EDIÇÃO EM LINGUAGEM XML DO ACERVO CARTAS PARA SEVERINO VIEIRA

Marinalda Silva Freitas [1]; Mariana Fagundes de Oliveira[2].

1. Marinalda Silva freitas, bolsista de iniciação científica PROBIC, Graduanda em Letras Vernáculas, Universidade Estadual de Feira de Santana e-mail: marinaldafreitas@gmail.com.br
2. Mariana Fagundes de Oliveira, Departamento de Letras e Artes, Universidade Estadual de Feira de Santana, e-mail: marianafag@gmail.com

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

INTRODUÇÃO

Como se sabe, as pesquisas desenvolvidas, no âmbito da Linguística Histórica, tratam-se de estudos bastante amplos e complexos, cujo objeto de estudo são amostras de língua presentes em textos escritos no passado. Tratando-se da pesquisa para a constituição do Português Brasileiro (PB), a falta de amostras diacrônicas constitui um dos maiores problemas ao trabalho do historiador da referida língua. Ciente disso, o projeto *Corpus Eletrônico de Documentos Históricos do Sertão* (CE-DOHS) (www.uefs.br/cedohs), que se trata de uma interface de outros projetos que já faziam linguística de *corpus*, porém em versão não eletrônica, desenvolve, atualmente, uma versão digital do Banco Documentos Históricos do Sertão (DOHS), do projeto Vozes do Sertão em Dados. Para colaborar com o CE-DOHS e com o Projeto para a História do Português Brasileiro/PHPB, realizou-se a edição, em linguagem XML, usando o E-dictor, da última parte das cartas pertencentes ao acervo *Cartas para Severino Vieira*, concluindo, assim, a edição de todo acervo. Este trabalho, bem como outros pertencentes ao projeto CE-DOHS, faz a transposição dos textos editados em Word para a versão em linguagem XML, usando a ferramenta computacional E-Dictor. Através do programa computacional E-Dictor, o qual foi desenvolvido por Kepler, Paixão de Souza e Faria (2007), para facilitar a edição eletrônica em linguagem *XML*, realiza-se a edição dos documentos. Esse programa amplia o alcance da linguagem *XML* e consegue unir a edição e a correção morfológica do texto em um mesmo ambiente. A primeira versão dessa ferramenta surgiu para atender a uma demanda obtida durante a criação do Projeto *Corpus Anotado do Português Tycho Brahe* (CTB/UNICAMP) e em atividades da equipe deste projeto com a equipe do *Projeto para a História da Língua Portuguesa* (PROHPOR/UFBA). Por saber da importância filológica do acervo utilizado nesta pesquisa, cuja edição é semi-diplomática, a edição em linguagem XML usando o E-Dictor resguarda a versão original dos textos; desta forma, preserva seu valor filológico. Para isso, são mapeadas todas as intervenções feitas no texto no que se refere à correção ortográfica, junção, segmentação e expansão de palavras, bem como outros tipos de intervenções necessárias.

