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

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

PARA A CONSTITUIÇÃO DE BANCO DE DADOS ELETRÔNICOS DE PORTUGUÊS BRASILEIRO FALADO CULTO: Edição com uso do eDictor de inquéritos do NURC-Salvador

Taine do Rosário[1]; Mariana Fagundes de Oliveira Lacerda[2]
1. Bolsista PIBIC/FAPESB, Graduanda em Letras com Inglês, Universidade Estadual de Feira de Santana, e-mail: taine.rosario99@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: Amostras orais; edição XML; português brasileiro

INTRODUÇÃO

O projeto Corpus Eletrônico de Documentos Históricos do Sertão (CE-DOHS), desde a sua criação, no ano de 2012, como um dos projetos que integram o Núcleo de Estudos de Língua Portuguesa (NELP), do Departamento de Letras e Artes (DLA) da Universidade Estadual de Feira de Santana (UEFS), desenvolve trabalhos voltados, no âmbito da Linguística de *Corpus*, para a elaboração e disponibilização de uma base documental significativa e representativa de diferentes períodos para o estudo do Português Brasileiro (PB) e, para isso, busca por fontes variadas e abundantes, uma vez que, como afirma Sardinha (2000), um dos elementos fundamentais para elaboração de um *corpus* é a sua extensão. Além dos aspectos relacionados ao tamanho, Sardinha (2000) também salienta outras características necessárias a um *corpus* ao apresentar a definição de Sanchez (1995).

> ‘Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise’ (Sanchez,1995, p. 8-9 apud Sardinha, 2000, p. 16)

Em consonância com a definição proposta, o CE-DOHS disponibiliza, gratuitamente, uma plataforma na qual é possível acessar uma vasta base documental, composta por dois conjuntos, nos quais se encontram textos escritos (cartas, anúncios, atas etc) e amostras de fala. Para organizar o processo de desenvolvimento da plataforma, foram adotadas agendas de trabalho que aglutinam tarefas específicas que são desenvolvidas por equipes estratégicas dentro do grupo de pesquisa. A edição modernizada, com o uso do eDictor, está inserida na segunda agenda e tem sido desenvolvida com êxito, especialmente por bolsistas de iniciação científica, possibilitando a inserção dos documentos representativos do falar baiano, sejam eles do sertão ou da capital, no universo das Humanidades Digitais, a partir da Nova Filologia.

Dessa forma, mostra-se pertinente o trabalho aqui apresentado, uma vez que a disponibilização da edição modernizada do projeto NURC-Salvador contribui diretamente para novas perspectivas no que diz respeito ao tratamento digital de materiais representativos da língua falada, pois, como é possível perceber, a maior parte dos documentos já editados contemplam a língua escrita.

MATERIAL E MÉTODOS OU METODOLOGIA (ou equivalente)

O *corpus* utilizado no presente plano de trabalho foi composto por sete inquéritos do projeto Norma Linguística Urbana Culta do Brasil (NURC-Salvador) – século XX. O Projeto NURC teve início, em âmbito nacional, no ano de 1969, sob a coordenação de Nelson Rossi, também coordenador da equipe de Salvador, tendo como principal objetivo o mapeamento do falar culto brasileiro a partir da coleta de amostras de fala de cinco capitais e, posteriormente, a organização das mesmas em um banco de dados. As amostras orais com as quais o trabalho aqui descrito foi desenvolvido foram cedidas ao CE-DOHS pela equipe do NURC-Salvador.

Na cidade de Salvador, as amostras foram coletadas entre 1973 e 1983, uma década de trabalho que teve como resultado 360 inquéritos, com registro de 461 informantes. As modalidades dos inquéritos foram três, sendo 58 inquéritos em elocuções em situação formal, 201 em diálogos entre informante e documentador e 101 em diálogos entre dois informantes. Como mencionado anteriormente, o trabalho desenvolvido contemplou apenas sete inquéritos já transcritos segundos os critérios do projeto NURC.

Tabela 1 – Descrição do Inquéritos

Nº do inquérito	Informantes	Data	Tipo	Duração	Tema
052	52	15.09.1974	EF	45 min	Ecologia – Atuação do homem na Biosfera.
095	110 / 111	17.10.1974	D2	lh 20min	O tempo cronológico. Terreno. Vegetais. Animais.
098	115 / 116	22.10.1974	D2	lh 20min	Transportes. Viagens.
150	186 / 187	16.05.1975	D2	lh 20min	A cidade. O comércio. Transportes e viagens. Os meios de comunicação e difusão. Cinema. TV. Rádio. Teatro.
204	266 / 267	08.04.1976	D2	lh 20min	Meteorologia. Terreno. Vegetais. Animais.
298	375 / 376	21.09.1977	D2	lh 20min	A cidade. Os transportes.
357	453 / 454	05.09.1978	D2	lh 20min	A família. O ciclo da vida. Vida social. Diversões.

Fonte: elaboração própria

A edição modernizada trata-se do primeiro estágio do tratamento filológico digital ao qual os textos são submetidos com o objetivo de que, posteriormente, possam ser usados na realização de buscas automáticas. No CE-DOHS, o referido processo é desenvolvido com o uso da ferramenta computacional eDictor (PAIXÃO DE SOUSA, KEPLER E FARIA, 2007). A ferramenta, no âmbito das Humanidades Digitais, executa

o que Gonçalves e Banza (2013, p. 4) chamam de “feliz congraçamento”, resultado da intersecção das inovações tecnológicas ao rigor científico já conhecido da Filologia.

