CINTIL-Treebank Searcher

Patricia Nunes Gonçalves, António Branco

NLX-Natural Language and Speech Group, Lisbon University, Portugal

{patricia.nunes, antonio.branco}@di.fc.ul.pt

Abstract

This is a short note to support the demonstration of the CINTIL-Treebank Searcher.

Index Terms: treebank, syntactic analysis, search, Portuguese

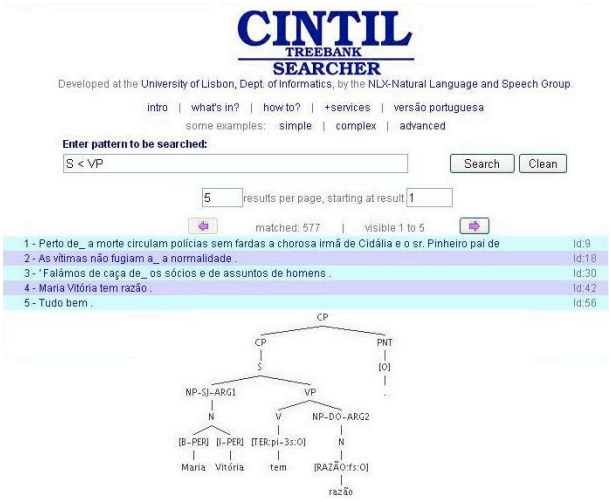
1. The CINTIL-Treebank Searcher

The CINTIL-Treebank Searcher is a freely available online service that permits to search the CINTIL-treebank and to visualize the syntactic analysis of the selected sentences.

This service is made available aiming at supporting research and development in the realm of natural language science and technology. It is well suited to be used as a research tool on the syntactic structure of Portuguese both by students and advanced researchers working on Linguistics, Natural Language Processing, or any other area involving the grammatical study of the Portuguese language.

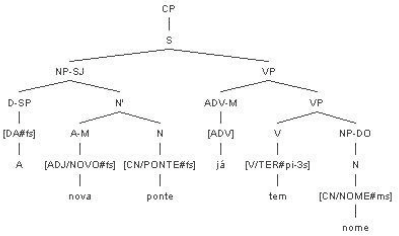
This online service for treebank searching and visualization was developed and is being maintained and extended by the NLX-Natural Language and Speech Group (`http://nlx.di.fc.ul.pt`), of the University of Lisbon, Department of Informatics.

This service receives a description of a syntactic structure pattern as input, entered by the user, and returns the list of sentences whose syntactic representation conforms to that pattern. Subsequently, by clicking on one of the listed sentences, the user obtains the syntactic tree of that sentence.



To describe the syntactic pattern he wants to search for, the user resorts to a description language based on regular expressions extended to easily capture basic relationships among the inner components and tags of a syntactic tree. The search system builds on the Tregex library [3], available from Stanford University, as the underlying search engine for tree query.

In order to briefly illustrate the syntax of the language used for describing syntactic patterns to be searched for in the treebank, consider the input expression `S < VP << NP-DO`. This query matches any tree containing a top-to-bottom node path where the sentence node (`S`) immediately dominates a verb phrase (`VP`), which in turn dominates (possibly non immediately) a noun phrase (`NP`) bearing a direct object grammatical function (`NP-DO`). The figure below shows a tree found by the search based on the query:



A fully-fledged description of the query language, together with key examples, are provided in one of the web pages made available in the The CINTIL-Treebank Searcher site.

2. The CINTIL Treebank

The CINTIL-Treebank is a corpus of sentences annotated with their syntactic trees, that encode the constituency relations among their elements. The treebank is composed of sentences from the CINTIL-International Corpus of Portuguese [1] and it is developed at the NLX Group.

The annotation of the CINTIL Treebank is performed by experts in Linguistics according to the mainstream method of annotation that is deemed to ensure a more reliable outcome: multiple annotation by independent annotators, followed by adjudication.

The annotation work is supported and its quality and consistency is ensured by resorting to a computational grammar. Each sentence is automatically analyzed by LXGram [2], an advanced grammar for the deep linguistic processing of Portuguese. Once a parse forest is obtained for a given sentence, independent annotators choose the analysis they each consider to be correct. In case of divergence between annotators, an adjudicator make a final decision.

3. References

[1] Barreto, F.; Branco, A.; Ferreira, E.; Mendes, A.; Nascimento, M.; Nunes F. and Silva J., 2006, "Open Resources and Tools for the Shallow Processing of Portuguese", Proceedings of the 5th LREC, 2006. Genova, Italy.

[2] Branco, A. and Costa, F., "A Computational Grammar for Deep Linguistic Processing of Portuguese: LXGram, version A.4.1", University of Lisbon, 2008.

[3] Levy, R. and Andrew, G., "Tregex and Tsurgeon: tools for querying and manipulation tree data structures", In Proceedings LREC, 2006.