A PropBank for Portuguese: the CINTIL-PropBank

António Branco, Catarina Carvalheiro, Sílvia Pereira, Mariana Avelãs, Clara Pinto, Sara Silveira, Francisco Costa, João Silva, Sérgio Castro, João Graça

University of Lisbon
Edifício C6, Departamento de Informática
Faculdade de Ciências, Universidade de Lisboa
Campo Grande, 1749-016, Portugal
{antonio.branco, catarina.carvalheiro, silvia.pereira, mariana.avelas,
clara.pinto, sara.silveira, fcosta, jsilva, sergio.castro}@di.fc.ul.pt

Abstract

With the CINTIL-International Corpus of Portuguese, an ongoing corpus annotated with fully flegded grammatical representation, sentences get not only a high level of lexical, morphological and syntactic annotation but also a semantic analysis that prepares the data to a manual specification step and thus opens the way for a number of tools and resources for which there is a great research focus at the present. This paper reports on the construction of a propbank that builds on CINTIL-DeepGramBank, with nearly 10 thousand sentences, on the basis of a deep linguistic grammar and on the process and the linguistic criteria guiding that construction, which makes possible to obtain a complete PropBank with both syntactic and semantic levels of linguistic annotation. Taking into account this and the promising scores presented in this study for inter-annotator agreement, CINTIL-PropBank presents itself as a great resource to train a semantic role labeller, one of our goals with this project.

Keywords: propbank, portuguese, annotated corpus

1. Introduction

Following the important methodological breakthrough that took place in Language Technology with the advent of statistical approaches, the development of annotated corpora has been deployed around adding increasingly more complex linguistic information, e.g. concerning phrase constituency (aka TreeBanks (Marcus et al., 1993)), syntactic functions (aka DependencyBanks (Böhmová et al., 2001)), and phrase-level semantic roles (aka PropBanks (Palmer et al., 2005)), just to mention a few salient examples.

To keep advancing along this trend and to develop corpora that are annotated with deep linguistic representations, the construction of annotated corpora faces a challenge that demands a new qualitative step: the fully fledged grammatical representation to be assigned to each sentence is so complex and so specific to that sentence that it cannot be reliably crafted manually piece by piece and the annotation cannot be performed without some supporting application, viz. a computational grammar.

This paper discusses the solutions we developed to construct a propbank on the basis of a deep linguistic grammar and its companion deep linguistic treebank (Branco et al., 2010), with a central goal: the construction of a high quality data set with semantic information that could support the development of automatic semantic role labellers (Baker et al., 2007; Carreras and Màrquez, 2005) for Portuguese.

Section 2 reports on the construction of a propbank on the basis of a corpora annotated with a deep linguistic grammar. In Section 3, we describe the extraction of semi-annotated constituency trees with automatic semantic roles that assist the manual completion step of our dynamic propbank, presented in Section 4. In Section 5, we enumerate some aplications of the PropBank, and Section 6 presents the concluding remarks.

2. A PropBank supported by a deep linguistic grammar

The deep linguistic grammar used for the initial semi-automatic propbanking was LXGram, a grammar for the computational processing of Portuguese (Branco and Costa, 2010; Branco and Costa, 2008a; Branco and Costa, 2008b), developed under the grammatical framework of HPSG (Pollard and Sag, 1994) which uses MRS (Copestake et al., 2005) for the representation of meaning and the Grammar Matrix (Bender et al., 2002) for the initial type system. In a first phase, the parses obtained with LXGram and manually selected by human annotators were gathered in the CINTIL-DeepGramBank, a corpus of deep grammatical representations, composed by sentences taken from the CINTIL-International Corpus of Portuguese with 1 million tokens of written and spoken linguistic materials (Branco et al., 2010).

The construction of the CINTIL-DeepGramBank was performed adopting the annotation procedure where independent annotators produce primary data and their decisions are validated in a subsequent adjudication phase by a third independent annotator. More specifically, each sentence was automatically processed by LX-Suite (Silva, 2007) and analysed by LXGram (Branco and Costa, 2010): once a set of grammatical analysis is obtained (parse forest), two independent annotators choose the analysis each one of them considers to be correct. In case of divergence between their decisions, a third independent adjudicator reviews their options and makes the final choice. The annotators and adjudicators are language experts with post-graduate degrees in Linguistics.

