not** the *npart* that precedes it; and, on the other hand, a second whole-part relation should also be established between the determinative *npart* and the *Nbp*, although this *npart* is not, by itself, an *Nbp*. Example (4) illustrates this situation:

(4) *O Pedro tocou com a ponta da língua no gelado da Ana*
'Pedro touched with the tip of the tongue the ice cream of Ana'

`WHOLE-PART(Pedro,língua)` - correct; `WHOLE-PART(língua,ponta)` - correct; `WHOLE-PART(Pedro,ponta)` - incorrect.

The set of *npart* varies according to the *Nbp* and each set has to be established a priori. For example, for the *Nbp pé* 'foot' we can include the nouns *peito* 'instep', *alto* 'top', *cova or arco* 'arch', *dorso* 'instep', *planta* 'sole', and *ponta* 'tiptoe'. This is done by way of rules that add the feature *npart* to the nouns in the set associated to each *Nbp*, in the context of a determinative complement *de N* 'of N' of that *Nbp*. So far, 54 rules were built to associate the *Nbp* with their parts. All in all, 27 general rules have been built and implemented in STRING in order to extract whole-part relations involving *Nbp*.

We now turn to another type of meronymic relation. In some cases, a whole-part relation is only implicit, and though *Nbp* are involved, they are not mentioned directly (*gastritis*-'stomach'). In these cases, we decided that, nevertheless, a whole-part relation between the human entity and the "hidden" *Nbp* should be established. At this time, we focus on some predicative nouns designating specially localized *diseases* (*Nsick*). High lexical constraints apply in this relation: for each of these disease predicative nouns, the specific *Nbp* that is involved must be explicitly indicated in the lexicon. For example, the case where a disease noun is built with the support verb *ter* 'have', example (5):

(5) *O Pedro tem uma gastrite* 'Pedro has gastritis'

The rule that captures the meronymy relation between *Pedro* and *estômago* 'stomach' is given below:

```
IF( CDIR[POST](#1[lemma:ter],#2[lemma:gastrite]) & SUBJ(#1,#3) & ~WHOLE-PART(#3,?) )
    WHOLE-PART[hidden=+](#3,##noun#[surface:estômago,lemma:estômago])
```

The rule itself reads as follows: first, the system checks if the disease noun (in this case, *gastrite* 'gastritis') is the direct object (`CDIR`) of the verb *ter* 'have' (variable `#1`); secondly, the system verifies if there is an explicit subject (variable `#3`) for the verb; and if there is still no `WHOLE-PART` relation between that subject and the other node; in this case, the system builds the `WHOLE-PART` dependency between the subject of the verb and the 'hidden' *Nbp*, for which it creates a new (dummy) `noun` node. To express that a "hidden" noun is involved in this relation, a special tag '`hidden`' is also introduced in the dependency.

So far, 29 different pairs (*disease nouns*, *Nbp*) have been encoded in the lexicon.

To conclude, we have also addressed the issue of ambiguity raised by idioms involving *Nbp*. As it is well known, there are many frozen sentences (or idioms) that include *Nbp*. However, for the overall meaning of these expressions, the whole-part relation is often irrelevant, as in the next example:

(6) *O Pedro perdeu a cabeça* (lit: Pedro lost the [=his] head) 'Pedro got mad'

The overall meaning of this expression has nothing to do with the *Nbp*, so that, even though we may consider a whole-part relation between *Pedro* and *cabeça* 'head', this has no bearing on the semantic representation of the sentence, equivalent in (6) to 'get mad'. STRING's strategy to deal with this situation is, first, to capture frozen or fixed sentences, and then, after building all whole-part dependencies, exclude/remove only those containing elements that were also involved in fixed sentences' dependencies. In this way, two general modules, for fixed sentences and whole-part relations, can be independently built, while a simple "cleaning" rule removes the cases where meronymy relation is irrelevant.

