APLICAÇÃO DE TECNOLOGIA COMPUTACIONAL EM BANCO DE DADOS EM LINGUÍSTICA HISTÓRICA: OS DOCUMENTOS DO CE-DOHS

Igor Leal Souza [1]; **Zenaide de Oliveira Novais Carneiro**[2]

1.Bolsista PIBIC/CNPq, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: igorengcomp@gmail.com
2. Orientadora, Departamento de Letras e Artes, Universidade Estadual de Feira de Santana, e-mail: zenaide.novais@gmail.com

Palavras-chave: Linguística de *Corpus*, *XML*, E-dictor

INTRODUÇÃO:

O projeto *Corpus Eletrônico de Documentos Históricos do Sertão* (CE-DOHS) (www.uefs.br/cedohs) tem o propósito de construir um banco de dados eletrônico. Nesse banco, serão disponibilizados documentos editados em linguagem *XML* com o auxílio da ferramenta E-Dictor, a qual permite que se façam edições de acordo com as necessidades das ferramentas para análise linguística, garantindo que a versão original possa ser recuperada caso seja necessário. A ideia central dessa técnica de anotação é o mapeamento das intervenções realizadas no texto. A nossa proposta foi investigar as vantagens da linguagem *XML* em banco de dados linguísticos históricos e o que precisa ser aprimorado na ferramenta em questão, o E-Dictor. A nossa base de aplicação foi o *corpus* do projeto CE-DOHS. O acervo que foi alvo dessa investigação é composto por 211 anúncios de jornais publicados no jornal *Folha do Norte*, datados de 1909 a 2006. Esse acervo foi editado e fac-similado em versão semi-diplomática e foi publicado no cd intitulado *Anúncios no Folha do Norte (1910-2006)*, da coletânea *PUBLICA-SE EM FEIRA DE SANTANA, Dos anúncios e das cartas de leitores e redatores no O Progresso e no Folha do Norte (1901-2006).*

METODOLOGIA:

A metodologia baseia-se fundamentalmente na metodologia do 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 as ferramentas e modelos que estão subsidiando o projeto CE-DOHS, um corpus voltado a um banco inédito de textos do Sertão da Bahia.

RESULTADO:

Com o auxílio dos recursos oferecidos pelo projeto, lançamos mão das novas tecnologias para dar um tratamento computacional aos textos editados, multiplicando assim suas finalidades, que além de construir um *corpus* histórico linguístico educacional, colabora também com diversas áreas de conhecimento, ampliando e otimizando o uso dos materiais históricos, garantindo também sua conservação.

Experimentando e observando o E-dictor com a aplicação no *copus*, e fazendo comparações com a edição tradicional, conseguimos constatar que a Edição Semi-diplomática fac-similada (Figura 1), além de ser mais trabalhosa por não oferecer o

recurso de fazer intervenções uma única vez, em uma palavra que necessite de padronização, por exemplo, e que se repita ao longo do texto, não destaca essas intervenções realizadas pelo editor, tornando-se propícia a erros, diferente do E-dictor (Figura 2), que oferece um mapeamento de todas as modificações realizadas. Um outro problema é não disponibilizar outras versões de edição, restrita apenas a edição Semi-Diplomática fac-similada, enquanto o E-Dictor disponibiliza várias versões do mesmo texto, além da Semi-Diplomática, a fac-similada, a edição Diplomática, o léxico de edições, e a ficha catalográfica do documento.

Novo Cinema em Feira

A direção do Cinema que brevo se inaugurará na Rua Castro Alves, atendendo a inú meras sugestões recebidas de várias procedências, propõe às pessoas de bem gosto de Feira as duas denominações «CINE MADRI» e «CINE GUARÁ» das quais aproveitará uma com que batizará o novo cinema. As cartas com a escolha ou mesmo novas sugestões devem ser encaminhadas à Rua Castro Alves, 1641 — Cinema Novo — no horário das 8 às 17 horas.

Agradece

A direção

N. 377-1-1-P

Anúncio 140

Novo Cinema| em Feira| A direção do Cinema que bre-|ve se inaugurará na Rua Cas-|tro Alves, atendendo a inúme-|ras sugestões recebidas de vá-|rias procedências, propõe às| pessoas de bem gosto de Fei-|ra as duas denominações <CI-|NE MADRI> e <CINE GUARÁ>| das quais aproveitará uma com| que batizará o novo cinema.|| As cartas com a escolha ou| mesmo novas sugestões devem| ser encaminhadas à Rua Cas-|tro Alves, 1641 – Cinema No-|vo – no horário das 8 às 17| horas.|| Agradece|| *A direção*|| N.377.1-1.P ||

Folha do Norte, 11 de janeiro de1958 (Ano XLVIII, N° 2530 p.1)

Figura 1: Modelo de edição semi-diplomática fac-similar.

Quando iniciamos a investigação, essa documentação encontrava-se em processo de edição em Linguagem XML. A edição já foi concluída e encontra-se disponível no banco eletrônico do CE-DOHS (www,uefs.br/cedohs). Em paralelo ao processo de edição, realizamos o trabalho de investigação das vantagens da edição em XML feita através do E-dictor e buscamos adequá-lo para uma melhor aplicação.