The workbench used to support this process of annotation was [incr tsdb()] (Oepen, 2001), which permits to parse, select and collect fully fledged deep grammatical represen-

tations for the respective sentences. Annotation speed is roughly 80 to 100 sentences per day. At the moment, last stable version 3 of the CINTIL-DeepGramBank is composed of 5422 sentences. For this version, the level of inter-annotator agreement (ITA) scores 0.86 in terms of the specific inter-annotator metric we developed for this kind of corpora and annotation (Castro, 2011). Since the CINTIL-DeepGramBank keeps being developed, we have an additional 4047 sentences in the ongoing version 4, with 0.80 of inter-annotator agreement.

3. Extracting semi-annotated constituency trees

Propbanks are syntactic constituency treebanks whose trees have their constituents labeled with semantic role tags (Palmer et al., 2005). Propbanks are thus annotated corpora that result from the extension of the annotation associated to the sentences in treebanks by means of an extra layer of linguistic information for semantic roles.

After the manual selection of the correct analyses (described in the previous section), the CINTIL-DeepGramBank was processed in order to obtain only the syntactic constituency trees.[1] To achieve this, the tool lkb2standard (Silva et al., 2010) was developed to extract these trees from the files exported by [incr tsdb()]. These are trees that are then ready to be extended to form the CINTIL-PropBank, by means of their enrichment with appropriate semantic role tags.

Some of the semantic role labels in the tag set used in this PropBank can be obtained directly from the deep grammatical representations and through this extraction tool. This is done by resorting to the feature structures that describe the semantics of the sentence in the CINTIL-DeepGramBank, namely those used to represent the arguments of predicators, ARG1 to ARG*n*. Furthermore, the extraction tool lkb2standard was designed to play a role that goes beyond the mere extraction of the constituency tree annotated with these ARG1 to ARG*n* labels. By resorting to the details of the deep grammatical representation, it permits to label phrases with a number of further labels that account for phrases that, on the surface level, are associated with more than one argument (see Figure 1, for example):

- **ARG1** – Argument 1
  e.g. *O João deu uma flor à Maria.* ("João gave Maria a flower.")
- **ARG2** – Argument 2
  e.g. *O João deu uma flor à Maria.* (*idem*)
- **ARG3** – Argument 3
  e.g. *O João deu uma flor à Maria.* (*idem*)
- **ARG11** – Argument 1 of subordinating predicator and Argument 1 in the subordinate clause (semantic function of Subjects of so called Subject Control predicators)
  e.g. *As crianças não querem dormir.* ("The children don't want to go to sleep.")
- **ARG21** – Argument 2 of subordinating predicator and Argument 1 in the subordinate clause (semantic function of Subjects of so called Direct Object Control predicators)
  e.g. *Uma oferta obrigou o João a tomar medidas.* ("An offer made João take action.")
- **ARG*n*cp** – Argument *n* in complex predicate constructions
  e.g. *O cliente podia estar mais confiante.* ("The client could have been more confident.")
- **ARG*n*ac** – Argument *n* of anticausative readings
  e.g. *O doente acordou.* ("The patient woke up.")

4. Manual PropBanking: completing the annotation

Building on the information made explicit by the deep linguistic grammar, the remaining phrases that are modifiers, associated with non argumental positions, are left with the semantic role tag M, as we can see in Figure 1.

There are two further tools supporting this manual phase of annotation described below aimed at specifying the semantic role of modifiers: one converts trees into an annotation format compatible with the annotation interface (see Figure 2); and a reverser tool for the inverse operation (transformed trees, such as the one shown in Figure 3).[2]

As the outcome of the operation of the first of them, the set of sentences to be annotated can be presented in a spreadsheet file, with each sentence in a different sheet. For each suite of treebanked sentences, a spreadsheet is created with as many sheets as there are sentences in that suite. If a given sentence happens not to have received a parse, its sheet only contains its identification number and that sentence.

