ABSTRACT

Zenaide de Oliveira Novais Carneiro (UEFS) (zenaide.novais@gmail.com)
Mariana Fagundes de Oliveira Lacerda (UEFS) (marianafag@gmail.com)
Igor Leal Souza (UNICAMP – Master student) (lealsigor@gmail.com)
Priscila Starline Estrela Tuy Batista (USP - PHD student) (priscilatuy@gmail.com)
Shirley Cristina Guedes dos Santos (UNICAMP) (shirleycgs@gmail.com)

This paper discusses the process of a database constitution inside the universe of Digital Humanities, presenting the experience in annotated *corpora* composition and its linguistic and computational aspects within the scope of the Electronic *Corpus* of Historical Documents of the Sertão Project – CE-DOHS (www.uefs.br/cedohs). The CE-DOHS Project is part of the Interdisciplinary Research Group on Digital Humanities of the State University of Feira de Santana (UEFS) and aims to carry out the digital edition of text from the Historical Documents of the Sertão, in partnership with of the Project Voices from Sertão in Data: history, people and formation of Brazilian Portuguese, one of the projects of the Studies in Portuguese Language Center of UEFS, as well as their morphological and syntactic annotation, drafting one annotated diachronic *corpus* that serve as electronic resource for the linguistic study of Brazilian Portuguese. It is intended to contribute with the studies of Brazilian Portuguese History Project (PHPB) on different theoretical perspectives and through a technological partnership with the *Tycho Brahe Parsed Corpus of Historical Portuguese* (www.tycho.iel.unicamp.br). To edit the collections in the XML language, it is used the eDictor (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010), a computational tool specially developed for philological work and automatic linguistic analysis[1]. The CE-DOHS Project has a set of documents mainly originated from the Bahian semi-arid area, the total partial: 16th century – letters (prospection); 17th century – 517 documents (letters, minutes, requirements, opinions and other), with 49.463 words; 18th century – 132 documents (letters, minutes, books from farms and other); 19th century – 440 letters, with 155.146 words; 20th century – 1,270 documents (letters, letters from readers and editors advertising and surveys), with 914.838 works. In all, 2,359 documents and 1,119,447 words. In order to constitute the CE-DOHS database, the project has been developed in phases: **Phase I** – Completed. As a result, several collections are available, especially handwritten letters organized by level of education and by degree of writing ability of the writer; 1084 private letters (1808-2000), totaling 350,850 words written by 422 writers (born between 1724 and 1980), most extracted from Carneiro et al (2011); Set of printed newspapers from *Folha do Norte* and *O Progresso*; and oral data extracted from the Collection *Samples of Spoken Language in the Bahian Semi-Arid* (Almeida and Carneiro, 2008). The collections edited in XML language are available in the CE-DOHS in different versions with respective document facsimiles: semi-diplomatic, modernized and in XML language. **Phase II** – In progress. In this phase, the collections available in XML language are in the process of **morphological**[2] and **syntactic**[3] annotation, based on the methodology

used by the *Tycho Brahe Project*[4]. Subsequently, the documents will receive the **annotation of chains of referents** proposed by Paixão de Sousa (2016) [5]. In addition, the number of documents has been expanded, incorporating documents from the 16th, 17th and 18th centuries (manuscripts and printed). All documents are organized by community, by type of linguistic contact and by trend (popular and cultured). This enlargement of the *corpus* "essentially favors a descriptive linguistics, strongly supported by the new technologies, and allows to take the quantitative analysis of authentic data as a starting point of the description, as it is done in other scientific domains". In addition to these materials, there are still the following **ongoing projects**: 1) A *corpus* for the 17th century (from 1617) documents written by Brazilians: Vieira Ravasco family and other contemporaries; 2) Letters and Minutes produced by "good men" of the City Council of Salvador, from the 17th century; 3) Application of linguistic and web-semantic annotation techniques in CE-DOHS; 4) Going back to the 18th century: private documents from the *Feira do Capuame* (1729-1830) and the *Sobrado do Brejo Seco* (1755-1830); 5) Thematic project: documents written by unskilled hands; 6) Insertion of the indigenous in the world of writing; 7) Oral databases of Afro-descendant communities in the West of Bahia and the tools; 8) Morphological and syntactical annotation of oral database: popular Portuguese of the Bahian Sertão and Salvador; 9) Refinando os *corpora*: polarização sociolinguística; separação por normas; níveis de escolaridade; normas, capital/interior; diferenciação diatópico-diacrônica e por gêneros textuais; 10) Refining the *corpora*: sociolinguistics polarization; organization by trends; schooling levels; linguistics norm, capital/interior; diatopic-diachronic differentiation and by textual genres; 11) Elaboration of computational tools (E-Corp and others) for construction and use of CE-DOHS. These documents have been used as a basis for the composition of a *PHPB Corpora Online Platform* (https://sites.google.com/site/corporaphpb), coordinated by Afrânio Barbosa, from the Federal University of Rio de Janeiro (UFRJ), and by Marcelo Módulo, from the University of São Paulo (USP). The documents of CE-DOHS are representative of diachronic varieties of Brazilian Portuguese (PB), from the 16th to the 20th century, from different regions of the country and different degrees of schooling. The material available at the database serves not only researchers interested in analysis of linguistic aspects, but also aspects of the diffusion of writing, reading, textual, historical, political, economic and social transmission, among others.