The Tycho Brahe Corpus of historical Portuguese: methodology and results
Charlotte Galves, University of Campinas

In the first part of the talk, the *Tycho Brahe Corpus* will be presented. This corpus is currently composed of 57 texts (2,547.503 words) written in Portuguese by authors born between 1386 and 1845 (cf. www.tycho.iel.unicamp.br/~tycho/corpus). At the present time, the great majority of the texts are from Portuguese authors, but a great number of Brazilian texts are currently being included, in cooperation with Brazilian teams working on the edition of non-literary texts produced in Brazil from the 17th century on. All the texts are formatted in XML, which allows them to be presented both in their original version and in a standardized version. The annotation systems were adapted to Portuguese from the *Penn Parsed Corpora Project*, and they could be easily transferred to other Romance languages. An automatic Part of Speech (POS) tagger was developed, whose accuracy is 95%. Presently, 33 texts are available with a corrected POS tagging. As for the syntactic parsing, the team has been training Dan Bickel's universal parser. The output of the parser is corrected partly by hand, and partly by using the revision function of the automatic search tool *Corpus Search*. Presently, 16 corrected parsed texts (c. 750,000 words / 34,280 sentences) are available for syntactic research, and many other texts, representative of different periods and genres, will be soon released. It must be emphasized that the unrestricted access to the texts is an important feature of the *Tycho Brahe Corpus*, since it allows one to study in detail the way syntax and information structure interact, both at the level of entire periods and internal to individual authors.

In the second part of the talk, results obtained thus far, thanks to the syntactically annotated part of the Corpus, will be presented. These results mainly concern aspects of Portuguese that have changed over time: clitic placement, subject position, the position of the verb, the use of determiners, and the syntax of infinitival clauses. The quantitative analyses made possible by the access to large quantities of data gave rise to a much more precise picture of the evolution of Portuguese grammar from the 16th century on, and locate the change from Classical to Modern European in authors born at the beginning of the 18th century, in which the syntax of clitics and the syntax of subjects concomitantly change. Moreover, thanks to the development of the Brazilian part of the corpus, we are now able to initiate a comparative history of the syntax of Brazilian and European Portuguese from the 16th century on, which was totally inexistent up to now.