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 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 Computacional, Edição Eletrônica, Português Brasileiro.

INTRODUÇÃO

Com o objetivo de desenvolver estudos sobre a história do português brasileiro (PB), muitos pesquisadores têm-se dedicado à *Linguística de Corpus*. Neste trabalho, em particular, damos seguimento ao Plano de Trabalho intitulado *a edição XML – estudo e aplicação de controle de edições em Documentos do sertão da Bahia (século XX - 1ª metade)*, desenvolvido no âmbito da bolsa de pesquisa Fapesb/Edital referência 2010, 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). Neste trabalho, em particular, como bolsista Pibic/CNPq, vamos mostrar os primeiros resultados do estudo sobre problemas operacionais e suas soluções da ferramenta integrada *E-dictor*, (Kepler e Paixão e Souza, 2004, 2009), principal ferramenta de interface de uso da linguagem XML no *corpus* do CE-DOHS que foi aplicado à edição para geração de edições eletrônicas, através do Termo Aditivo de Transferência de Tecnologia com o *Corpus Histórico do Português Tycho Brahe* (www.tycho.iel.unicamp.br).

METODOLOGIA

A metodologia baseia-se fundamentalmente na metodologia do *Corpus Histórico do Português Tycho Brahe* (www.tycho.iel.unicamp), sob a coordenação de Charlotte Galves, no Instituto de Estudos da Linguagem da Unicamp, composto por um banco 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. Esse tipo de banco de dados vem se mostrando uma tendência mundial.

RESULTADOS

Podemos ver como resultado, a aplicação das novas tecnologias de processamento de texto com aplicação na *Lingüística de Corpus* formando um banco de dados eletrônicos para fins linguísticos, visando a otimização no acesso aos documentos históricos do sertão da Bahia, através do uso da linguagem XML (*eXtended Markup Language*) (Figura 1) e da ferramenta integrada de anotação de corpus, E-dictor (Figura 2) desenvolvido por Faria, Kepler e Paixão e Souza (2004-2009).

O XML é uma linguagem de marcação que permite a edição/etiquetagem de uma palavra ou frase. Essa técnica permite que se façam edições de acordo com as necessidades das ferramentas para análise linguística, facilitando o trabalho do editor dando maior liberdade na edição. Dessa forma, as palavras podem ser facilmente modificadas. A estrutura XML para codificação no E-Dictor, tem dois objetivos primordiais: ser o mais neutra possível no que concerne ao conteúdo textual codificado e atender a necessidades linguísticas e filológicas.

A edição em linguagem XML possibilita o controle e o mapeamento das intervenções realizadas nos documentos, de forma que podemos editá-los e prepará-los para buscas automáticas de análise linguística e, ao mesmo tempo, garantir a recuperabilidade das formas originais. Por ser a linguagem XML de difícil operação para iniciantes em linguagens computacionais, foi desenvolvida a ferramenta E-dictor (cf. PAIXÃO DE SOUZA & KEPLER, 2007) para fazer a mediação entre o editor e o XML, a qual lançamos mão para realizar as edições. O E-dictor amplia o alcance do XML e torna-o mais confiável e simples de usar, agilizando o trabalho de edição dos textos, já que não é preciso a modificação dos textos diretamente com a linguagem.

O E-Dictor permite que outros conteúdos sejam codificados, como: metadados, elementos do texto em geral, classes de palavras, níveis de edição filológica e comentários do editor, enriquecendo o material. O E-dictor é flexível o suficiente para contribuir em outros contextos de construção de *corpora* de textos. Assim, transformamos os documentos, que estão fac-similados em *word*, para o *XML*, visando à composição do *corpus* eletrônico do Projeto *Vozes*, o *Corpus Eletrônico de Documentos Históricos do Sertão* (CE-DOHS/FAPESB, Processo 5566/2010/Consepe:202/2010).

O CE-DOHS em fase final da primeira parte possibilitará a busca automática de dados. O objetivo com esse tipo de banco de dados em universidades brasileiras, como a UEFS, é contribuir com o estudo do português brasileiro, com materiais editados de forma segura, usando a tecnologia confiável, garantindo os aspectos tradicionais na preparação de amostras linguísticas históricas (expansão de abreviaturas, uniformização de pontuação, modernização de grafia etc.), mas, agora, em linguagem manipulável, com ferramentas de busca.

