ESTUDO E APLICAÇÃO DE TÉCNICAS PARA ACERVOS DIGITAIS PARA USOS LINGUÍSTICOS EM LINGUAGEM XML

Amanda Lopes de Souza Martins [1]; Zenaide de Oliveira Novais Carneiro[2]

1. Bolsista FAPESB de Iniciação Científica Jr, Estudante do 3º Ano do Ensino Médio, Universidade Estadual de Feira de Santana, e-mail: nana54321@hotmail.com

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

PALAVRAS-CHAVE: Banco de Dados, Edição Eletrônica, Português Brasileiro.

INTRODUÇÃO

A linguística de *corpus* tem despertado o interesse de muitos pesquisadores que buscam estudar da história do português brasileiro (PB). Este trabalho de edição de textos feito a partir do banco DOHS do projeto *Vozes do Sertão em Dados: história, povos e formação do português brasileiro* (CNPq. Processo 401433/2009-9/Consepe: 102/2009), especificamente em parte do Acervo *Cartas para Severino Vieira, governador da Bahia (1901-1902),* editadas por Carneiro (2005) em uma versão computacional eletrônica em em linguagem XML como parte do Plano de Trabalho de I.C. Junior no Edital Ação Referência FAPESB/2010 no âmbito do projeto *Corpus Eletrônico de Documentos Históricos do Sertão* (CE-DOHS), (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). Esse banco eletrônico é feito em parceria com o *Projeto para a História do Português Brasileiro* (PHPB) e o *Corpus Histórico do Português Tycho Brahe* (http://Tycho.iel.unicamp.br/~tycho/corpus/).

METODOLOGIA

A metodologia baseia-se na Linguística Computacional para uso de banco de dados. A mesma utilizada pelo *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://Tycho.iel.unicamp.br/~tycho/corpus/, onde estão definidos as ferramentas e os modelos que estão subsidiando o projeto CE-DOHS, um *corpus* voltado a um banco de eletrônico. O trabalho é desenvolvido por fases, primeiramente realizamos a edição XML através do uso da ferramenta E-dictor (PAIXÃO DE SOUZA; KEPLER; FARIA, 2009) *(Figura 1).*

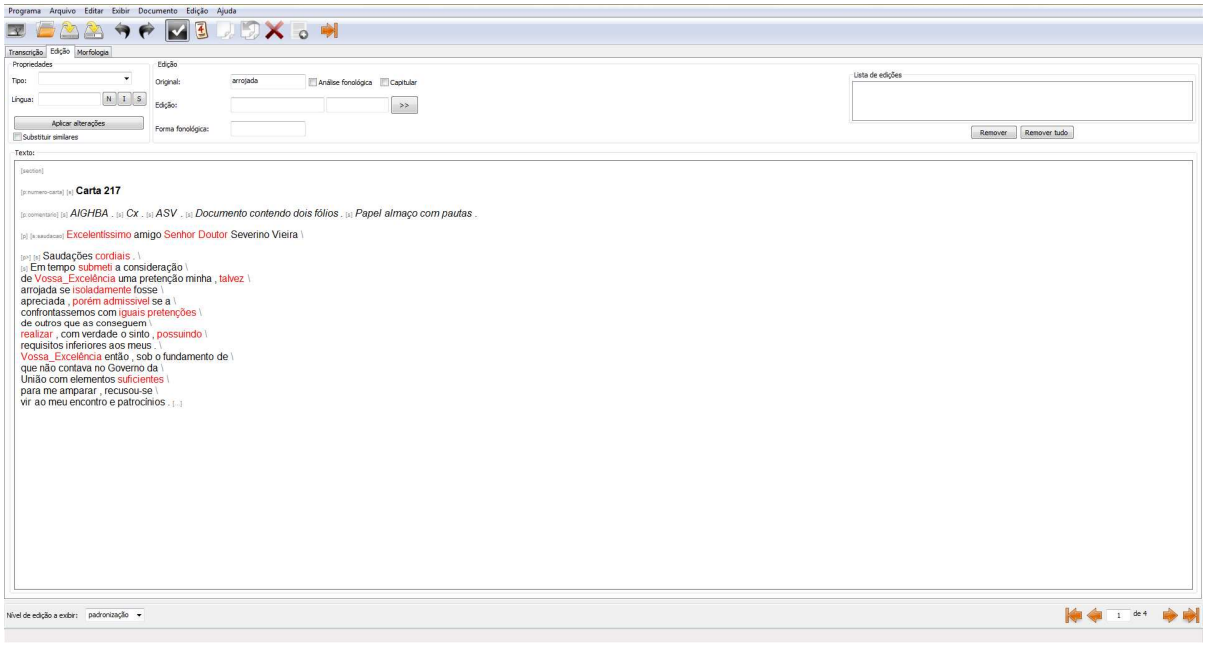

Figura 1: Modelo de edição utilizando o E-dictor.

Após gerar a versão XML da edição Semi-Diplomática Fac-similada, começamos a editar o documento corrigindo erros ortográficos, padronizando palavras e preenchendo os metadados *(Figura 2)*.

