Floresta Sintá(c)tica: Bigger, Thicker and Easier

Cláudia Freitas[1], Paulo Rocha[2], and Eckhard Bick[3]

[1,2]Linguateca, DEI, Universidade de Coimbra, Portugal
{freitas,parocha}@dei.uc.pt
[3]Syddansk Universitet, Odense, Denmark
eckhard.bick@mail.dk

Abstract. In this paper, we describe the resumption of activities of Floresta Sintá(c)tica, a treebank for Portuguese. We present some underlying guidelines around the project and how they influence our linguistic choices. We then describe the new texts added to the treebank, proceed to mention the new syntactic information added to the old texts, and finally describe the new user-friendly search system and the plans for its expansion.

Keywords: Treebank, corpus, syntax, Portuguese language.

1 Introduction

The Floresta Sintá(c)tica[1] is a publicly available treebank for Portuguese. It was created in 2000 as a collaboration between Linguateca[2] and VISL Project[3], and consists of European and Brazilian Portuguese-language texts automatically annotated by the parser PALAVRAS (Bick, 2000). As the project resumed in 2007, the goal of this paper is to present Floresta's new features, namely, (i) additional texts; (ii) linguistic information; and (iii) search interface. A detailed description of the project, as well as its main motivations, objects, building process and usefulness were described elsewhere (see Afonso et al, 2001 and the Floresta documentation page, at the website).

Floresta has a subset corpus, Bosque, manually revised. Since 2007, Bosque has been undergoing a re-revising process, which guarantees more consistent material, regarding not only annotation aspects, but also the documentation of the underlying linguistic choices. In addition, in this new phase we created Selva, an intermediate corpus between Floresta and Bosque, in both size and degree of revision. Finally, we're developing a new search interface, Milhafre.

Although the usefulness of a treebank like Floresta has already been documented (Afonso et al. 2001), we would like to reinforce here the underlying ideas that guide Floresta's choices: to reflect a consensus among the possible syntactic analysis of a given phenomenon, or, at least, to offer an informed choice. As a result, we expect to be able to (i) offer material to the widest possible range of users; (ii) serve as a research space, and not as a one-theory demonstration space (though of course we are

[1] http://www.linguateca.pt/Floresta/
[2] http://www.linguateca.pt/
[3] http://visl.sdu.dk

A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, pp. 216–219, 2008.

aware that we can not escape from an underlying theory to the syntactic annotation). So, we have to balance (a) the need for a grammar that is rich and complex enough in order to process real language (our corpora); (b) the absence of a consensual syntactic model; and (c) the linguistic background of the users. In other words, one of our challenges is to make the material useful, regardless of the "quantity and quality" of the users' linguistic background.

The remainder of the paper is organized as follows: in section 2 we describe Selva; in section 3, we describe some of the new linguistic information that is available; section 4 presents Milhafre, a new search system and its interface for queries; finally, section 5 shows our conclusions and directions for future work.

2 Bigger: The "Selva"

We are aware that Bosque is limited, from both the linguistic and the computational-statistical point of view, by its small size. Additionally, both Bosque and Floresta are composed only of newspaper texts from two single sources. Therefore, we decided to build Selva, a corpus that contains around 300.000 words and 30.000 sentences, divided into three roughly equal shares of scientific, literary and transcribed spoken texts, further subdivided in approximately equal shares of Portuguese and Brazilian texts. These texts were mainly selected for their free availability, which means that the literary texts are mainly late 19th century and early 20th century works (around 10.000 words by each of five Portuguese and five Brazilian authors), while the spoken texts are composed of interviews previously included in the AC/DC project (Santos & Bick 2000) and parliamentary transcripts. Scientific texts were mainly taken from Wikipedia articles on scientific subjects and a small set of academic theses. Selva is intended to be a partially reviewed corpus, where some characteristics of the corpus are reviewed one by one, instead of the complete annotation being revised tree by tree as in Bosque.

3 Thicker

One of our tasks was to map the new tags from the parser into the previously reviewed files of Bosque, and then review them manually; Selva had those tags from the start.