As we can see in Figure 2, each line has cells automatically filled in, and others to be manually filled in by the annotator. Each line includes: in column (A), the syntactic category and grammatical function; in column (B), the semantic role assigned by the grammar; in (C), the cell to be filled in by the human annotator; in (D), the constituent being tagged, and in (E) the possible observations from the annotator.

A completion step followed that consists in the manual specification of the occurrences of this portmanteau tag M in terms of one of the semantic roles available for modifiers in our tag set:

- **LOC** – Location: to locate an action in place, whether physical or abstract (see Figure 4)
- **EXT** – Extension: to use with strings with an extension notion, mainly numerical. Includes measures, percentages, quantifiers, and comparative expressions (see Figure 4)
- **CAU** – Cause: to determine a cause, a reason of an action (see Figure 5)
- **TMP** – Temporal: to locate an action in time, including the frequency, duration, and repetition (see Figure 4)

[1] For a detailed account of the linguistic options that are behind the syntactic constituency, see (Branco et al., 2011).

[2] For a more detailed account of this annotation environment and process, see (Branco et al., 2009).

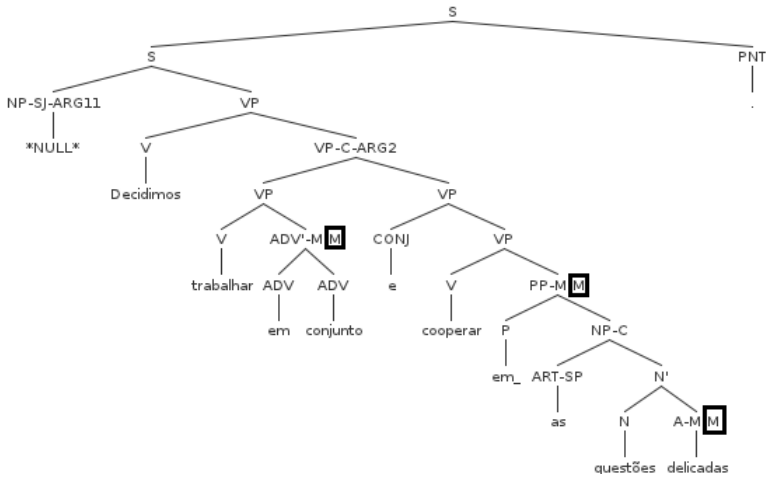

Figure 1: CINTIL-DeepGramBank constituency tree with semantic M tags highlighted for: *Decidimos trabalhar em conjunto e cooperar nas questões delicadas* ("We decided to work together and cooperate on the delicate issues").

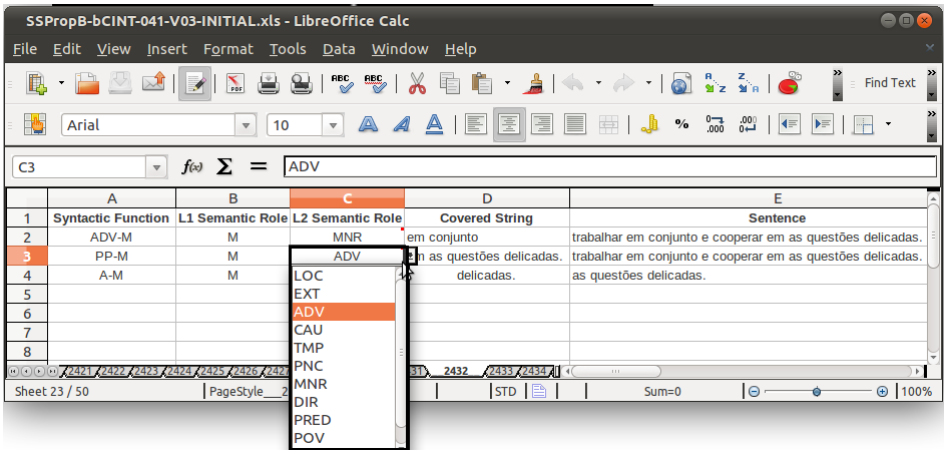

Figure 2: Spreadsheet annotation interface for specifying semantic roles of the M tags for: *Decidimos trabalhar em conjunto e cooperar nas questões delicadas* ("We decided to work together and cooperate on the delicate issues").

