Automatic Identification of 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,a48654}@ualg.pt

2 **Instituto Superior Técnico, Universidade de Lisboa**
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
Nuno.Mamede@ist.utl.pt

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

Abstract

In this paper, we improve the extraction of semantic relations between textual elements as it is currently performed by STRING, a hybrid statistical and rule-based Natural Language Processing chain for Portuguese, by targeting *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 meronymy involving human entities and *body-part nouns*.

1998 ACM Subject Classification I.2.7 Natural Language Processing

Keywords and phrases whole-part relation, meronymy, body-part noun, disease noun, Portuguese

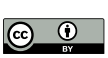
Digital Object Identifier 10.4230/OASIcs.SLATE.2014.225

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 (QA), Text Summarization (TS), Machine Translation (IR), Information Extraction (IE), Information Retrieval (IR) and others [9, 10, 15].

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] [17]. At this time, only the first steps have been taken in the direction of semantic parsing. 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 meronymy involving human entities and *Nbp*. 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 from this work.

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



3rd Symposium on Languages, Applications and Technologies (SLATE'14).
Editors: Maria João Varanda Pereira, José Paulo Leal, and Alberto Simões; pp. 225–232

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Related Work

Meronymy is a complex relation that "should be treated as a collection of relations, not as a single relation" [14]. In NLP, various information extraction techniques have been developed in order to capture whole-part relations from texts.

Hearst [12] 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 [5]. 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." [12, p. 544]. The author claims that he tried applying the same technique to meronymy, but without great success.

Girju *et al.* [9, 10] 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.* [11] developed a method for learning whole-part relations from vocabularies and text sources. 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 [23] was developed in order to harvest semantic relations in a text. 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.

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 [13, 26]. Well-known, existing lexical ontologies for Portuguese are Portuguese WordNet.PT [18, 19], later extended to WordNet.PT Global (Rede Léxico-Conceptual das Variedades do Português) [20]; MWN.PT-MultiWordNet of Portuguese[3] [25]; PAPEL (Palavras Associadas Porto Editora Linguateca)[4] [22]; and Onto.PT[5] [21]. 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)[6].

[2] http://www.fao.org/agrovoc [last access: 04.05.2014].
[3] http://mwnpt.di.fc.ul.pt/ [last access: 04.05.2014].
[4] http://www.linguateca.pt/PAPEL/ [last access: 04.05.2014].
[5] http://ontopt.dei.uc.pt/ [last access: 04.05.2014].
[6] Only after submission of this paper, we were alerted for the work of Cláudia de Freitas for annotating

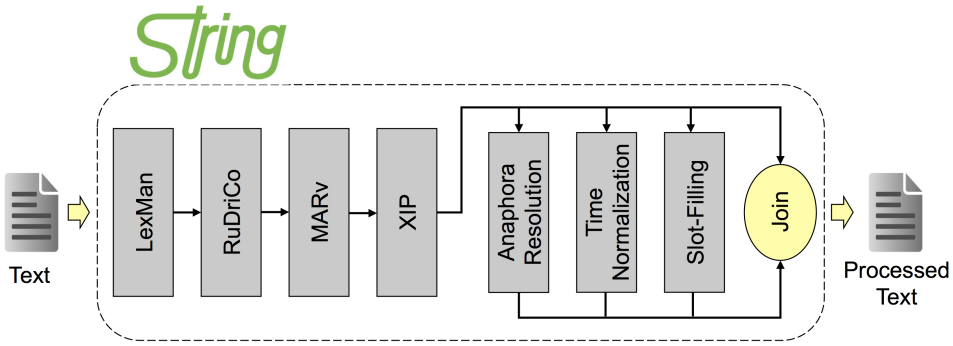


Figure 1 STRING Architecture.

Attention was also paid to two well-known parsers of Portuguee, in order to discern how did they handle the whole-part relations extraction: the PALAVRAS parser [2], consulted using the Visual Interactive Syntax Learning (*VISL*) environment[7], and LX Semantic Role Labeller[8] [3]. Judging from the available on-line versions/demos of these systems, apparently, none of these parsers extracts whole-part relations, at least explicitly.

