EDIÇÃO EM XML DO ACERVO CARTAS PARA SEVERINO VIEIRA

Marinalda Silva Freitas [1]; Zenaide de Oliveira Novais Carneiro [2].

1. Marinalda Silva freitas, bolsista de iniciação científica PEVIC, Graduanda em Letras Vernáculas, Universidade Estadual de Feira de Santana e-mail: marinaldafreitas@gmail.com.br

2. Zenaide de Oliveira Novais Carneiro, Departamento de Letras e Artes, Universidade Estadual de Feira de Santana, e-mail: zenaide.novais@gmail.com

PALAVRAS-CHAVE: Linguística de *corpus*, *corpus* eletrônico, Português brasileiro.

INTRODUÇÃO

A pesquisa sobre a história do Português Brasileiro (PB) constitui-se uma área bastante ampla, mas, ainda, com pouco material disponível para estudo e, por isso, com muitas lacunas a serem preenchidas. Cientes desta situação, há alguns anos, vem se desenvolvendo, no âmbito do banco de Documentos Históricos do Sertão, o DOHS, o Projeto Linguística de Corpus, prezando pela constituição de amostras empíricas de língua que se apresentam através de textos escritos no passado. Neste sentido, o Projeto Vozes do Sertão em Dados: história, povos e formação do português brasileiro (CNPq. Processo 401433/2009-9/Consepe: 102/2009), coordenado por Zenaide de Oliveira Novais Carneiro tem se dedicado à construção de banco de dados. Entretanto, como as amostras constituídas estão restritas a meios impressos ou digitais em formato não manipulável, o nosso objetivo é colaborar no transporte do meio impresso e digital de amostras já constituídas para meios eletrônicos no âmbito do Projeto CE-DOHS Corpus Eletrônico de Documentos Históricos do Sertão (www.uefs.br/cedohs), (FAPESB, Processo 5566/2010/Consepe:202/2010), coordenado por Zenaide de Oliveira Novais Carneiro e Mariana Fagundes de Oliveira, sediado no Núcleo de Estudos de Língua Portuguesa (NELP), na Universidade Estadual de Feira de Santana (UEFS). Trata-se da edição em formato XML, via utilização da ferramenta computacional E-dictor (PAIXÃO DE SOUZA; KEPLER; FARIA, 2009), através do Termo Aditivo de Transferência de Tecnologia do Corpus Histórico do Português Tycho Brahe (www.tycho.iel.unicamp.br), sediado no Instituto de Estudos da Linguagem (IEL) da Universidade Estadual de Campinas/UNICAMP, coordenado por Charlotte Marie Chambelland Galves. Com isso, esperamos contribuir para além da geração de edições na rede mundial de computadores, com bases para outras análises em nível de bancos anotados morfológica e sintaticamente de grande importância para o avanço dos estudos linguísticos no Brasil.

MATERIAIS E MÉTODOS

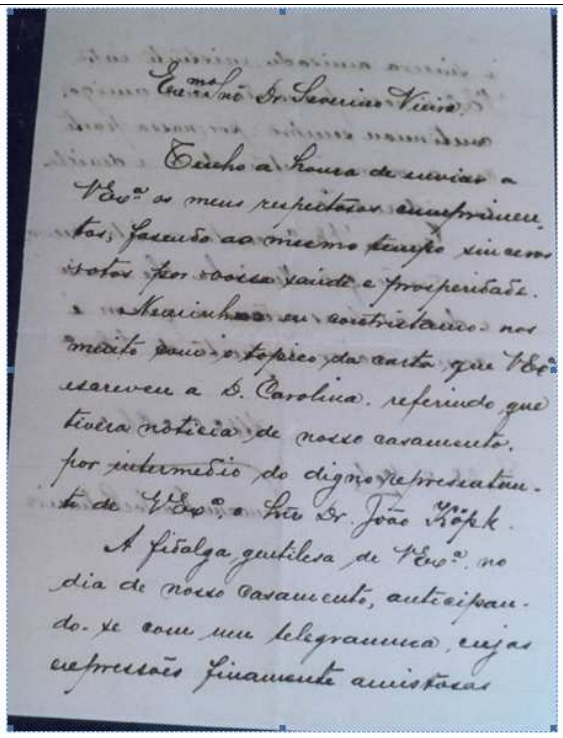
O material utilizado nesta pesquisa faz parte do banco de dados Documentos Históricos do Sertão/DOHS, do Projeto Vozes do Sertão em Dados (disponível em: www.uefs.br/nelp). Trata-se de uma documentação composta por um maço de 102 cartas (209-310), extraídas de Carneiro (2005), enviadas a Severino Vieira por 60 remetentes (57 homens e 3 mulheres), a maioria letrada e, sobretudo, citadina. As cartas