A Figura 1 traz um exemplo do uso da linguagem XML e a Figura 2, uma imagem da ferramenta E-dictor. E a Figura 3, uma imagem do site já com o *corpus* disponível.

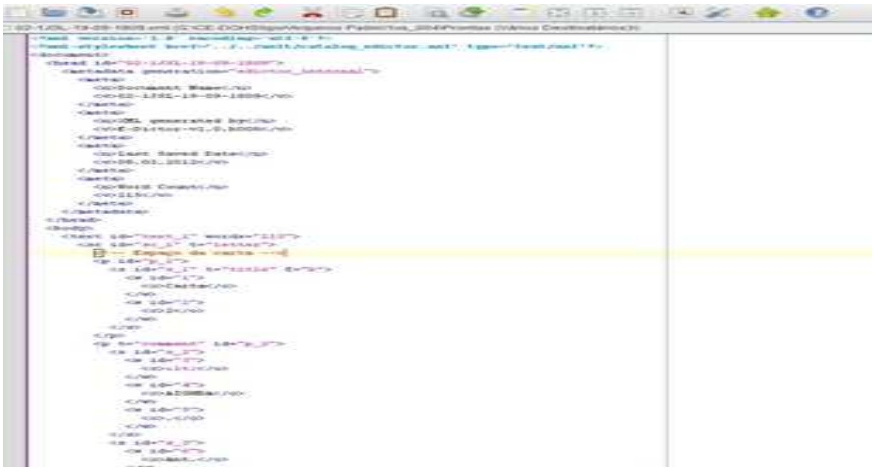

Figura 1: Exemplo da linguagem *XML*

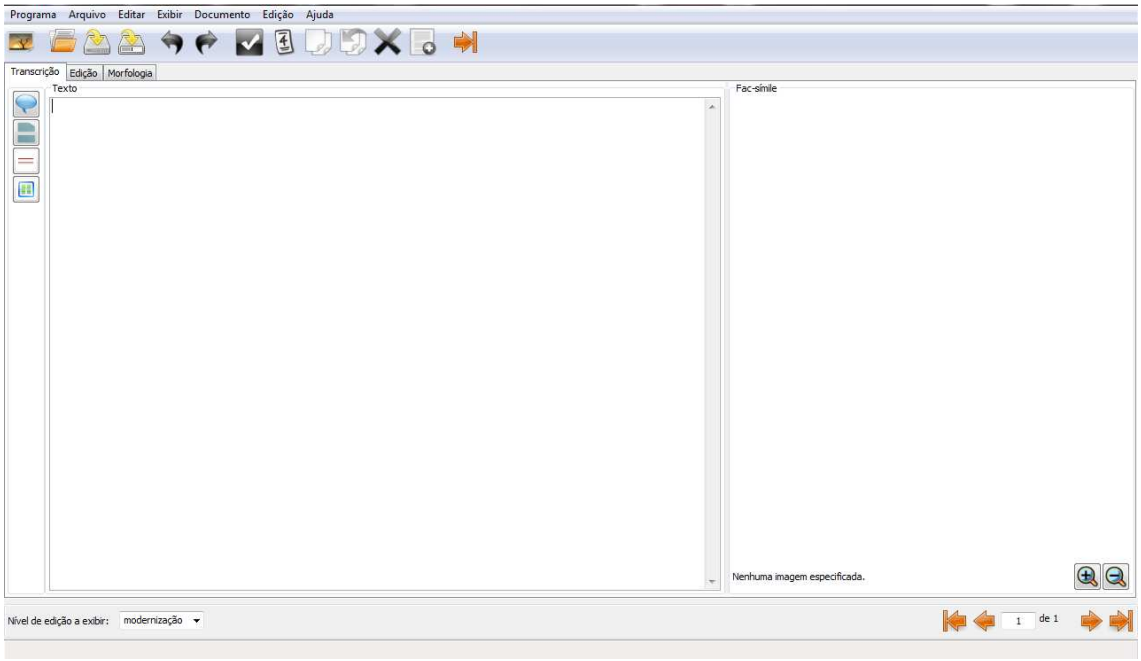

Figura 2: Ferramenta E-dictor.