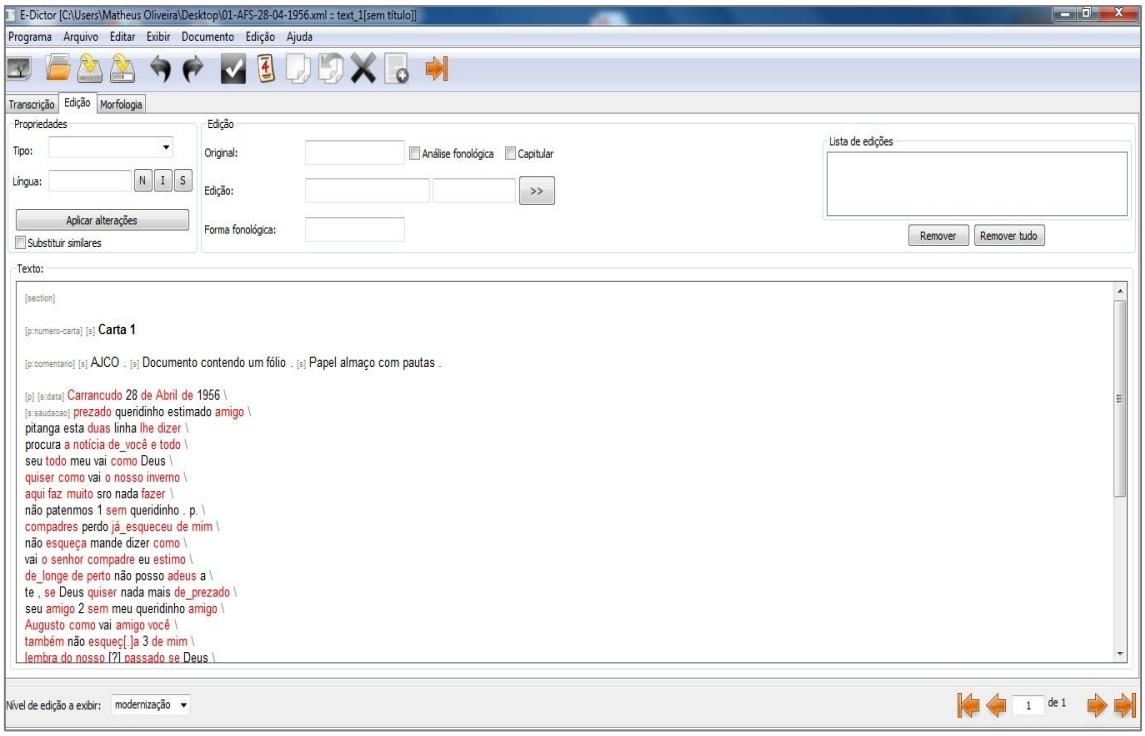

Figura 2: Modelo de edição em XML utilizando o E-Dictor.

Como resultado do trabalho desenvolvido, temos também a apresentação de pôster intitulado *"Periodics collection: letters from readers and redactors and advertisements of newspapers from Bahia"* no **Workshop Construction and use of large annotated corpora** realizado de 09 a 13 de setembro de 2013, no Instituto de Estudos da Linguagem da Universidade Estadual de Campinas. A participação nesse evento foi muito importante, uma vez que lá pudemos discutir as melhorias para o E-Dictor, chegando a conclusão de que precisamos desenvolver um aplicativo para dar suporte a edição com o E-Dictor, facilitando a visualização e o uso.

CONSIDERAÇÕES FINAIS:

O objetivo dessa investigação foi listar algumas opções de uso e aplicação da ferramenta E-Dictor. Acreditamos que as contribuições dadas foram significativas, tanto no incentivo da aplicação da ferramenta, que de fato otimiza a edição de documentos, quanto nos pontos a serem aprimorados.

REFERÊNCIAS:

CARNEIRO, Z. O. N. (2008). *Vozes do sertão em dados: história, povos e formação do português brasileiro.* In: VI Feira do Semi-Árido: desertificação, perspectivas de autonomia produtiva frente aos desafios socioambientais, 2008, Feira de Santana. VI Feira do Semi-Árido: desertificação, perspectivas de autonomia produtiva frente aos desafios socioambientais. Feira de Santana, v. 1.

CARNEIRO, Zenaide Novais (2005). *Cartas brasileiras* (1809-1907): um estudo filológico-linguístico. Campinas: UNICAMP. Tese de doutorado inédita.

CARNEIRO, Z. O. N. (2010). CE-DOHS. *Documentos Históricos do Sertão*. Projeto de Pesquisa.

CARNEIRO, Zenaide de Oliveira Novais; ALMEIDA, Norma Lucia F. de (2006*). A criação de escolas a partir de critérios demográficos na Bahia do século XIX: uma viagem ao interior*. In: LOBO, Tânia; RIBEIRO, Ilza; CARNEIRO, Zenaide de O. N.; ALMEIDA, Norma Lucia F. de. Para a história do português brasileiro: novos dados, novas análises. Salvador: Edufba, vol. 6, 1-2, p. 649-673.

CARNEIRO, Zenaide de Oliveira Novais; ALMEIDA, Norma Lucia F. de. (2007. *Elementos para uma sócio-história do semi-árido baiano.* In: RAMOS, J.; ALKMIM, Mônica A. Para a história do português brasileiro: estudos sobre mudança lingüística e história social. Belo Horizonte: Faculdade de Letras da UFMG, v.5. p. 423-442.

PAIXÃO DE SOUSA, M.C. *Memórias do Texto*. Revista Texto Digital. n. 2. Universidade Federal de Santa
Catarina. 2006.

PAIXÃO DE SOUSA, Maria Clara. *Projeto Memórias do Texto*. FAPESP-UNICAMP, 2004.

PAIXÃO DE SOUSA, M. C. & KEPLER, F. (2007). *E-Dictor: Uma ferramenta integrada para a anotação de edição e classe de palavras*. Comunicação ao VI Encontro de Lingüística de Corpus. USP. São Paulo, 7 de setembro.

SAXON. <http://saxon.sourceforge.net/>