datam de 1901-1902, sendo 41 cartas de 1901 e 58 cartas de 1902, período que corresponde ao primeiro biênio do mandato de Severino Vieira, Governador da Bahia (1901-1904). A documentação é de extrema importância para história do português brasileiro, pois se trata de uma amostra diacrônica do português culto e semi-culto falado no Brasil, no século XIX.

A metodologia baseia-se fundamentalmente na metodologia do Projeto *Corpus Histórico do Português Tycho Brahe*, composto por um *corpus* eletrônico anotado de textos em português escritos por autores nascidos entre 1435 e 1845, desenvolvido desde 1998, em http://www.tycho.iel.unicamp.br/~tycho/corpus/, onde estão definidos os modelos e as ferramentas que estão subsidiando o Projeto CE-DOHS, um *corpus* voltado a um banco inédito de textos do sertão da Bahia. Esse tipo de banco de dados vem-se mostrando uma tendência mundial, com um grande número de projetos, sendo o pioneiro o projeto *Penn Helsinki Parsed Corpus of Middle English* (http://www.ling.upenn.edu/hist-corpora/), coordenado por Anthony Kroch na Universidade da Pensilvânia.

Os documentos submetidos à edição eletrônica foram feitos no formato de edições semidiplomáticas e fac-similadas pertencentes ao Projeto Vozes do sertão em dados supracitado acima. (c.f. figura1).

Figura 1:



Carta 267

AIGHBA. Cx. ASV. Documento contendo dois fólios. Papel almaço com pautas. Tarja preta nas bordas do 1º fólio.

Excelentissimo Senhor Dr. Severino Vieira |

Tenho a honra de enviar a| VossaExcelência os meus respeitosos cumprimen-|tos, fazendo ao mesmo tempo sinceros| votos por vossa saúde e prosperidade.|

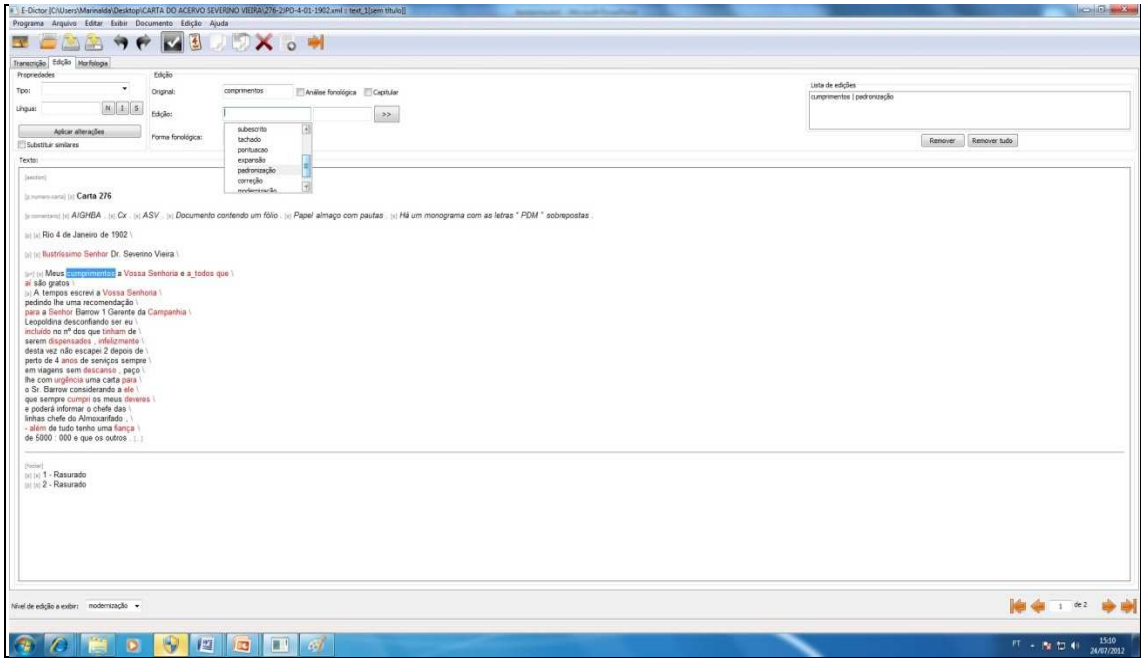
Neusinha e eu constristamo-nos| muito com o topico da carta, que VossaExcelência| escreveu a D. Carolina, referindo, que| tivera noticia de nosso casamento,| por intermedio do digno representan-|te de VossaExcelência, o Senhor Dr. João Käpk. |