O eDictor combina um editor filológico digital a um anotador morfossintático. No processo de etiquetagem, denominação que é dada para as alterações realizadas no texto ao longo da edição modernizada, o eDictor utiliza a linguagem XML, que proporciona o mapeamento e a decodificação dos níveis de edição, além da geração de diversas versões de um mesmo texto, o que garante a preservação do texto original. Ainda em função do uso da XML, é possível definir regras específicas para cada tipo de texto, funcionalidade que é essencial para bancos de dados como o CE-DOHS, que dispõe de vários tipos de texto, cada um com suas especificidades.

Figura 1 – Página principal do site do eDictor

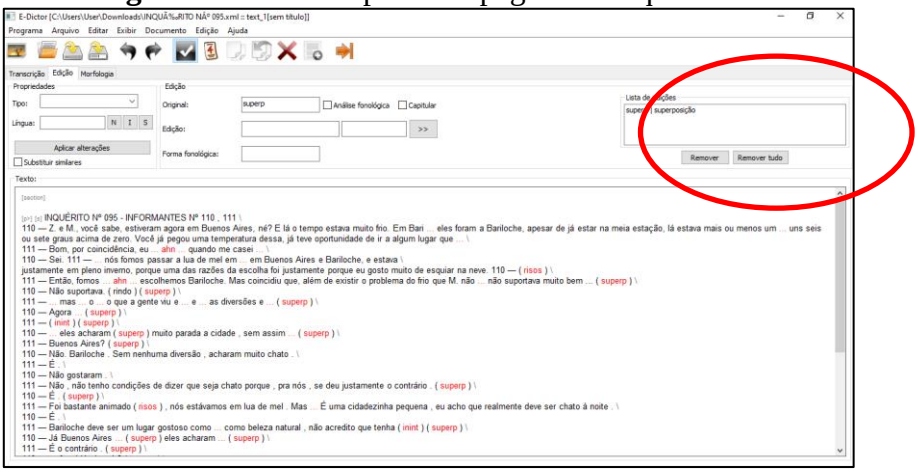
Fonte: https://edictor.net/

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

No modelo original, o arquivo que acompanha o eDictor dispõe de onze etiquetas, organizadas hierarquicamente. Para a edição das amostras do NURC-Salvador, foram adicionadas outras seis etiquetas, a saber: pausa, superposição, marcador conversacional, risos, hesitação e repetição. A partir da adição, foi possível desenvolver o processo de edição de maneira a preservar os aspectos essenciais para os estudos linguísticos que poderão ser desenvolvidos com o *corpus*, bem como prepará-lo corretamente para a etapa seguinte de anotação. Sendo assim, é necessário ter muita atenção ao executar adaptações e realizar escolhas, sempre buscando o equilíbrio que favorecerá ambos os campos de pesquisa.

Após o carregamento do arquivo de configuração, trata-se da transcrição do texto na aba de transcrição do eDictor, nesse caso, não foi necessário transcrever, o texto foi apenas copiado e colado do arquivo docx, para a aba de transcrição da ferramenta, onde a quebra de páginas foi realizada. Posteriormente, o XML do arquivo é gerado e então a edição é iniciada, onde são feitas as intervenções necessárias através do processo de etiquetagem. Como mencionado em momentos anteriores, o eDictor preserva o texto original, para isso, as palavras ou elementos que sofreram algum tipo de modificação ficam em vermelho, dessa forma é possível identificar que houve intervenção de ordem externa.

Figura 8 – XML da primeira página do inquérito 095



Fonte: elaboração própria

Como é possível observar na imagem, algumas palavras estão em vermelho, essa é a indicação visual de intervenção, mas, além dessa sinalização, o eDictor, no canto superior esquerdo da interface apresenta uma lista de edições, onde são mostradas todas as mudanças que foram aplicadas a determinada palavra. Em se tratando dos inquéritos do NURC-Salvador, a maioria das etiquetas originais não foram usadas, principalmente em função das mesmas terem sido desenvolvidas com enfoque no tratamento de textos escritos. Entretanto, como mencionado anteriormente, não houve dificuldades para a realização da edição das amostras, visto que o arquivo foi adaptado com a adição das etiquetas necessárias

CONSIDERAÇÕES FINAIS (ou Conclusão)

A partir da edição das amostras de fala do projeto NURC-Salvador foi possível obter resultados muito frutíferos. O caráter inovador do plano, uma vez que foi a primeira edição modernizada com o referido tipo de documento, trouxe contribuições significativas tanto para o banco de dados do CE-DOHS, com a sua expansão, quanto para o projeto NURC-Salvador que teve, ainda que parcialmente, sua inserção no universo das Humanidades Digitais, bem como para a ferramenta eDictor que, mais uma vez, foi utilizada para e aprimorada, possibilitando o uso posterior para trabalhos de natureza semelhante, inclusive com os outros inquéritos que compõe o acervo do NURC.

REFERÊNCIAS

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.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. eDictor. 2007 (disponível em: https://edictor.net/)

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

GONÇALVES, M. F.; BANZA, A. P. (org.). Património textual e humanidades digitais: da antiga à nova Filologia. Évora: CIDEHUS, 2013