EDIÇÃO XML DE ANÚNCIOS DO JORNAL FOLHA DO NORTE (SÉCULOS XX E XXI): CONTRIBUIÇÕES À RECONSTRUÇÃO DA HISTÓRIA DO PORTUGUÊS BRASILEIRO

Matheus Santos Oliveira[1]; Mariana Fagundes de Oliveira Lacerda[2]

1. Bolsista PEVIC, Graduando em Licenciatura em Letras com Língua Espanhola, Universidade Estadual de Feira de Santana, e-mail: matheusuefs@live.com

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

PALAVRAS-CHAVE: Edição eletrônica, anúncios, português brasileiro.

INTRODUÇÃO

A busca de possíveis explicações para a origem do Português Brasileiro (doravante PB) e de como ele se distanciou do Português Europeu (PE) tem sido o assunto principal sobre o qual se debruça um sem número de linguistas históricos nos últimos cem anos, segundo Dante Lucchesi (2001).

Há, pelo menos, três grandes teorias para as origens do PB. Daí a importância de se construírem grandes bancos de dados como o primeiro passo para análises linguísticas e consequentes tentativas de explicação para a história dessa vertente da língua portuguesa na América. Isso só será alcançado graças ao trabalho conjunto de linguistas históricos (o que está sendo feito no âmbito do PHPB – Projeto para a História do Português Brasileiro –, do qual o CE-DOHS é parceiro), do que resultará uma melhor interpretação da realidade sócio-histórica da língua portuguesa do Brasil.

Este trabalho é, portanto, um produto do projeto CE-DOHS – *Corpus* Eletrônico de Documentos Históricos do Sertão, do Núcleo de Estudos de Língua Portuguesa (NELP), do Departamento de Letras e Artes (DLA) da Universidade Estadual de Feira de Santana (UEFS). O CE-DOHS (versão eletrônica do *corpus* do projeto DOHS – Documentos Históricos do Sertão) tem, por objetivo, implementar o acervo digital de um grande *corpus* em formato de texto computacionalmente manipulável, em linguagem XML, que permita recuperar informações gráficas de cunho filológico dos documentos originais, além disso gerar bases anotadas (morfológica e sintática) para análise linguística. Ou seja, um *corpus* que permita o uso de mecanismos que gerem versões diversas, acessíveis em processamentos de buscas automáticas. Tem, pois, nesse sentido, um diálogo bastante interessante entre a

Linguística e a Engenharia da Computação (ou, ainda, entre a Linguística de *corpus* e a Linguística Computacional).

METODOLOGIA

O material faz parte do banco de dados DOHS, do Projeto Vozes, conforme fora salientado.

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, conforme supracitado, 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.

O *corpus* editado faz parte do acervo *Anúncios do Jornal Folha do Norte (século XX-XXI)*. Trata-se de uma coletânea de 212 anúncios publicados entre 1910 e 2006.

Em termos práticos, esses anúncios, que já faziam parte do acervo do DOHS, já editados em versões semi-diplomáticas (com intervenção pequena do editor), receberam, agora, versões eletrônicas, utilizando o programa E-dictor. Esta ferramenta computacional, desenvolvida por Paixão de Souza, Kepler e Faria (2009), é especialmente voltada ao trabalho filológico e às análises linguísticas automáticas. A ferramenta combina um editor de XML e um etiquetador morfossintático, e permite a geração automática de versões correspondentes a edições diplomáticas, semi-diplomáticas e modernizadas (em HTML), e de versões com anotação morfossintática (em texto simples e XML).

RESULTADOS

As figuras a seguir mostram o passo a passo da edição XML desenvolvida: a edição fac-similar e semi-diplomática (forma como foi editada no projeto DOHS) – figura 1; edição eletrônica, utilizando a ferramenta E-dictor – figura 2; forma como os textos serão

disponibilizados no site do CE-DOHS – figura 3; *home* do site do CE-DOHS, onde é possível visualizar as diversas versões dos textos editados – figura 4.

Novo Cinema
em Feira

A direção do Cinema qu breve se inaugurará na Rua Castro Alves, atendendo a inúmeras 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, Nº 2530 p.1)