3 Whole-Part Dependency Extraction Module in STRING

3.1 STRING Overview

STRING [17] is a fully-fledged NLP chain that 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 MARv module then performs POS-disambiguation, using HMM and the Viterbi algorithm; and, finally, (iv) the XIP parser (Xerox Incremental Parser) [1] segments sentences into chunks (or elementary sentence constituents: `NP`, `PP`, etc.) and extracts dependency relations among chunks' heads (`SUBJect`, `MODifier`, etc.). 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 executes *dependency rules*. Dependency rules extract different types of dependencies between 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

the human body semantic features in the AC/DC corpora, so we did not considered it here; please refer to: http://www.linguateca.pt/acesso/Esqueleto.pdf [last access: 04.05.2014].

[7] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/dependency.php [last access: 04.05.2014].

[8] http://lxcenter.di.fc.ul.pt/services/en/LXSemanticRoleLabeller.html [last access: 04.05.2014].

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 the XIP parser. 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. Different typical, syntactic-semantic situations targeted by the meronymy extraction module could be sketched out, but for space limitations only the most simple will be presented here. Example (1) is a simple case where there is a determinative `PP` complement *de N* (of N), 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)
```

The rule itself 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`); 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`); if all these conditions are met, then, the parser builds the `WHOLE-PART` relation between the human determinative complement and the *Nbp*. A similar rule is required for a dative complement, as in sentence *O Pedro partiu um braço ao João/O pedro partiu-lhe um braço* (Pedro broke him an arm).

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.

Another interesting case is 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:

(3) *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 (3) to 'get mad'.

The STRING 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 (ambiguous idioms, e.g. *à cabeça* (on/at the head), must be addressed in another way). 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 (3) 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, from 10 formal classes of idioms.

4 Evaluation

The first fragment of the CETEMPúblico corpus [27] 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* was adapted to be used within the UNITEX corpus processor [24] along with the remaining available resources for European Portuguese, distributed with the system.

Using the *Nbp* (151 lemmas) dictionary 16,746 *Nbp* 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 consisted of incomplete utterances, or included more than one *Nbp* per sentence. A certain number of particularly ambiguous *Nbp*; *e.g.*, *arcada* (arcade), *articulação* (articulation), etc., which 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. The output sentences were 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. For example, for (1) the annotators would produce `WHOLE-PART(João,braço)`.

From the 100 sentences that were annotated by all the participants in this process, we calculated the Average Pairwise Percent Agreement, the Fleiss' Kappa [6], and the Cohen's Kappa coefficient of inter-annotator agreement [4] using ReCal3: Reliability Calculator [8], for 3 or more annotators.[9] The four annotators achieved the following results. First, the

[9] http://dfreelon.org/utils/recalfront/recal3/ [last access: 04.05.2014].

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

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

Average Pairwise Percent Agreement, that is, the percentage of cases each pair of annotators agreed with each other is shown in Table 1. The Average Pairwise Percent Agreement is 85.031%, which is relatively high. 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 shown in Table 2. The Average Pairwise Cohen's Kappa is 0.629. According to Landis and Koch [16] this figures correspond to the lower bound of the "substantial" agreement; however, according to Fleiss [7], 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, and will use it for the evaluation of the system output.

The system performance was evaluated using the usual evaluation metrics of Precision, Recall, F-measure, and Accuracy. The results are shown in Table 3, where `TP`=*true-positives*; `TN`=*true-negatives*; `FP`=*false-positives*; `FN`=*false-negatives*. The number of instances (`TP`, `TN`, `FP` and `FN`) is higher than the number of sentences, as one sentence may involve several instances. 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 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 and are quite far away from each other. 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