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

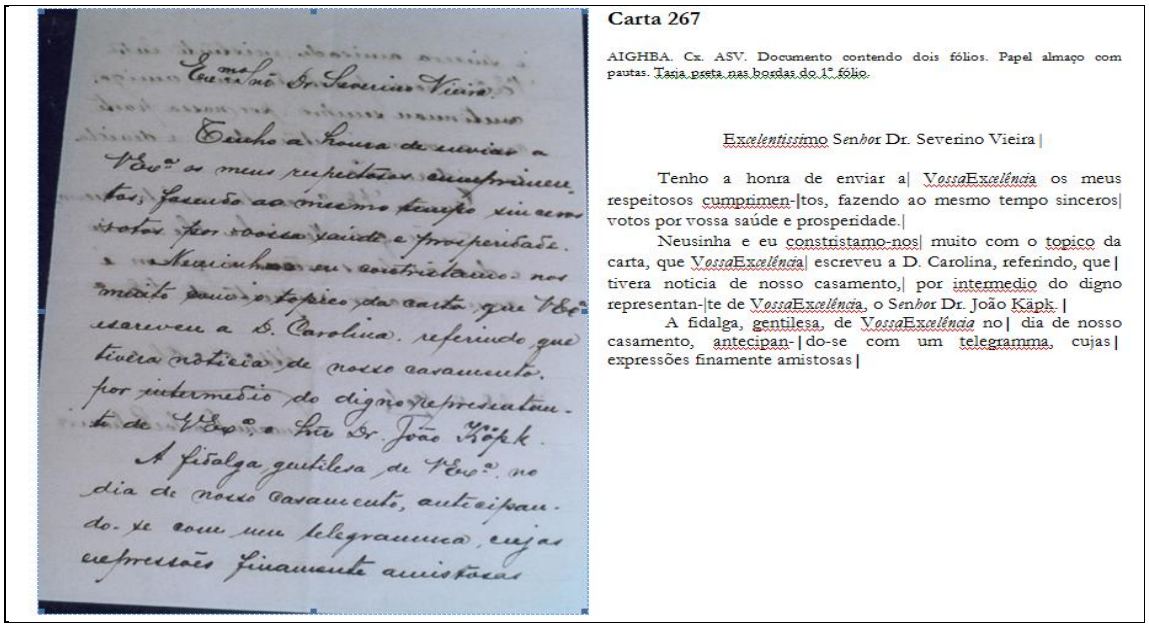
datam de 1901-1902, sendo 41 cartas de 1901 e 58 cartas de 1902, período que corresponde ao primeiro biênio do mandato de Severino Vieira, Governador da Bahia (1901-1904). A documentação é de extrema importância para a história do português brasileiro, pois se trata de uma amostra diacrônica do português 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 ofMiddleEnglish* (http://www.ling.upenn.edu/hist-corpora/), coordenado por Anthony Kroch na Universidade da Pensilvânia.

No plano de trabalho anterior, como PEVIC, deu-se início à edição de parte deste acervo. Como bolsista PROBIC, o plano de trabalho objetiva a edição total do acervo *Cartas para Severino Vieira.* Isso aconteceu antes do tempo previsto para execução de todo o plano de trabalho, o que possibilitou a contribuição na edição de outros acervos que também integrarão o banco eletrônico do CE-DOHS.

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: Imagem da edição semidiplomática.



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.|

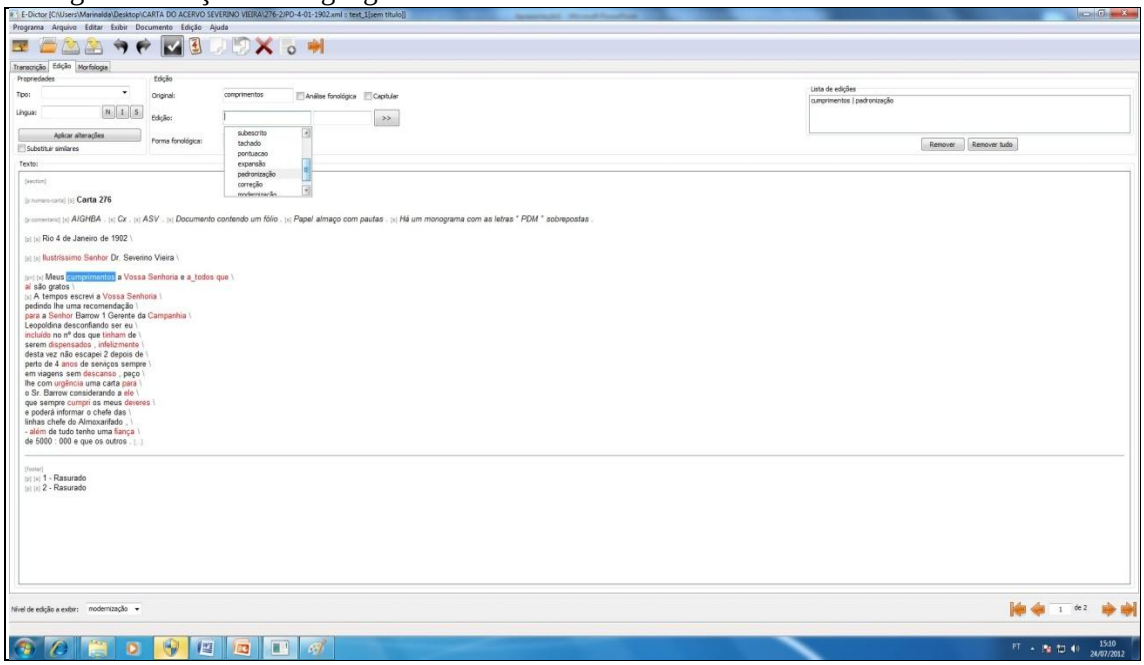
Neusinha e eu constristamo-nos| muito com o topico da carta, que VossaExcelência| escreveu a D. Carolina, referindo, que| tivera noticia de nosso casamento,| por intermedio do digno representan-|te de VossaExcelência, o Senhor Dr. João Käpk. |