Figura 1: edição fac-similar e semi-diplomática

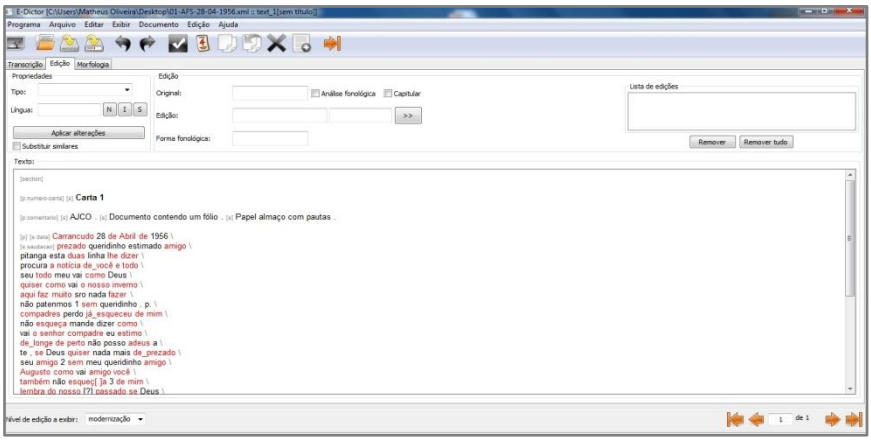
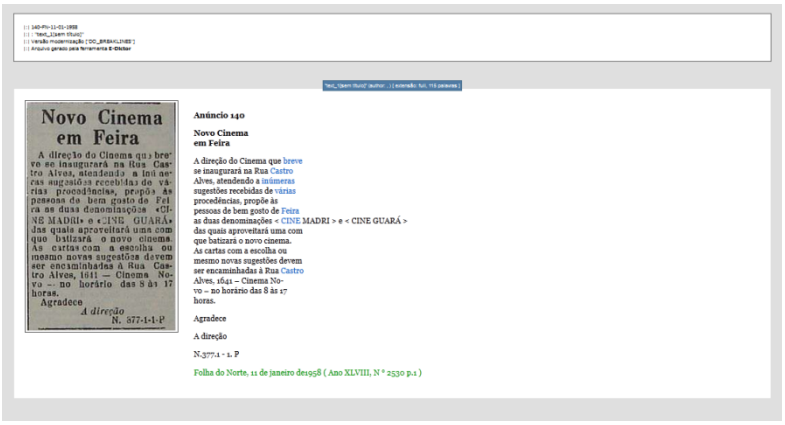


Figura 2: edição, usando o E-dictor



Novo Cinema
em Feira

A direção do Cinema qu breve se inaugurará na Rua Castro Alves, atendendo a inúmeras 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 breve
se inaugurará na Rua Castro
Alves, atendendo a inúmeras
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 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 3: edição XML pronta

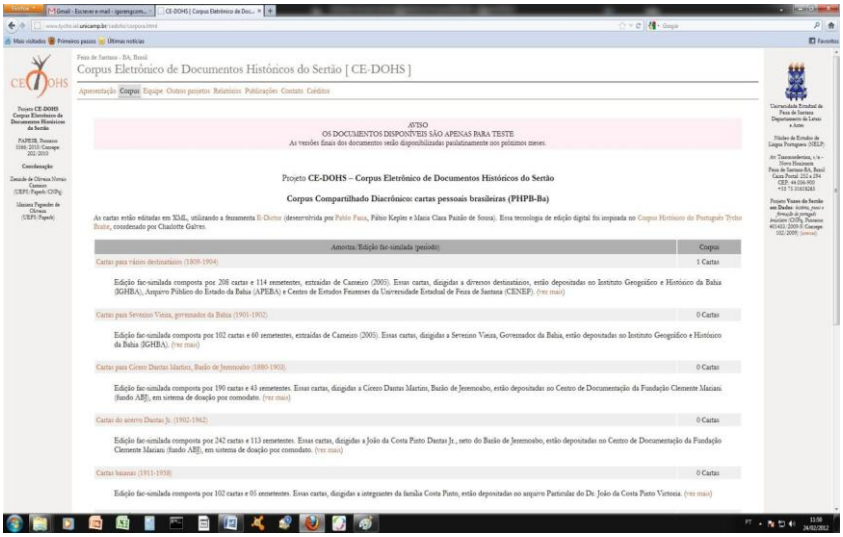


Figura 4: site do CE-DOHS

CONCLUSÕES

O trabalho ora apresentado teve, como dito, por objetivo, constituir um *corpus* linguístico sobre o qual um linguista pode debruçar-se com bastante facilidade para fazer estudos diacrônicos ou sincrônicos sobre a variação e a mudança linguística no PB. Dessa forma, contribuiu-se, aqui, com a primeira agenda de trabalho do PHPB: a construção de bancos de dados, atividade para a qual chamava a atenção a importante linguista histórica e visionária Rosa Virgínia Mattos e Silva (2008).

REFERÊNCIAS

CARNEIRO, Zenaide de Oliveira Novais; OLIVEIRA, Mariana Fagundes (Org.) *Publica-se em Feira de Santana*: dos anúncios e das cartas de leitores e redatores em *O Progresso* e no *Folha do Norte* (1901-2006). Feira de Santana: Editora UEFS , 2012.

LOBO, Tânia. Arquivos, acervos e a reconstrução histórica do português brasileiro. In: OLIVEIRA, Klebson; CUNHA E SOUZA, Hirão F.; SOLEDADE, Juliana (Org.). Do português arcaico ao português brasileiro: outras histórias. Salvador: EDUFBA, 2009.

MATTOS E SILVA, Rosa Virgínia. *Caminhos da linguística histórica: ouvir o inaudível*. São Paulo: Parábola editorial, 2008.

PAIXÃO DE SOUZA M. C. "Mémoria do Texto Revista Texto Digital Universidade Federal de Santa Catarina: 2006.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. *E-dictor: Novas perspectivas na codificação e edição de corpora de textos históricos*. In: VIII Encontro de Linguística de Corpus, 2009, Rio de Janeiro. Resumos, 2009. (a sair em: 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, 2009. p. 69-105.

PROJETO CORPUS ELETRÔNICO DE DOCUMENTOS HISTÓRICOS DO SERTÃO (disponível em www.uefs.br/cedohs), 2011.