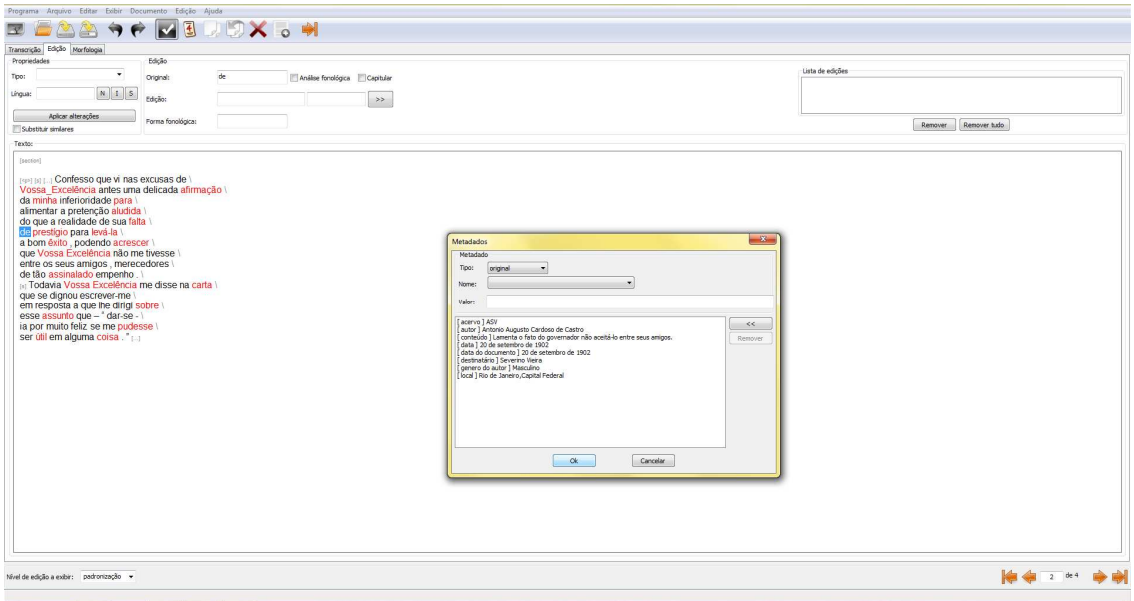

Figura 02: Modelo de edição e junto os metadados correspondentes feitos no E-dictor.

Os metadados servem como uma ficha catalográfica, contendo informações sobre a carta em questão, como o acervo, autor, data, conteúdo. Enfim, utilizando as técnicas de codificação digital para que o texto em questão possa ser lançado no banco de dados podendo ser visualizado no mundo inteiro.

RESULTADOS

Após o primeiro contato com o *corpus*, o trabalho da edição XML começou finalmente a ser desenvolvido e após a edição de parte do acervo de Severino Vieira, as cartas foram postadas no site. Esse resultado pode ser visto em http://www.tycho.iel.unicamp.br/cedohs/corpora/catalog-SV.html. Um trabalho que levou cerca de dois semestres para ser realizado, mas que proporcionou uma tamanha gratificação e a certeza de um trabalho bem desenvolvido sobretudo porque como I.C júnior, uma estudante do Ensino Médio e já pode contribuir com a comunidade universitária.

CONSIDERAÇÕES FINAIS

Temos enfim a edição em linguagem XML do Acervo *Cartas para Severino Vieira, governador da Bahia (1901-1902),* que constitui o *corpus* do CE-DOHS – Corpus Eletrônico de Documentos Históricos do Sertão (http://www2.uefs.br/cedohs/). O objetivo de contribuir para o estudo do Português Brasileiro com a composição de *corpora* anotados foi alcançado.

REFERÊNCIAS

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

CARNEIRO, Z. & C. GALVES (2010) "Variação e Gramática: Colocação de clíticos na história do português brasileiro", a sair em Revista de Estudos da Linguagem, UFMG.

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.

GALVES, Charlotte. *Ensaios sobre as gramáticas do português*. Campinas: Editora da Unicamp, 2001.

GALVES, C. (2010) Periodização e competição de gramáticas: o caso do português médio, a sair em LOBO, Tânia; CARNEIRO, Zenaide; RIBEIRO, Silvana; SOLEDADE, Juliana; ALMEIDA, Ariadne. (Orgs.) Coletânea de estudos em homenagem a Rosa Virgínia Mattos e Silva. Salvador: EDUFBA. (no prelo)

MATTOS E SILVA, Rosa Virgínia. (2002). Para a história do português culto e popular brasileiro: sugestões para uma pauta de pesquisa. In: ALKMIM, Tânia M. *Para a história do português brasileiro*: novos estudos. São Paulo: Humanitas/FFCHL/USP:FAPESP, v. 2, p. 443-464.

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.

LOBO, Tânia Conceição Freire . A questão da periodização da história lingüística do Brasil. In: Ivo Castro; Inês Duarte. (Org.). Razões e emoção. Miscelânia de estudos em homenagem a Maria Helena Mira Mateus.. Lisboa: Imprensa Nacional; Casa da Moeda, 2003, v. 1, p. 395-409.