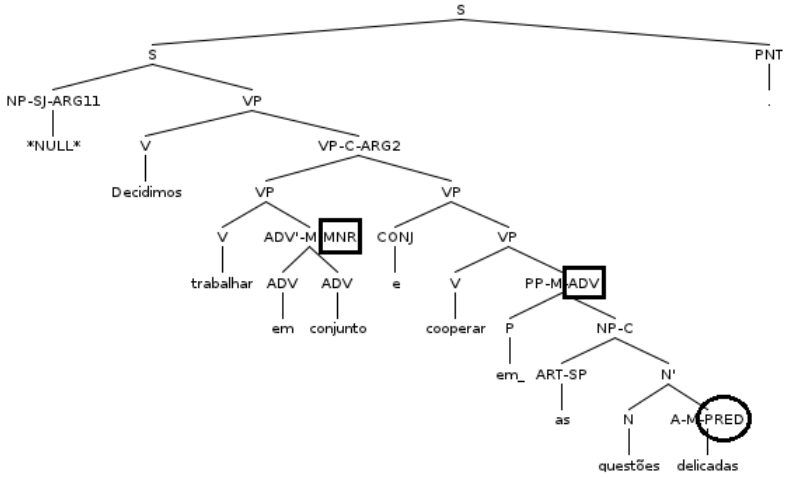

Figure 3: CINTIL-PropBank tree with manual MNR and ADV tags and automatic PRED tag highlighted for: *Decidimos trabalhar em conjunto e cooperar nas questões delicadas* ("We decided to work together and cooperate on the delicate issues").

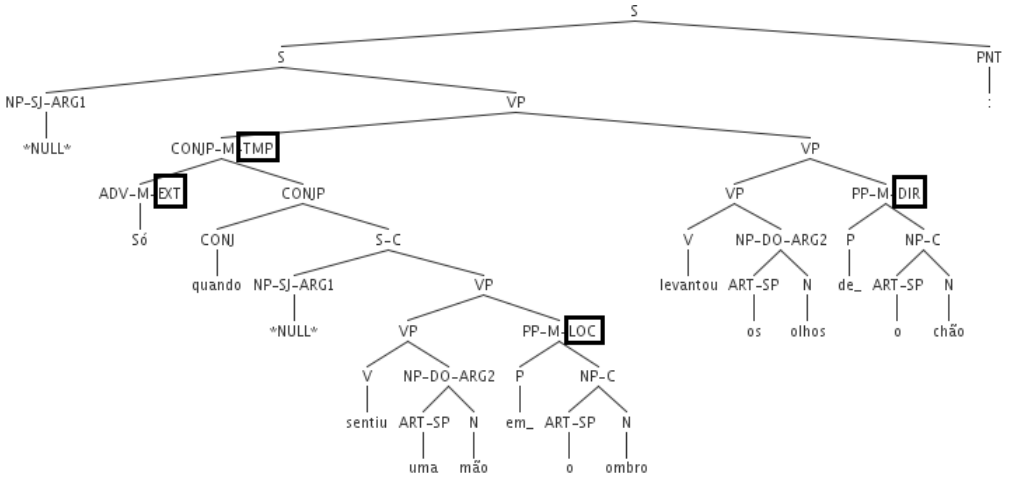

Figure 4: CINTIL-PropBank tree with the semantic roles EXT, TMP, LOC, and DIR highlighted for: *Só quando sentiu uma mão no ombro levantou os olhos do chão* ("Only when he felt a hand on his shoulder did he raise his eyes from the floor").

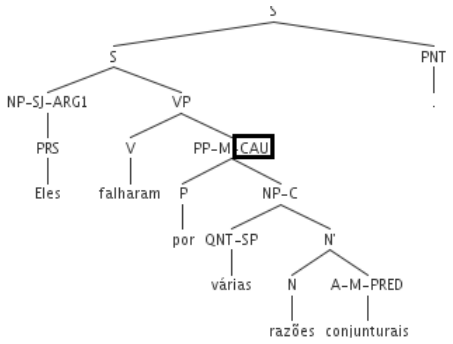

Figure 5: CINTIL-PropBank tree with the semantic role CAU highlighted for: *Eles falharam por várias razões conjunturais* ("They failed for many conjunctural reasons").

