UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto n°17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXVII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA – 2023

CORPUS PEPP: EDIÇÃO MODERNIZADA PARA BANCO ELETRÔNICO DE AMOSTRAS DE FALA

Mirian Galindo Marques[1]; Mariana Fagundes de Oliveira Lacerda[2]
1. Bolsista PIBIC/CNPq, Graduada em Letras Vernáculas, Universidade Estadual de Feira de Santana, e-mail: gmirian924@gmail.com
2. Orientadora, Departamento de Letras e Artes, Universidade Estadual de Feira de Santana, e-mail: marianafagundes@uefs.br

PALAVRAS-CHAVE: Amostras orais; português popular; Linguística de Corpus.

INTRODUÇÃO

A tradição dos estudos de Linguística Histórica é marcada pela natureza atomística das análises feitas nesse campo de estudo da língua. Esse caráter atomístico das análises dos fatos linguísticos, que, inicialmente, refletia concepções igualmente atomizadas (pré-saussurianas) do fenômeno linguístico, manteve-se na Linguística Histórica, mesmo quando essas concepções que o fundamentavam já estavam superadas, em boa medida devido à dificuldade de proceder a uma observação sistemática, e, da forma como possível, exaustiva, dos materiais disponíveis.

A constituição de Banco de Textos visa exatamente a romper com essa tendência nos estudos de história da língua, possibilitando, com a facilidade de um amplo acesso aos materiais, a aplicação das novas teorias que propugnam uma apreensão globalizante do objeto através de sua estrutura interna (linguística) e daquelas que, ainda mais globalizantes, propõem a apreensão dos fatos através da interação do sistema de relações linguísticas com as disposições e relações nas quais esse sistema se atualiza, as relações sociolinguísticas. A ampliação dos *corpora*, como afirma Bacelar do Nascimento (2004, p.1), "favorece essencialmente uma Linguística descritiva, fortemente apoiadapelas novas tecnologias, e permite tomar como ponto de partida da descrição, a análise de quantidade significativa de dados autênticos, à semelhança do que se faz noutros domínios científicos. O uso de *corpora* permite a realização de descrições linguísticas de base empírica e promove, com isso, a discussão de questões teóricas solidamente fundamentadas."

Hoje, contando com melhores recursos tecnológicos, no universo das Humanidades Digitais, projetos que desenvolvem a Linguística de *Corpus* vêm disponibilizando não somente edições semidiplomáticas, em PDF, ou transcrições de amostras de fala, mas também edições em linguagem XML desses materiais, usando o eDictor, programa

computacional desenvolvido por Paixão de Souza, Kepler e Faria (2009), 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)

O objetivo primordial deste trabalho, no universo das humanidades digitais, foi realizar a edição modernizada do *corpus* do Programa de Estudos sobre o Português Popular de Salvador (PEPP), da Universidade do Estado da Bahia (UNEB), coordenado pela professora doutora Norma da Silva Lopes, segundo a qual "O PEPP tem possibilitado a ampliação da pesquisa sobre os usos linguísticos em Salvador, e na Bahia, com pesquisas nos diversos *campi* da UNEB e em outras instituições, contribuindo para o conhecimento do português brasileiro." (2018, p. 29). A edição acha-se conlcuída, restando apenas ser feita sua revisão final, para, então, ser o material disponibilizado na página *online* do projeto Corpus Eletrônico de Documentos Históricos do Sertão (CE-DOHS) http://www.uefs.br/cedohs -, do Núcleo de Estudos de Língua Portuguesa (NELP) da Universidade Estadual de Feira de Santana (UEFS) - <https://nelp.uefs.br> -, do qual o PEPP é parceiro, apresentando também os dados extratextuais e toda informação pertinente.

O PEPP, constituído entre os anos de 1998 e 2000, surgiu para preencher uma necessidade de amostras da fala de pessoas com pouca ou média escolarização em Salvador. Conforme Lopes (2018, p. 24), "O PEPP se destinou a não só fazer registros da língua de uso real da fala de Salvador na época de sua constituição, mas também promover pesquisas linguísticas em diversos níveis no estado da Bahia, dando, assim, sua contribuição para o entendimento do português brasileiro."

MATERIAL E MÉTODOS

O *corpus* PEPP é formado por 48 gravações de entrevista de, aproximadamente, 40 minutos, transcritas segundo as normas definidas pelo Projeto NURC/Salvador. Os informantes estão distribuídos em quatro faixas etárias: de 15 a 24 anos, de 25 a 35 anos, de 45 a 55 anos e de 65 anos em diante. Quanto à escolaridade dos informantes, variou entre mínima, de 1 a 4 anos (pouca escolaridade), e máxima, de 11 anos de permanência na escola (média escolaridade). (LOPES, 2008)

A edição modernizada foi realizada com o uso do eDictor, desenvolvido por Kepler, Paixão de Sousa e Faria (2007). Essa ferramenta combina um editor de XML e um etiquetador morfossintático, e permite a geração automática de versões correspondentes a edições diplomáticas, semidiplomáticas e modernizadas (em html), e de versões com anotação morfossintática (em texto simples e xml). Trata-se de um feliz congraçamento entre as mais novas tecnologias e a antiga Filologia.

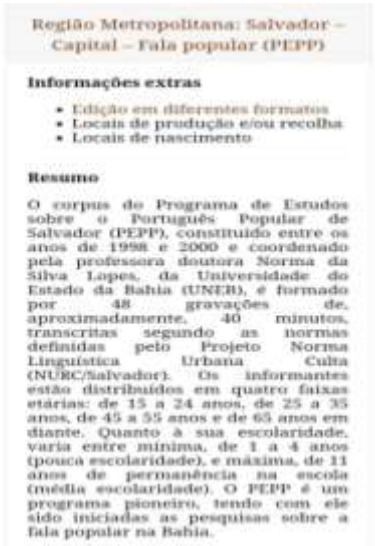
Figura 1: Logomarca da ferramenta eDictor.

