Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties

Diana Santos
Linguateca, SINTEF ICT

In this paper I present briefly Linguateca[1], an infrastructure project for Portuguese which is ten years old, and will show how it provides several possibilities to study grammatical and semantical differences between varieties of the language.

After a short history of Portuguese corpus linguistics, presenting the main projects in the area, I discuss in some detail the AC/DC project (Santos & Bick, 2000), the Floresta Sintáctica treebank (Afonso et al., 2002, Freitas et al., 2008, Bick, 2004), and sketch some ideas for parallel corpora as started in CorTrad[2] (Tagnin et al., 2009).

I will use three different kinds of examples: those related to known differences between variants, in both grammar and lexis, those related to diachronic differences, in that respect describing in detail Silva (2008, in press) model of Quantitative Lexicology and Variational Linguistics[3] in CONDIVport, and those that are in a way corpus-driven and for which novel functionalities of AC/DC have been devised, namely the comparison of two search expressions; and the pattern database.

The paper will be structured as follows

1. a short introduction to Linguateca, for an international audience instead of to a Portuguese-speaking one (Santos, 2009)
2. some history of Portuguese corpora and corpus linguistics, surveying the main projects internationally that provide similar data or systems
3. present in a nutshell three ongoing Linguateca projects that deal with syntactical analysis of running text (AC/DC, Floresta and CorTrad)
4. provide some examples of variety studies with the above projects
    a. how to go about studying known differences
    b. a model of convergence and divergence of national varieties
    c. new functionalities for corpus-based discovery

References

Susana Afonso, Eckhard Bick, Renato Haber & Diana Santos. 2002. "Floresta sintá (c)tica: a treebank for Portuguese". In Manuel González Rodrigues & Carmen Paz Suarez Araujo (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)* (Las Palmas de Gran Canaria, Espanha, 29-31 de Maio de 2002), Paris : ELRA, pp. 1698-1703.

Eckhard Bick. 2004. “Looking at the Floresta Sintá(c)tica with a CorpusEye: A user-friendly cross-language search interface”. http://www.linguateca.pt/Floresta/floresta-corpuseye_en.pdf.

Cláudia Freitas, Paulo Rocha & Eckhard Bick. 2008. "Floresta Sintá(c)tica: Bigger, Thicker and Easier". In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira

[1] www.linguateca.pt, financed currently by UMIC and FCCN.
[2] CorTrad is a subproject of COMET – Corpus Multilíngüe para Ensino e Tradução, from University of São Paulo, whose Web access is done through DISPARA (Santos, 2002), by Linguateca in cooperation with NILC, see http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html.
[3] See http://wwwling.arts.kuleuven.be/qlvl/ for the general approach.

& Paulo Quaresma (eds.), *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)* Vol. 5190, (Aveiro, Portugal, 8-10 September 2008), Springer Verlag, pp. 216-219.

Diana Santos. 2002. "DISPARA, a system for distributing parallel corpora on the Web", in Elisabete Ranchhod & Nuno J. Mamede (eds.), *Advances in Natural Language Processing (Third International Conference, PorTAL 2002, Faro, Portugal, June 2002, Proceedings)*, LNAI 2389, Springer, 2002, pp.209-218.

Diana Santos. 2009. "Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva". *Linguamática* **1** (2009), May 2009, pp. 25-59. http://linguamatica.com/index.php/linguamatica/article/view/20/9

Diana Santos & Eckhard Bick. 2000. "Providing Internet access to Portuguese corpora: the AC/DC project", in Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), pp. 205-210.

Augusto Soares da Silva. 2008. "O corpus CONDIV e o estudo da convergência e divergência entre variedades do português". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 25-28. http://www.linguateca.pt/LivroL10/Cap04-Costaetal2008-Silva.pdf

Augusto Soares da Silva. In press. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In: Geeraerts, Dirk, Gitte Kristiansen & Yves Peirsman (eds.), *Cognitive Sociolinguistics*. Berlin/New York: Mouton de Gruyter.

Tagnin, Stella E. O., Elisa Duarte Teixeira & Diana Santos. 2009. "CorTrad: a multiversion translation corpus for the Portuguese-English pair", *28th International conference on lexis and grammar* (Bergen, Norway, 30 September – 3 October 2009).