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 à edição em XML utilizando o *E-dictor*.

Figura 2: Imagem da edição em linguagem XML.



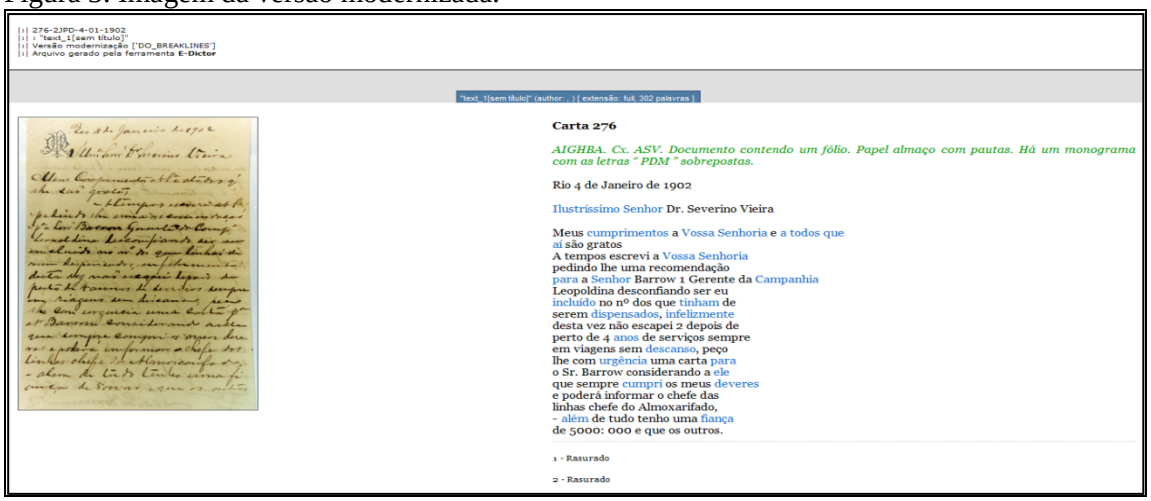
RESULTADOS E/ OU DISCUSSÃO

Encontra-se concluída a edição, em linguagem XML, de todo o acervo *Cartas para Severino Vieira*.

No site do CE-DOHS (http://www.uefs.br/cedohs/), está disponível, em linguagem XLM, o acervo completo *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. Esta última versão trata-se de uma lista com todas as palavras que sofrem algum tipo de modificação durante a edição em linguagem XML usando o *E-dictor*. O então chamado léxico de edição é composto por duas colunas: uma coluna mostra as palavras como estavam no original, e a outra coluna mostra como as palavras ficaram depois das intervenções realizadas durante a edição eletrônica. Todas as versões podem ser consultadas em http://www.tycho.iel.unicamp.br/cedohs/corpora/catalog-SV.html.

Figura 3: Imagem da versão modernizada.



CONSIDERAÇÕES FINAIS

O presente trabalho ainda não oferece a busca automática de dados, mas possibilita a geração *online* de vários tipos de edição, técnica, semidiplomática, modernizada, bem como o léxico de edição que é gerado, assim como as outras versões de edição, a partir da consulta às cartas pertencentes ao acervo eletrônico do CE-DOHS, disponível no *site* http://www.uefs.br/cedohs/), além de fichas de metadados. Nesta primeira fase do CE-DOHS, objetivou-se o início 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 linguística automática.

REFERÊNCIAS

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

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.