UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXIV SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2021

Banco CE-DOHS: elaboração dos metadados de amostras de fala do projeto Norma Linguística Urbana Culta do Brasil (NURC-Salvador) – século XX

Aracele Santos de Lima[1] e Mariana Fagundes de Oliveira Lacerda[2];

1. Aracele Santos de Lima, Graduanda em Letras com Língua Inglesa, Universidade Estadual de Feira de Santana, e-mail: aracele.santosdelima@gmail.com
2. Mariana Fagundes de Oliveira Lacerda, Departamento de Letras e Artes, Universidade Estadual de Feira de Santana, e-mail: marianafagundes@uefs.br

PALAVRAS-CHAVE: Linguística de Corpus. Português Brasileiro. Amostras de Fala.

INTRODUÇÃO

O projeto CE-DOHS: Corpus Eletrônico de Documentos Históricos do Sertão, que conta com ótimos recursos tecnológicos, no universo das Humanidades Digitais, disponibiliza edições semidiplomáticas, em PDF, e, por meio do estabelecimento de redes com projetos que desenvolvem a Linguística de Corpus – como o projeto Corpus Histórico do Português Tycho Brahe (http://www.tycho.iel.unicamp.br/~tycho/corpus/), coordenado por Charlotte Galves (UNICAMP) –, também edições em linguagem XML, usando o eDictor, programa computacional desenvolvido por Kepler, Paixão de Souza e Faria (2007), para facilitar a edição eletrônica de textos antigos, que, anotada sintaticamente, permite a busca automática de dados no estudo linguístico. Como se vê, "Do feliz congraçamento entre as mais recentes tecnologias e a antiga Filologia, surgiu um novo universo de possibilidades para a preservação, disponibilização e análise de textos antigos, universo em que é possível oferecer ao leitor mais de uma edição do mesmo texto, permitindo que tenha ao seu dispor o texto editado, em diferentes versões, e o seu original." (GONÇALVES; BANZA, 2013, p. 4) Segundo Shepherd et al. (2012, p. 11), "A ideia de coligir coleções de textos naturais com o objetivo de os submeter à análise linguística remonta ao trabalho dos estruturalistas norte-americanos da década de 1950, tais como Harris (1951) e Fries (1952). Com o Brown Corpus (Francis e Kucera, 1954), surgiria o primeiro corpus eletrônico compilado para este fim. Embora até hoje este corpus seja largamente utilizado, na altura praticamente não existiam textos escritos em formato digital, os computadores eram máquinas enormes e caras, que ocupavam salas inteiras, e os programas informáticos demoravam horas e até dias a correr."

Neste trabalho, realizou-se a elaboração dos metadados de 7 inquéritos do projeto Norma Linguística Urbana Culta do Brasil (NURC-Salvador) – século XX. Os metadados reúnem informações diversas, sobre dados do documento: autor, conteúdo, destinatário, local, data do documento, referência, fonte, gênero do autor, gênero do documento; e dados do processamento: revisão final da edição XML, primeira revisão da edição XML, edição semidiplomática, revisão da edição semidiplomática, edição XML, número de palavras.

MATERIAL E MÉTODOS OU METODOLOGIA (ou equivalente)

Os 7 inquéritos trabalhado aqui são do projeto Norma Linguística Urbana Culta do Brasil (NURC-Salvador) – século XX. O Projeto NURC foi criado em âmbito nacional, em 1969, com o propósito de reunir um banco de dados da fala culta de várias capitais do Brasil. O *corpus* de Salvador foi constituído entre 1973 e 1983 e reúne 360 inquéritos, com registro de 461 informantes; 58 inquéritos em elocuções em situação formal, 201 em diálogos entre informante e documentador e 101 em diálogos entre dois informantes. A fase de constituição do *corpus* de Salvador foi coordenada por Nelson Rossi, também coordenador nacional do Projeto NURC/Brasil em vários períodos.
A elaboração dos metadados será feita com base nos pressupostos da Linguística de Corpus e nos critérios do Projeto CE-DOHS e do projeto Tycho Brahe.