Frozen sentences are initially parsed as any ordinary sentence, and then the idiomatic expression is captured by a special dependency (`FIXED`), which takes as its arguments the main lexical items of the idiom. The number of arguments varies according to the type of idiom. In the example (6) above, this corresponds to the dependency: `FIXED(perdeu,cabeça)`, which is captured by the following rule:

```
IF (VDOMAIN(?,#2[lemma:perder]) & CDIR[post](#2,#3[surface:cabeça])) FIXED(#2,#3)
```

This rule captures any `VDOMAIN`, that is, a verbal chain of auxiliaries and the main verb whose lemma is *perder* 'loose', and a post-positioned direct complement whose head is the surface form *cabeça* 'head'.

In order to capture the idioms involving *Nbp*, we built about 400 of such rules[4], from 10 formal classes of idioms [3].

4 Evaluation

4.1 Evaluation Corpus

The 1st fragment of the CETEMPúblico corpus [32] was used in order to extract sentences that involve *Nbp*. This fragment of the corpus contains 14,715,055 tokens (147,567 types), 6,256,032 (147,511 different) simple words and 260,943 sentences. The existing STRING lexicon of *Nbp* and *Nsick* was adapted to to be used within the UNITEX corpus processor [28,29] along with the remaining available resources for European Portuguese, distributed with the system.

Using the *Nbp* (151 lemmas) and the *Nsick* (29 lemmas) dictionaries, 16,746 *Nbp* and 79 *Nsick* instances were extracted from the corpus (excluding the ambiguous noun *pelo* 'hair' or 'by-the', which did not appeared as an *Nbp* in this fragment). Some of these sentences were then excluded for they consist of incomplete utterances, or include more than one *Nbp* per sentence. A certain number of particularly ambiguous *Nbp*; *e.g.*, *arcada* 'arcade', *articulação* 'articulation', *etc.* that showed little or no occurrence at all in the *Nbp* sense were discarded from the extracted sentences. Finally, the sentences that lacked a full stop were corrected, in order to prevent errors from STRING's sentence splitting module. In the end, a set of 12,659 sentences with *Nbp* was retained for evaluation.

Based distribution of the remaining 103 *Nbp*, a random stratified sample of 1,000 sentences was selected, keeping the proportion of their total frequency in the corpus. This sample also includes a small number of disease nouns (6 lemmas, 17 sentences).

4.2 Inter-annotator Agreement

The output sentences were then divided into 4 subsets of 225 sentences each. Each subset was then given to a different annotator, and a common set of 100 sentences was added to each subset in order to assess inter-annotator agreement. For each sentence, the annotators were asked to append the whole-part dependency, as it was previously defined in a set of guidelines, using the XIP format.

Table 1. Average Pairwise Percent Agreement

Average pairwise percent agr.	Pairwise pct. agr. cols 1 & 4	Pairwise pct. agr. cols 1 & 3	Pairwise pct. agr. cols 1 & 2	Pairwise pct. agr. cols 2 & 4	Pairwise pct. agr. cols 2 & 3	Pairwise pct. agr. cols 3 & 4
85.031%	86.111%	**90.741%**	82.407%	81.481%	80.556%	88.889%

Table 2. Average Pairwise Cohen's Kappa

Average pairwise CK	Pairwise CK cols 1 & 4	Pairwise CK cols 1 & 3	Pairwise CK cols 1 & 2	Pairwise CK cols 2 & 4	Pairwise CK cols 2 & 3	Pairwise CK cols 3 & 4
0.629	0.65	**0.757**	0.59	0.558	0.518	0.699

For example, for (1) the annotators would produce `WHOLE-PART(João,braço)`. Annotators could also mark a `FIXED` dependency (in the case of idioms) or no dependency at all (if no whole-part relation was present).

From the 100 sentences that were annotated by all the participants in this process, we calculated the Average Pairwise Percent Agreement, the Fleiss' Kappa [10], and the Cohen's Kappa coefficient for inter-annotator agreement [8] using ReCal3: Reliability Calculator [12], for 3 or more annotators [3].

The four annotators achieved the following results. First, the Average Pairwise Percent Agreement, that is, the percentage of cases each pair of annotators agreed with each other, is 85.031%, which is relatively high, as it is shown in Table 1. Next, the Fleiss' Kappa inter-annotator agreement coefficient was calculated, and it equals 0.625; the observed agreement of 0.85 is higher than expected agreement of 0.601, which we deem as a positive result. Finally, the Average Pairwise Cohen's Kappa is 0.629. as shown in Table 2.