- **PNC** – Purpose, goal: to all strings that describe a goal or a proposal of a given action (see Figure 6)
- **MNR** – Manner: to all strings that specifies the way, manner how an action is realized or due (see Figure 3)
- **DIR** – Direction: to reference directions, covering both the source/origin and destination (see Figure 4)
- **POV** – Point of View: to strings that expresses an author position about a given event (see Figure 7)
- **PRED** – Secondary predication: to all cases of predicative structures, mainly past participles and resultative constructions
- **ADV** – Adverbial: to strings that do not fall into any of the other categories (see Figure 3)

At this point, it is important to note that, in the case of attributes and relative clauses with A-M, AP-M and CP-M tags at the constituency level, the tag M (at the third level, the semantic role) is automatically replaced by PRED at this step of conversion (see and compare Figures 1 and 3 for the phrase "nas questões delicadas").

This manual phase of the construction of the PropBank is always done by two independent annotators, who choose the tags each one of them consider to be correct. In case of divergence between annotators, a third independent adjudicator reviews their decisions and makes the final choice. The annotators are experts with post-graduations in Linguistics. The annotation speed is around 200 sentences per day. According to our latest data, from stable version 3 (5422 sentences), the level of inter-annotator agreement is over 0.75 in terms of the k-coefficient. For the ongoing version 4 (with an extra 4047 sentences), the level of inter-annotator agreement is 0.76.

When this manual propbanking is finalized, the sentences — now extended with the newly assigned tags for the semantic roles of modifiers — are reverted back into the original tree representation. This operation is ensured by a reverting tool that takes the data in the sheets of the spreadsheet and recombines the new information added by the human annotator with the original information (grammatical category and syntactic functions) about the parse tree of the sentence. We have now a complete PropBank with the two information levels: phrase constituency and phrase-level semantic roles. As can be seen in Figure 3, we have now all the M tags replaced by fully specified semantic values:

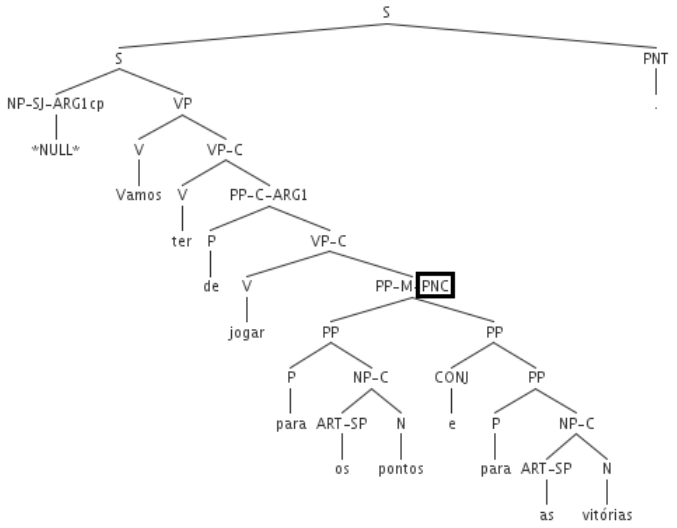

Figure 6: CINTIL-PropBank tree with the semantic role PNC highlighted for: *Vamos ter de jogar para os pontos e para as vitórias* ("We're going to have to play for points and victories").

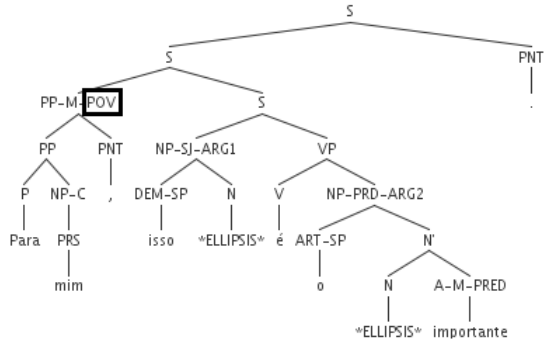

Figure 7: CINTIL-PropBank tree with the semantic role POV highlighted for: *Para mim, isso é importante* ("To me, that's important").