Fonte: <https://humanidadesdigitais.org/edictor/>

RESULTADOS E/OU DISCUSSÃO

Realizou-se a edição modernizada, com uso do eDictor, do conjunto de amostras orais do PEPP, bem como o estudo de trabalhos nos campos da Linguística Histórica, História do Português Brasileiro, Linguística de Corpus e Humanidades Digitais. Aprender a manusear a ferramenta digital eDictor (PAIXÃO DE SOUSA, KEPLER E FARIA, 2007) tratou-se de uma rica experiência. Na próxima etapa de trabalho, será feita a revisão geral da edição, que será, então, disponibilizada no site CE-DOHS, ampliando o número de amostras orais do banco, preparadas, em linguagem XML, para a anotação sintática, por pesquisadores interessados nesse uso dos materiais.
As amostras orais do CE-DOHS ficam disponíveis no *menu* Coleções Documentais do site CE-DOHS. Atualmente, no site, encontra-se disponível apenas a edição em XML, sem modernização.

Figura 2: Página do *corpus* PEPP no site CE-DOHS


Região Metropolitana: Salvador – Capital – Fala popular (PEPP)

Informações extras

- Edição em diferentes formatos
- Locais de produção e/ou recolha
- Locais de nascimento

Resumo

O corpus do Programa de Estudos sobre o Português Popular de Salvador (PEPP), constituído entre os anos de 1998 e 2000 e coordenado pela professora doutora Norma da Silva Lopes, da Universidade do Estado da Bahia (UNEB), é formado por 48 gravações de, aproximadamente, 40 minutos, transcritas segundo as normas definidas pelo Projeto Norma Linguística Urbana Culta (NURC/Salvador). Os informantes estão distribuídos em quatro faixas etárias: de 15 a 24 anos, de 25 a 35 anos, de 45 a 55 anos e de 65 anos em diante. Quanto à sua escolaridade, varia entre mínima, de 1 a 4 anos (pouca escolaridade), e máxima, de 11 anos de permanência na escola (média escolaridade). O PEPP é um programa pioneiro, tendo com ele sido iniciadas as pesquisas sobre a fala popular na Bahia.

Fonte: < http://www5.uefs.br/cedohs/view/colecoes_documentais.html#A1823>

CONSIDERAÇÕES FINAIS

Observou-se que o *corpus* explorado possui uma amostra valiosa da comunidade de fala, composta pela população não universitária de Salvador, na sua época de constituição, anos finais do séc. XX, a qual é de salutar importância para o entendimento da norma vernacular do português brasileiro. A realização das edições modernizadas desses inquéritos requer não só a atenção em relação à padronização ortográfica comum a manuscritos e impressos, mas solicita do pesquisador o cuidado em preservar tanto a morfossintaxe como as características da gramática da fala, como reduplicações, partes de palavras, marcadores conversacionais, hesitações, entre outros, com fidelidade aos dados linguísticos e à representatividade de um *corpus* cujo objetivo é registrar a língua de uso real (conferir ROSÁRIO (2023)). Da mesma sorte, a experiência promoveu a aprendizagem teórica e prática no manuseio da ferramenta digital eDictor (PAIXÃO DE SOUSA, KEPLER E FARIA, 2007), já consolidada entre os pesquisadores da área e aplicada no tratamento de textos para a constituição de bancos de dados linguísticos eletrônicos, os quais disponibilizam aos consulentes diferentes versões de edição, ficando sempre disponível ao consulente a versão original.

REFERÊNCIAS

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

CARNEIRO, Z. O. N.; LACERDA, M. F. O. (Org). **CE-DOHS - Corpus Eletrônico de Documentos Históricos do Sertão (2012-2025)**. URL: http://www.uefs.br/cedohs. Acesso em: 14 fev 2023.

GONÇALVES, M. F.; BANZA, A. P. Fontes de metalinguísticas para a história do português clássico. In: GONÇALVES, M. F.; BANZA, A. P. **Património Textual e Humanidades Digitais**: da antiga à nova filologia. Évora: CIDEHUS, 2013. p. 73-112.

LACERDA, M. F.; CARNEIRO, Z. O; SANTIAGO, H. S. (Org). **Núcleo de Estudos de Língua Portuguesa**. URL: <https://nelp.uefs.br/>. Acesso em: 14 fev. 2023.

LOPES, N. da S. O PEPP e os estudos sobre o português de Salvador. In: **A Cor das Letras (UEFS)**. v. 19, n. Especial, 2018. p. 23-39.

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, RJ. p. 69-105. 2009.

SANTIAGO, H. S.; LACERDA, M. F. O., BRITO, R. C.; CARNEIRO, Z. O. N. **CEDOHS**: um banco de dados sociolinguísticos para a história do português brasileiro. LaborHistórico, Rio de Janeiro, 7 (Especial): 311-329, 2021. DOI: https://doi.org/10.24206/lh.v7iespec.41640 2021. Acesso em: 15 jan. 2022.

ROSÁRIO, Taine do. **Edição Modernizada com uso do eDictor para o banco de textos do NELP/UEFS**: Documentos orais, manuscritos e impressos. Orientadora: Mariana Fagundes de Oliveira Lacerda. 2023, 44 p. Trabalho de Conclusão de Curso (Licenciatura em Letras com Inglês) – Universidade Estadual de Feira de Santana, 2023.