According to Landis and Koch [20] this figures correspond to the lower bound of the "substantial" agreement; however, according to Fleiss [11], these results correspond to an inter-annotator agreement halfway between "fair" and "good".

In view of these results, we can assume as a reasonable expectation that the remaining, independent and non-overlapping annotation of the corpus by the four annotators is sufficiently consistent, so it will be used for the evaluation of the system output.

4.3 Evaluation of the System's Overall Performance

The results of the system performance are showed in Table 3. The number of instances (`TP`=*true-positives*; `TN`=*true-negatives*; `FP`=*false-positives*; `FN`=*false-negatives*) is higher than the number of sentences, as one sentence may involve several instances, and we count 5 partial `TP` as 0.5. The relative percentages of the `TP`, `TN`, `FP` and `FN` instances are similar between the 100 and the 900 set of sentences. This explains the similarity of the evaluation results and seems to

[3] `http://dfreelon.org/utils/recalfront/recal3/` [last access: 08.02.2014].

Table 3. System's performance for *Nbp*.

Number of sentences	TP	TN	FP	FN	Precision	Recall	F-measure	Accuracy
100	8	73	7	14	0.53	0.36	0.43	0.79
900	73.5	673	55	118	0.57	0.38	0.46	0.81
Total:	81.5	746	62	132	0.57	0.38	0.46	0.81

confirm our decision to use the remaining 900 sentences' set as a golden standard for the evaluation of the system's output with enough confidence. The recall is relatively small (0.38), which can be explained by the fact that in many sentences, the *whole* and the *part* are not syntactically related. Precision is somewhat better (0.57). The accuracy is relatively high (0.81) since there is a large number of *true-negative* cases.

5 Conclusions

In this paper, we present a rule-based module for whole-part relations extraction involving human entities and body-part nouns (*Nbp*) in Portuguese, which has been implemented and integrated in the STRING NLP system. The most relevant syntactic constructions triggering this meronymic relations were described including the recovering of an implicit *Nbp* associated to several disease nouns. We also prevented idioms composed of *Nbp* to be captured, in view of their meaning non-compositionality, by capturing those idioms before the meronymy module. Around 17 thousand sentences with *Nbp* and disease nouns were extracted from a corpus. 4 Portuguese native speakers annotated a stratified random sample of 1,000 sentences and produced a golden standard, which was confronted against the system's output. The results show 0.57 precision, 0.38 recall, 0.46 F-measure, and 0.81 accuracy. In future work, we intent to improve recall by focusing on the *false-negative* cases already found, which shown that several syntactic patterns have not been paid enough attention, such as coordination.

Acknowledgments. This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013; and Erasmus Mundus Action 2 2011-2574 Triple I - Integration, Interaction and Institutions. We would like to thank the comments of the anonymous reviewers, which helped to improve this paper.

References

1. Abreu, S., Bonamigo, T., Vieira, R.: A review on Relation Extraction with an eye on Portuguese. J. Braz. Comp. Soc. 19(4), 553–571 (2013)
2. Ait-Mokhtar, S., Chanod, J., Roux, C.: Robustness beyond shallowness: Incremental dependency parsing. Natural Language Engineering 8(2/3), 121–144 (2002)