First, we reviewed some new function tags. The tags N<ARGS and N<ARGO were introduced to mark arguments of the head noun related to subjects and objects, respectively, when the head noun is a deverbal noun. We used N<ARG to those that are not related to deverbal nouns. Noun modifiers continue to be marked as N<, as in the examples below:

1. nenhuma delas tem *medo* de não encontrar — N<ARG
2. A *poluição* das águas — N<ARGO (= poluir águas)
3. A *participação* de ONGs — N<ARGS (= ONGs participam)
4. A *poluição* de origem humana — N<

Another novelty of Bosque is the "searchable" tags, added to either terminal or non-terminal nodes or both, and introduced to simplify the search for some complex

structures, which can now be found looking for a single tag. At clause level, "searchable" tags were implemented marking the presence of elliptic subjects and types of subclauses (relative clauses, comparative clauses, consecutive clauses, etc.). Other topics included complex verbal tenses (marked on the main verb), passives, and partitives. Focusing on non-verbal structures, we revised "searchables" related to relative-clauses, substantive clauses and partitive constructions.

4 Easier: Milhafre

Since its inception, the usefulness of Floresta has been somewhat limited by the absence of an effective search interface/tool. There are several interfaces available, mainly for Bosque, such as CorpusEye (Bick, 2005) and the in-house developed Águia (*eagle*). Besides, several different formats of Bosque can be obtained from the website (Vilela et al., 2005) for use with other tools - including the TigerXML format, for use with TigerSearch (Lezius 2002), or the PennTreebank, which can be used e.g., with TGrep2 (Rohde, 2005). However, we didn't consider these tools ideal, considering the richness of Floresta and its typical user.

As a first stage, we updated Águia to deal with the changes in format. Águia uses the CQP toolkit (Christ el al., 1999); this toolkit is however not appropriate for searches in tree structures, and doesn't handle well the nested structures which are usual in syntactic trees. Therefore, we chose to use Tgrep2, a tool appropriate to that kind of search, and developed an interface, Milhafre (*goshawk*), which allows the user to bypass both Tgrep2's complex syntax and the need to learn the extensive list of tags used in Floresta. This new JavaScript-based interface handles the users' requests and transforms them into a query to be answered by TGrep2.

Currently, the system handles not only searches for words, structures, PoS, and their functions, but for also lemma, morphology, and "searchables" mentioned above. Milhafre may return also aggregate results (like the distribution by function of NPs, or the distribution by lemma of prepositions following NPs). All results are made available in text format as well.

5 Concluding Remarks

In this paper, we presented some of the new features of Floresta Sintá(c)tica – its size, interface and linguistic information. We know that size is a crucial factor in a Treebank, as is a friendly search interface. That is the reason Selva continues to undergo revision, and the Milhafre search tool is still improving. Rather than subscribing to one specific school of syntax, our linguistic options try to suit the widest range of linguistic users, reinforcing our main role as resource providers for research on Portuguese PLN and corpus studies.

Acknowledgement

This work was done in the scope of the Linguateca, contract nº339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union.

References

Afonso, S., Bick, E., Haber, R., Santos, D.: Floresta sintá(c)tica: um treebank para o português. Actas do XVII Encontro da Associação Portuguesa de Linguística (APL) (2000)

Bick, E.: The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)

Bick, E.: CorpusEye: Et brugervenligt web-interface for grammatisk opmærkede korpora. In: Widell, P., Kunøe, M. (eds.) 10. Møde om Udforskningen af Dansk Sprog, Proceedings. Århus University (2005)

Christ, O., Schulze, B.M., Hofmann, A., Koenig, E.: The IMS Corpus Workbench: Corpus Query Processor (CQP): User'sManual. Institute for Natural Language Processing, University of Stutgart (CQP v2.2) (1999)

Lezius, W.: TIGERSearch - Ein Suchwerkzeug für Baumbanken. In: Busemann, S. (ed.) Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002). Saarbrücken (2002)

Rohde, D.: TGrep2 User Manual, version 1.15, May 10 (2005)

Santos, D., Bick, E.: Providing Internet access to Portuguese corpora: the AC/DC project. In: Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhauer, G. (eds.) Proceedings of LREC (2000)

Vilela, R., Simões, A., Bick, E., Almeida, J.J.: Representação em XML da Floresta Sintáctica. In: Ramalho, J.C., Simões, A., Correia Lopes, J. (eds.) XATA 2005 (2005)