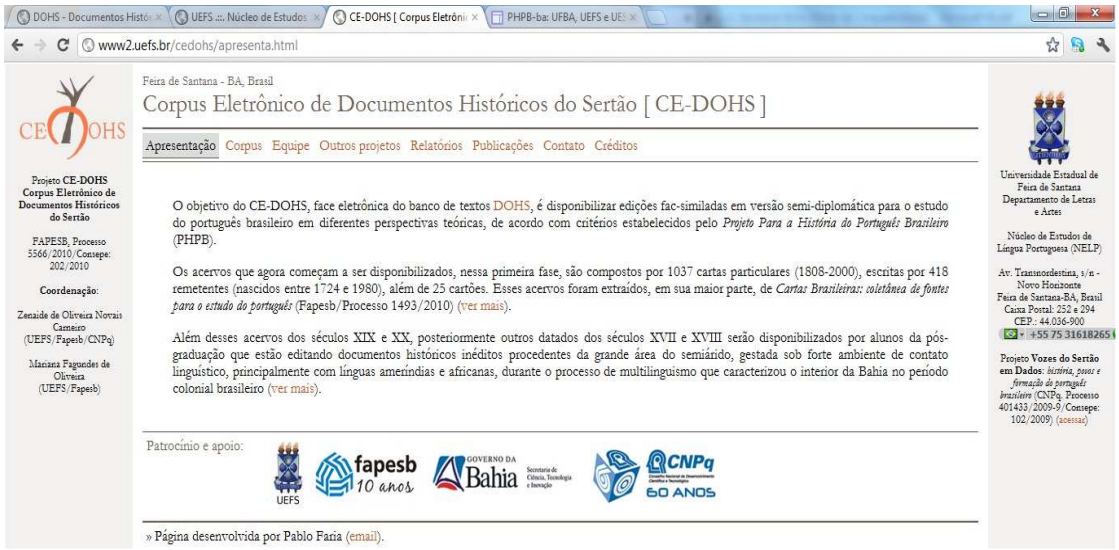

Figura 3: Site do projeto CE-DOHS

CONSIDERAÇÕES FINAIS

A experiência adquirida durante a bolsa Fapesb/Ação Referência 2010 e agora em nível mais aprofundado já com domínio de questões operacionais foi bastante significativa. Mesmo pertencendo ao curso de Engenharia de Computação, foi possível estabelecer relação com a área de Letras, especificamente, a área da Linguística Histórica, através do estudo e a aplicação do E-dictor, o uso de novas ferramentas computacionais e a manutenção do site do CE-DOHS (Figura 3). Essa parceria otimiza a proposta de

construção desse novo tipo de banco de dados, que tem se mostrado tendência mundial, a construção de *corpus* eletrônico.

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.

GALVES, Charlotte (http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/index.html) relatórios anuais. IDE, Nancy e Laurent Romary, 2003: Outline of the international standard linguistic annotation framework. Proceedings of ACL'03 Workshop on LOBO, Tânia / RIBEIRO, Ilza Ribeiro / CARNEIRO, Zenaide / ALMEIDA, Norma (Orgs. 2006). Para a história do português brasileiro: novos dados, novas análises. Salvador: Editora da Universidade Federal da Bahia, vol. VI, 2 tomos.

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.

PAIXÃO DE SOUSA, M.C. (2005a). *Memórias do Texto*. Participação na mesa-redonda “Bibliotecas e bancos de dados digitais de literatura”, II Simpósio Nacional de Literatura e Informática. Universidade Federal de Santa Catarina (UFSC), Florianópolis, outubro de 2005.

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.

TYCHO BRAHE. <http://www.tycho.iel.unicamp.br/~tycho/corpus/index.html>

W3C (1997). “Extensible Markup Language”. http://www.w3.org/XML