3. Baptista, J., Correia, A., Fernandes, G.: Frozen Sentences of Portuguese: Formal Descriptions for NLP. In: Workshop on Multiword Expressions: Integrating Processing. Intl. Conf. of the European Chapter of the ACL, Barcelona, Spain, pp. 72–79 (2004)
4. Baptista, J., Mamede, N., Markov, I.: Integrating Verbal Idioms into an NLP System. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T., das Graça Nunes, M. (eds.) PROPOR 2014. LNCS(LNAI), vol. 8775, pp. 250–255. Springer, Heidelberg (2014)
5. Berland, M., Charniak, E.: Finding parts in very large corpora. In: Proceedings of the 37th Annual Meeting of the ACL on Computational Linguistics, pp. 57–64. ACL, Morristown (1999)
6. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Ph.D. thesis, Aarhus Univ. Aarhus, Denmark: Aarhus Univ. Press (2000)
7. Costa, F., Branco, A.: LXGram: A deep linguistic processing grammar for portuguese. In: Pardo, T.A.S., Branco, A., Klautau, A., Vieira, R., de Lima, V.L.S. (eds.) PROPOR 2010. LNCS (LNAI), vol. 6001, pp. 86–89. Springer, Heidelberg (2010)
8. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960)
9. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT, Cambridge (1998)
10. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psych. Bull. 76(5), 378–382 (1971)
11. Fleiss, J.: Statistical methods for rates and proportions, 2nd edn. John Wiley, New York (1981)
12. Freelon, D.: ReCal: Intercoder Reliability Calculation as a Web Service. Intl. J. of Internet Science 5(1), 20–33 (2010)
13. Girju, R., Badulescu, A., Moldovan, D.: Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In: Proceedings of HLT-NAACL, vol. 3, pp. 80–87 (2003)
14. Girju, R., Badulescu, A., Moldovan, D.: Automatic discovery of part-whole relations. Computational Linguistics 21(1), 83–135 (2006)
15. van Hage, W.R., Kolb, H., Schreiber, G.: A method for learning part-whole relations. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS (LNAI), vol. 4273, pp. 723–735. Springer, Heidelberg (2006)
16. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conf. on Computational Linguistics, COLING 1992, vol. 2, pp. 539–545. ACL, Morristown (1992)
17. Hirst, G.: Ontology and the lexicon. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 209–230. Springer (2004)
18. Iris, M., Litowitz, B., Evens, M.: Problems of the Part-Whole Relation. In: Evens, M. (ed.) Relational Models of the Lexicon: Representing Knowledge in Semantic Networks, pp. 261–288. Cambridge Univ. Press (1988)
19. Khoo, C., Na, J.C.: Semantic Relations in Information Science. Annual Review of Information Science and Technology 40, 157–229 (2006)
20. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (1977)

21. Mamede, N., Baptista, J., Diniz, C., Cabarrão, V.: STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In: PROPOR 2012, vol. Demo Session. Paper available at http://www.propor2012.org/demos/DemoSTRING.pdf
22. Marrafa, P.: WordNet do Português: Uma base de dados de conhecimento linguístico. Instituto Camões (2001)
23. Marrafa, P.: Portuguese WordNet: general architecture and internal semantic relations. DELTA 18, 131–146 (2002)
24. Marrafa, P., Amaro, R., Mendes, S.: WordNet.PT Global – extending WordNet.PT to Portuguese varieties. In: Proceedings of the 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, pp. 70–74. ACL Press, Edinburgh (2011)
25. Oliveira, H.: Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese. Ph.D. thesis, Univ. of Coimbra/FST (2012)
26. Oliveira, H.G., Santos, D., Gomes, P., Seco, N.: PAPEL: A Dictionary-based Lexical Ontology for Portuguese. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 31–40. Springer, Heidelberg (2008)
27. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: Proceedings of Conf. on Computational Linguistics/ACL (COLING/ACL-06), Sydney, Australia, pp. 113–120 (2006)
28. Paumier, S.: De la reconnaissance de formes linguistiques à l'analyse syntaxique. Ph.D. thesis, Université de Marne-la-Vallée (2000)
29. Paumier, S.: Unitex 3.1beta, User Manual. Univ. Paris-Est Marne-la-Vallée (2014)
30. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: developing an aligned multilingual database. In: 1st Intl. Conf. on Global WordNet (2002)
31. Prévot, L., Huang, C., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A.: Ontology and the lexicon: A multi-disciplinary perspective (introduction). In: Huang, C., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prévot, L. (eds.) Ontology and the Lexicon: A Natural Language Processing Perspective. Studies in Natural Language Processing, ch. 1, pp. 3–24. Cambridge Univ. Press (2010)
32. Rocha, P., Santos, D.: CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Nunes, M. (ed.) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000), pp. 131–140. ICMC/USP, São Paulo (2000)