RESULTADOS E/OU DISCUSSÃO (ou Análise e discussão dos resultados)

Apresentam-se aqui os resultados obtidos do plano vinculado ao projeto Plataforma de Corpus Eletrônico de Documentos Históricos do Sertão (XVI-XX): contribuições ao entendimento da relação entre contato linguístico e mudança paramétrica na formação do português brasileiro, que conta com ótimos recursos tecnológicos, no universo das Humanidades Digitais, disponibilizando edições semidiplomáticas, em PDF, e, por meio do estabelecimento de redes com projetos que desenvolvem a Linguística de Corpus – como o projeto Corpus Histórico do Português Tycho Brahe (http://www.tycho.iel.unicamp.br/~tycho/corpus/), coordenado por Charlotte Galves (UNICAMP) –, também edições em linguagem XML, usando o eDictor, programa computacional desenvolvido por Kepler, Paixão de Souza e Faria (2007), para facilitar a edição eletrônica de textos antigos, que, anotada sintaticamente, permite a busca automática de dados no estudo linguístico. Como se vê, "Do feliz congraçamento entre as mais recentes tecnologias e a antiga Filologia, surgiu um novo universo de possibilidades para a preservação, disponibilização e análise de textos antigos, universo em que é possível oferecer ao leitor mais de uma edição do mesmo texto, permitindo que tenha ao seu dispor o texto editado, em diferentes versões, e o seu original." (GONÇALVES; BANZA, 2013, p. 4)
Neste plano, realizou-se a elaboração dos metadados de 7 inquéritos do projeto Norma Linguística Urbana Culta do Brasil (NURC-Salvador) – século XX. Os metadados reúnem informações diversas, sobre dados do documento: informante, conteúdo, local, data, referência, fonte, gênero do informante, gênero do documento; e dados do processamento: edição XML, primeira revisão da edição XML, revisão final da edição XML, número de palavras.
A edição XML dos referidos inquéritos está sendo realizada por outra bolsista do CE-DOHS, Taine do Rosário, com quem trabalhei em parceria durante o desenvolvimento da presente pesquisa.
Os resultados serão, em breve, divulgados na página do projeto CE-DOHS, tanto a edição, feita por Taine do Rosário, quanto os metadados dos 7 inquéritos, preenchidos por mim: http://www5.uefs.br/cedohs/

Figura 1: página inicial do projeto CE-DOHS.

Conferir, a seguir, a título de ilustração, um *print* dos metadados do documento e um *print* dos metadados do processamento dos documentos:

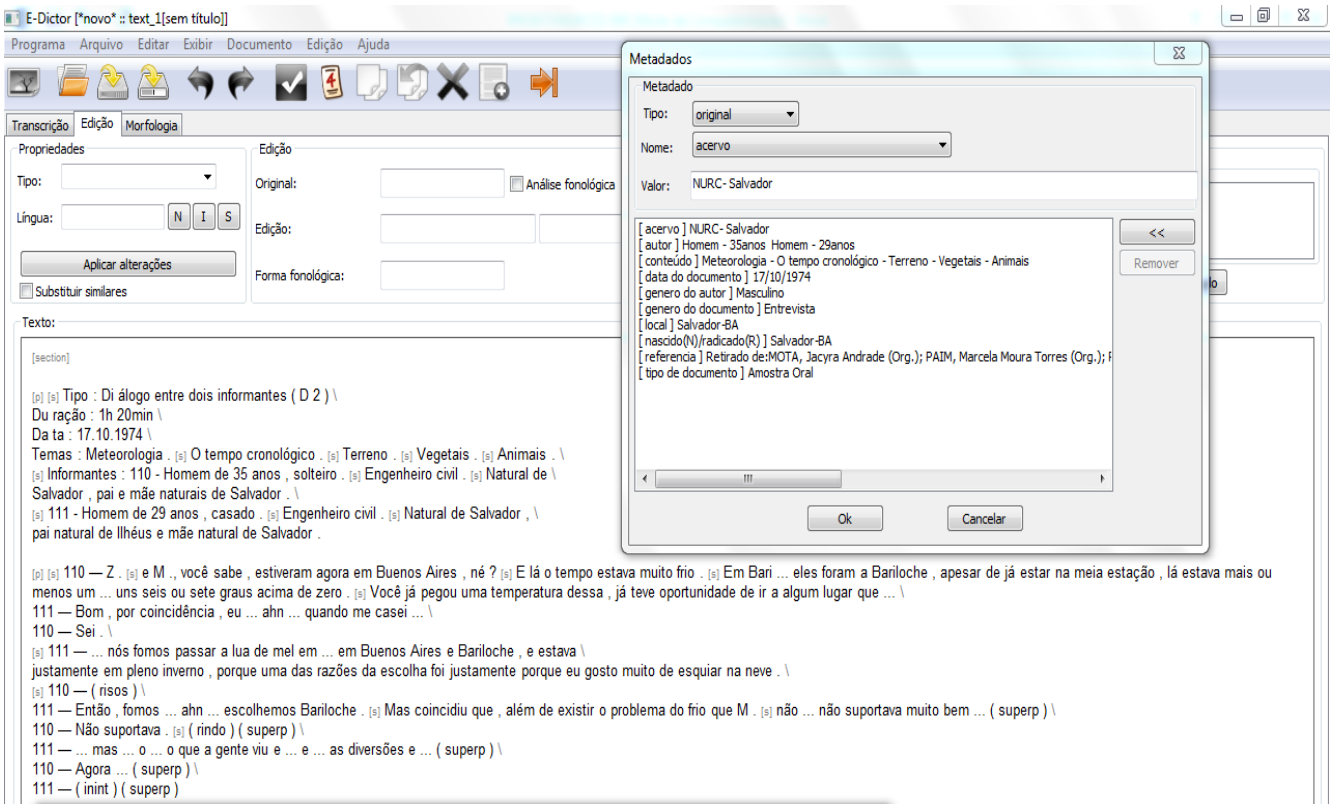

Figura 2: Metadados dos documentos

CONSIDERAÇÕES FINAIS (ou Conclusão)

A edição XML de amostras orais constitui uma das agendas de trabalho do projeto CE-DOHS, e os resultados do presente plano colaboram com essa agenda. Para a história do PB, os pesquisadores se debruçam sobre um conjunto de dados, tanto escritos como de fala. Para maiores informações sobre o conjunto de amostras de fala com que o CE-DOHS trabalha, acessar o site: http://www5.uefs.br/cedohs/amostrasorais/index.html.

REFERÊNCIAS

BACELAR DO NASCIMENTO, M. F. O lugar do corpus na investigação lingüística. Disponível em: http://www.clul.ul.pt/equipa/berlim-2000-nascimento.pdf. Acesso em: 20 abr. 2004.

CASTILHO, A. T. de (Org.). Para a história do português brasileiro: primeiras idéias. São Paulo: Humanitas Publicações / FFLCH/USP, 1998, v.1. GONÇALVES, M. F.; BANZA, A. P. (org.). Património textual e humanidades digitais: da antiga à nova Filologia. Évora: CIDEHUS, 2013.

MATTOS E SILVA, R. V. 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, 2002, v. 2, p. 443-464.

MATTOS E SILVA, R. V. Ensaios para uma sócio-história do português brasileiro. São Paulo: Parábola, 2004.

MOTA, J. ; ROLLEMBERG, V. A linguagem falada culta na cidade de Salvador: materiais para seu estudo. Salvador : Instituto de Letras da UFBA, 1994.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. E-dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. In: Anais do VIII Encontro de Linguística de Corpus, realizado na UERJ, 13 a 14 de novembro de 2009. Rio de Janeiro, 2009. p. 69-105.

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

SHEPHERD, T. M. G.; SARDINHA, T. B.; PINTO, M. V. (Org.) Caminhos da Linguística de Corpus. Campinas: Mercado de Letras, 2012.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. E-dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. In: Anais do VIII Encontro de Linguística de Corpus, realizado na UERJ, 13 a 14 de novembro de 2009. Rio de Janeiro, 2009. p. 69-105.