A fidalga, gentilesa, de VossaExcelência no| dia de nosso casamento, antecipan-|do-se com um telegramma, cujas| expressões finamente amistosas|

O programa que intermedia a linguagem XML nos textos editados em Word, o E-dictor, permite o controle das várias intervenções realizadas no texto, bem como a conservação da versão original do texto – o que valida o método entre os filólogos pelo fato da conservação do valor filológico dos textos.

Abaixo, as figuras 2 e 3 referem-se a edição em XML utilizando o E-dictor.

Figura 2:



RESULTADOS E/ OU DISCUSSÃO

Ao termino do trabalho, conseguiu-se concluir a edição de mais da metade do acervo Cartas para Severino Vieira em linguagem XML, contabilizando o total de 60 cartas.

No site do CE-DOHS (http://www.uefs.br/cedohs/), está disponível, em linguagem XLM, parte do acervo Cartas para Severino Vieira, bem como cartas de outros acervos, todos integrantes do Banco DOHS, do Projeto Vozes do Sertão em Dados.

A figura 3, abaixo, traz a imagem da carta em linguagem XML na forma como está disponível no site através da geração de edições, nesse caso, uma do tipo modernizada, além de outras em versão semidiplomática, técnica e também do léxico de edições que podem ser consultadas em http://www.tycho.iel.unicamp.br/cedohs/corpora/catalog-SV.html.

Figura 3:

CONSIDERAÇÕES FINAIS

O presente trabalho, no estágio atual, ainda não oferece, através de seus textos, a busca automática de dados, mas possibilita a geração *on line* de vários tipos de edição, técnica, semidiplomática, modernizada e o léxico, além de fichas de metadados. Nesta primeira fase do CE-DOHS, objetivou-se o inicio da constituição de um banco de dados eletrônicos para, futuramente, receberem anotação morfológica e sintática. A partir das anotações, será possível a busca automática de dados, a qual facilitará o trabalho dos linguistas, pois proporcionará a análise lingüística automática.

REFERÊNCIAS

CARNEIRO, Zenaide. *Cartas Brasileiras*: *um estudo lingüístico-filológico.* Tese de Doutorado, Campinas: Unicamp, 2005.

CE-DOHS – Documentos Históricos Do Sertão Em Dados (disponível em http://www2.uefs.br/cedohs/), 2011.

CORPUS DOHS. Documentos Históricos do Sertão (disponível em http://www.uefs.br/dohs/), 2010.

PAIXÃO DE SOUZA, M.C., KEPLER, F.N. & FARIA, P. (a sair) *"E-Dictor: novas perspectivas na codificação e edição de corpora de textos históricos"*. In: Shepherd, T., Berber Sardinha, T. e Veirano Pinto, M. (2009) (Org.). *Linguística de Corpus: Sínteses e Avanços. Anais do VIII Encontro de Linguística de Corpus*, realizado na UERJ, 13 a 14 de novembro de 2009. Rio de Janeiro, RJ.

PAIXÃO DE SOUSA, M.C. *"Memórias do Texto"*. Revista Texto Digital. Universidade Federal de Santa Catarina: 2006.

PROJETO VOZES DO SERTÃO EM DADOS (disponível em http://www.uefs.br/nelp/), 2010.