EDIÇÃO EM LINGUAGEM *XML* DO ACERVO PARTICULAR DE DR. JOÃO DA COSTA PINTO VICTORIA: 2ª FASE

Priscila Starline Estrela Tuy Batista[1]; Zenaide de Oliveira Novais Carneiro[2]

1. Bolsista PROBIC, Graduada em Licenciatura em Letras Vernáculas, Universidade Estadual de Feira de Santana, e-mail: priscilatuy@gmail.com
2. Orientadora, 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

Nesta pesquisa, usamos a base de dados do Projeto *Vozes do Sertão em Dados: história, povos e formação do português brasileiro* (CNPq. Processo 401433/2009-9/Consepe: 102/2009) (www.uefs.br/nelp), coordenado por Zenaide de Oliveira Novais Carneiro, banco DOHS (www.uefs.br/dohs), especificamente em um subconjunto de textos da 2ª metade do século XX, o APJCPV – Acervo Particular Dr. João da Costa Pinto Victorio e a convertemos em formato eletrônico no âmbito 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). Foi desenvolvido o trabalho de edição em linguagem XML, utilizando a ferramenta computacional E-Dictor (PAIXÃO DE SOUZA; KEPLER; FARIA, 2009). Essa ferramenta amplia o alcance da linguagem XML e consegue unir a edição do texto e a correção em um mesmo ambiente, além de mapear as intervenções realizadas. A linguagem XML possibilita a codificação completa dos textos, tanto no cabeçalho, com fins de informações, catálogo e busca, quanto na estrutura dos textos, como paragrafação e paginação.

METODOLOGIA

Nossa metodologia baseia-se naquela 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/, sítio eletrônico no qual estão definidos os modelos e as ferramentas que estão subsidiando o projeto CE-DOHS, um *corpus* voltado a um banco de dados inédito de textos do sertão baiano, através do Termo Aditivo de Transferência de Tecnologia do *Corpus Histórico do Português Tycho Brahe* (www.tycho.iel.unicamp.br), sediado no Instituto de Estudos da Linguagem (IEL) da Universidade Estadual de Campinas/UNICAMP, coordenado por Charlotte Marie Chambelland Galves.

RESULTADOS

Como resultado do trabalho que foi desenvolvido, temos a edição do APJCPV em sua totalidade. As 102 cartas na versão Semi-Diplomática fac-similada, agora encontram-se em linguagem *XML* (Figura 1) e estão disponíveis no site do projeto CE-DOHS (www.uefs.br/cedoh), o qual, em uma primeira versão, encontra-se em edição Semi-Diplomática Fac-Símilada (Figura 2). Esse é um trabalho minucioso, exigindo que o editor esteja sempre atento. Ao término de cada edição, antes de ser destinado à alimentação do banco de dados, o documento foi revisado pela orientadora, para garantir que não haja erros na edição. O XML disponibiliza as versões Diplomática e Semi-Diplomática do documento, além da versão técnica, da ficha com metadados e do léxico de edições.

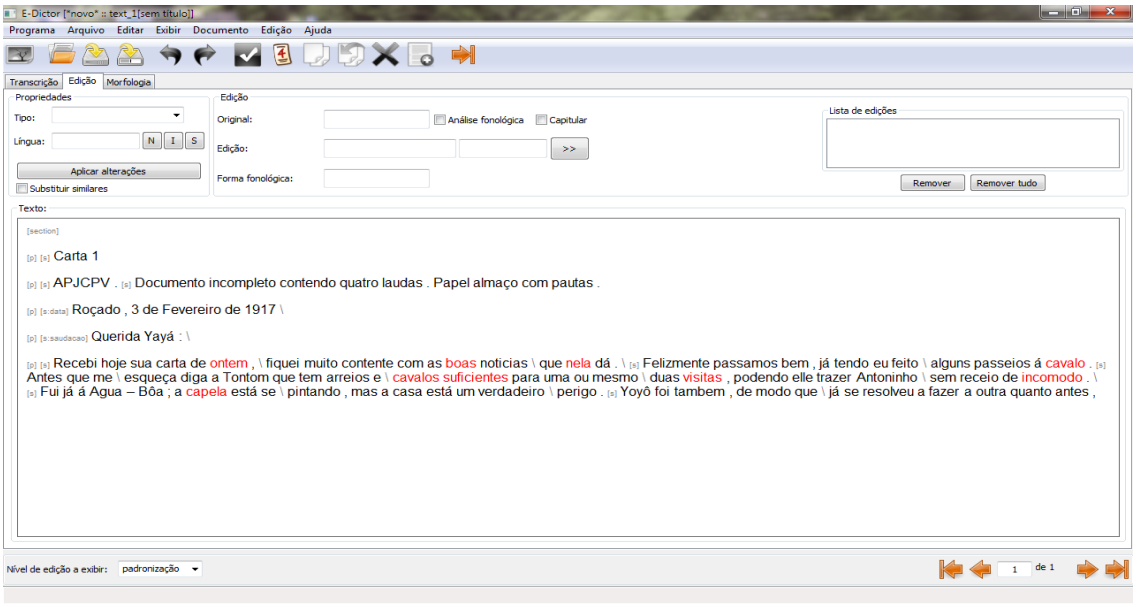

Figura 1: Modelo de edição em XML utilizando o E-Dictor.