ADV and MNR tags. Recall that, in this case, the PRED tag was automatically assigned at the conversion step that generated the spreadsheet.

At this point, with all propbanking guidelines, criteria, and process succinctly described, we are able to attest how do the labels enumerated in previous section work through examples illustrating their assignment (see Figures 4 to 7).

5. Some applications of the PropBank

It is important to note that with the automatic PropBanking phase it was already possible to extract treebanks and dependencybanks since CINTIL-DeepGramBank already contains the constituency structure with syntactic information, which is enough to extract a treebank, and syntactic functions tags, which can be used to build a dependencybank. With the second PropBanking step — manual specification of semantic role tags — we now have an opportunity to get an added value, a resource to train a semantic role labeller.[3] A semantic role labeller allows to correctly identify the various semantic roles in a sentence enabling the recognition of relations between their elements, such as who did what, what happened to whom, etc. With these semantic values, we have a world of new possibilities to improve or create tools and resources for areas such as question answering, information extraction, summarization, machine learning, and information retrieval on the web which opens the possibility for semantic web searching.

[3]This application is currently under development and testing with the current version of the CINTIL-PropBank.

6. Concluding remarks

In this paper, we reported on the solutions we followed to develop a propbank with almost 10 thousand sentences. This propbank was built with the help of a deep linguistic grammar which permitted to construct a high quality and reliable data set with semantic information that will support the training of semantic role labellers for Portuguese. This resource has also the potential to benefit many other natural language processing applications, such as information extraction, question-answering, summarization, machine translation, information retrieval, among others.

7. References

Colin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval'07 task 19: frame semantic structure extraction. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07, ACL)*, pages 9–104, Stroudsburg, PA, USA.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The Grammar Matrix: An open-source starter-kit for the development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Procedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2001. The Prague Dependency Treebank: A three-level annotation scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 103–127. Kluwer Academic Publishers.

António Branco and Francisco Costa. 2008a. A computational grammar for deep linguistic processing of portuguese: LXGram, version A.4.1. Technical Report TR-2008-17, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática.

António Branco and Francisco Costa. 2008b. LXGram in the shared task "comparing semantic representations" of STEP 2008. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 299–314. College Publications.

António Branco and Francisco Costa. 2010. A deep linguistic processing grammar for Portuguese. In *Lecture Notes in Artificial Intelligence*, volume 6001, pages 86–89. Springer, Berlin.

António Branco, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto, and Francisco Costa. 2009. Dynamic propbanking with deep linguistic grammars. In *Proceedings, TLT009 — The 8th International Workshop on Treebanks and Linguistic Theories*, pages 39–50, Milan.

António Branco, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto, and João Graça. 2010. Developing a deep linguistic databank supporting a collection of treebanks: the CINTIL DeepGramBank. In *Proceedings of LREC2010 - The 7th international conference on Language Resources and Evaluation*, La Valleta, Malta.

António Branco, João Silva, Francisco Costa, and Sérgio Castro. 2011. CINTIL TreeBank handbook: Design options for the representation of syntactic constituency. Technical Report TR-2011-02. Available at: http://docs.di.fc.ul.pt/.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: semantic role labeling. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, CONLL'05, pages 152–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sérgio Castro. 2011. Developing reliability metrics and validation tools for datasets with deep linguistic information. Master's thesis, Universidade de Lisboa, Faculdade de Ciências, Departamento de Informática, Portugal.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language & Computation*, 3(4):281–332, December.

Mitchell Marcus, Mary Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June.

Stephan Oepen. 2001. [incr tsdb()] — competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany. in preparation.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Stanford: Chicago University Press and CSLI Publications.

João Silva, António Branco, Sérgio Castro, and Ruben Reis. 2010. Out-of-the-box robust parsing of Portuguese. In *Proceedings of the 9th Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR)*, pages 75–85.

João Silva. 2007. Shallow processing of Portuguese: From sentence chunking to nominal lemmatization. Master's thesis, MSc thesis, University of Lisbon. Published as Technical Report DI-FCUL-TR-07-16, Portugal.