This paper addressed the problem of whole-part relations extraction involving human entities and body-part nouns (*Nbp*) in Portuguese. A rule-based meronymy extraction module has been built and integrated in the grammar of the STRING system. It contains 27 general rules addressing the most relevant syntactic constructions triggering this type of meronymic relations. A set of 400 rules had also been devised to prevent the whole-part relations being extracted in the case the *Nbp* are elements of idiomatic expressions. From a relatively large corpus, about 17 thousand sentences with *Nbp* were extracted. A stratified random sample of 1,000 sentences was independently annotated by 4 Portuguese native speakers in order to produce a golden standard and confront it 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.

Acknowledgements. 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 S. Ait-Mokhtar, J. Chanod, and C. Roux. Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(2/3):121–144, 2002.

2 E. Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus Univ. Aarhus, Denmark: Aarhus Univ. Press, 2000.

3 A. Branco and F. Costa. A Deep Linguistic Processing Grammar for Portuguese. In Pardo et al., editor, *Computational Processing of Portuguese*, LNAI 6001, pages 86–89. Springer, 2010.

4 J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

5 C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT, Cambridge, 1998.

6 J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psych. Bull.*, 76(5):378–382, 1971.

7 J. L. Fleiss. *Statistical methods for rates and proportions (2nd ed.)*. New York: John Wiley, 1981.

8 D. Freelon. ReCal: Intercoder Reliability Calculation as a Web Service. *Intl. J. of Internet Science*, 5(1):20–33, 2010.

9 R. Girju, A. Badulescu, and D. Moldovan. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In *Proceedings of HLT-NAACL*, volume 3, pages 80–87, 2003.

10 R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, 21(1):83–135, 2006.

11 W. Van Hage, H. Kolb, and G. Schreiber. A method for learning part-whole relations. *The Semantic Web – ISWC 2006, LNAI/LNCS*, 4273:723–725, 2006.

12 M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conf. on Computational linguistics*, volume 2 of *COLING 92*, pages 539–545. ACL Morristown, NJ, USA, 1992.

13 G. Hirst. Ontology and the lexicon. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, pages 209–230. Springer, 2004.

14 M. Iris, B. Litowitz, and M. Evens. Problems of the Part-Whole Relation. In M. Evens, editor, *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, pages 261–288. Cambridge Univ. Press, 1988.

15 C. Khoo and J.-C. Na. Semantic Relations in Information Science. *Annual Review of Information Science and Technology*, 40:157–229, 2006.

16 J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

17 N. Mamede, J. Baptista, C. Diniz, and V. Cabarrão. STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In *Intl. Conf. on Computational Processing of Portuguese (PROPOR 2012)*, volume Demo Session, Paper available at `http://www.propor2012.org/demos/DemoSTRING.pdf`, 2012.

18 P. Marrafa. *WordNet do Português: uma base de dados de conhecimento linguístico*. Instituto Camões, 2001.

19 P. Marrafa. Portuguese WordNet: general architecture and internal semantic relations. *DELTA*, 18:131–146, 2002.

20 P. Marrafa, R. Amaro, and S. Mendes. 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*, pages 70–74, Edinburgh, Scotland. ACL Press, 2011.

21 H. Gonçalo Oliveira. *Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*. PhD thesis, Univ. of Coimbra/FST, 2012.

22 H. Gonçalo Oliveira, P. Gomes, D. Santos, and N. Seco. PAPEL: A Dictionary-based Lexical Ontology for Portuguese. In *Computational Processing of the Portuguese Language, 8th Intl. Conf., Proceedings (PROPOR 2008)*, volume 5190, pages 31–40, Aveiro, Portugal. Springer, 2008.

23 P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of Conf. on Computational Linguistics/ACL (COLING/ACL-06)*, pages 113–120. Sydney, Australia, 2006.

24 S. Paumier. *Unitex 3.1.beta, User Manual*. Univ. Paris-Est Marne-la-Vallée, 2014.

25 E. Pianta, L. Bentivogli, and C. Girardi. MultiWordNet: developing an aligned multilingual database. In *1st Intl. Conf. on Global WordNet*, 2002.

26 L. Prévot, C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, and A. Oltramari. Ontology and the lexicon: a multi-disciplinary perspective (introduction). In C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, and L. Prévot, editors, *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 1, pages 3–24. Cambridge Univ. Press, 2010.

27 P. Rocha and D. Santos. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In M. G. Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, pages 131–140. São Paulo: ICMC/USP, 2000.