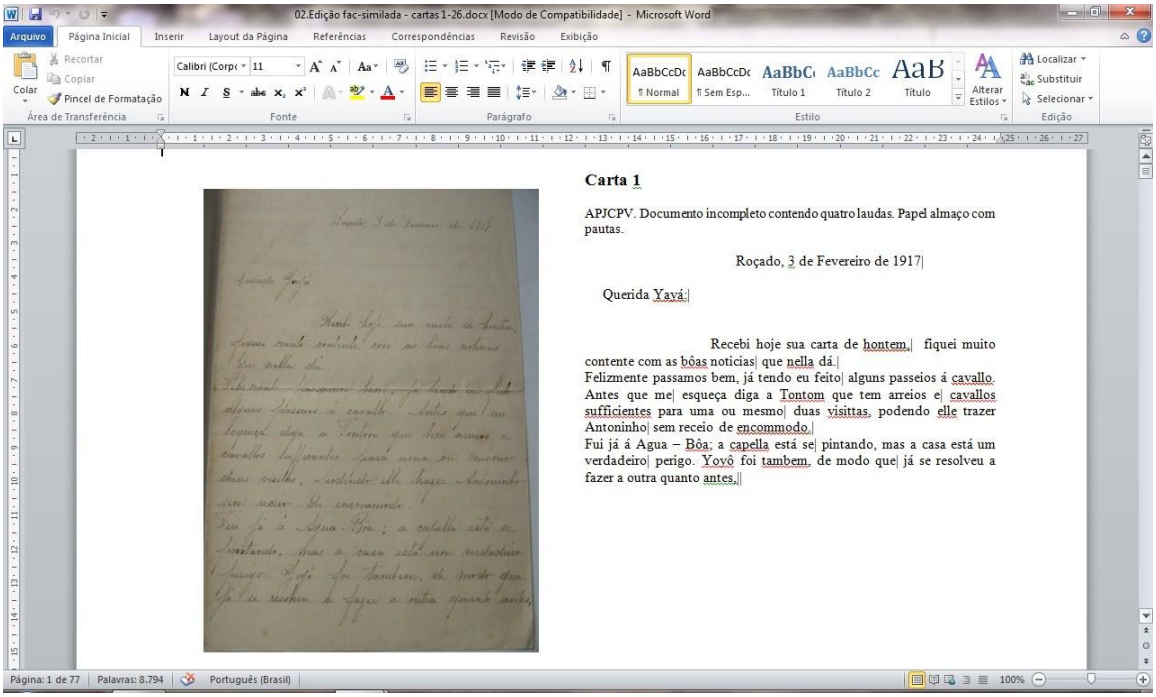

Figura 2: Modelo da edição Semi-Diplomática fac-similada.

Em paralelo ao processo de edição do APJCPV, finalizamos a revisão para publicação do Acervo *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)*. Esse acervo, também resultado de bolsistas anteriores, compõe o *corpus* do banco de dados DOHS, futuramente será editado em linguagem XML e disponibilizado no site do CE-DOHS. Houve, ainda, a colaboração na revisão do acervo ***Correspondências amigas: o acervo de Valente, Bahia (1980-1993)***, Cartas brasileiras (1809-2000): coletânea de fontes para o estudo do português / Zenaide de Novais Carneiro (Org.).

Também como resultado do trabalho e dos estudos, realizamos a apresentação de painel intitulado "Edição fac-similada de cartas do Arquivo Particular de Dr. João da Costa Pinto Victória: versão XML", no Castilho - *II Congresso Internacional de Linguística Histórica*, ocorrido de 07 à 10 de fevereiro de 2012, na Universidade de São Paulo (USP). Além disso, o trabalho foi apresentado na modalidade de

comunicação, no *I Congresso Internacional de Estudos Filológicos*, realizado de 29 de julho a 02 de agosto de 2012, na Universidade Federal da Bahia (UFBA).

Temos também a apresentação de pôster intitulado *"brasilian letters: edition in XML format"* 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.

CONSIDERAÇÕES FINAIS

O objetivo desta edição, feita em linguagem XML, foi disponibilizar esse material para pesquisadores interessados em estudar a história do Português Brasileiro (PB), e compor o banco eletrônico de dados, o CE-DOHS – Corpus Eletrônico de Documentos Históricos do Sertão (http://www2.uefs.br/cedohs/), a partir do corpus do DOHS – Documentos Históricos do Sertão. De forma mais geral, o nosso trabalho colabora com a composição de *corpora* anotados de longo alcance de busca em parceria com o Projeto Corpus Histórico do Português Tycho Brahe, além de outros parceiros.

REFERÊNCIAS

BARBOSA, Afrânio G. et al. ***Normas de Transcrição de Documentos Manuscritos e Impressos.*** In. MATTOS E SILVA, Rosa Virgínia (org.) (2001). *Para a história do português brasileiro:* primeiros estudos. São Paulo. Humanitas/FFCHL/USP. FAPESP, Vol. II, tomo II.

BARBOSA, Afrânio G. ***Linguística de corpus e sociolinguística histórica: o lugar dos grupos de fatores externos.*** In: XV Congresso Internacional de La Asociación de Linguística y Filología de América Latina/ALFAL. Montevideo, 2008.

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.

DUARTE, Maria Eugênia Lamoglia & Callou, Dinah (org.). Para a História do Português Brasileiro – **Notícias de corpora e outros estudos – Vol. IV**. Faculdade de Letras da UFRJ/FAPERJ, Rio de Janeiro, 2002.

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)

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

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. (2007b). **Linguística de Corpus e História da Língua Portuguesa: Propostas, Resultados e Desafios**, Coordenação de Mesa Redonda no V Congresso Internacional da Associação Brasileira de Lingüística – ABRALIN. Belo Horizonte, 2 de março de 2007.

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.