laborhistórico

ISSN 2359-6910
https://revistas.ufrj.br/index.php/lh/

ARTIGO

Recebido em 21 de fevereiro de 2021
Aprovado em 25 de abril de 2021

CE-DOHS: um banco de dados sociolinguísticos para a história do português brasileiro

CE-DOHS: a sociolinguistic database for the history of Brazilian Portuguese

DOI: https://doi.org/10.24206/lh.v7iespec.41640

Huda da Silva Santiago

Professora Assistente da Universidade Estadual de Feira de Santana. Doutorado em Língua e Cultura pela Universidade Federal da Bahia (2019).

E-mail: huda_santiago@uefs.br
ORCID: https://orcid.org/0000-0002-5523-642X

Mariana Fagundes de Oliveira Lacerda

Professora Titular da Universidade Estadual de Feira de Santana. Doutorado em Letras e Linguística pela Universidade Federal da Bahia (2009).

E-mail: marianafagundes@uefs.br
ORCID: https://orcid.org/0000-0003-4335-3458

Rosana Carvalho Brito

Doutoranda pelo Programa de Pós-Graduação em Estudos Linguísticos da Universidade Estadual de Feira de Santana.

E-mail: rosanacarvalhobrito@gmail.com
ORCID: https://orcid.org/0000-0002-8315-2434

Zenaide de Oliveira Novais Carneiro

Professora Plena da Universidade Estadual de Feira de Santana. Doutorado em Linguística pela Universidade Estadual de Campinas (2005).

E-mail: zenaidenovais@gmail.com

ORCID: https://orcid.org/0000-0001-5990-4854

RESUMO

Neste artigo, apresentamos o controle sócio-histórico que é feito na constituição do Corpus Eletrônico de Documentos Históricos do Sertão (CE-DOHS), do Departamento de Letras e Artes (DLA) da Universidade Estadual de Feira de Santana (UEFS), permitindo análises histórico-diacrônicas do português brasileiro. Trata-se de um banco de dados sociolinguísticos, que reúne mais de um milhão de palavras, disponibilizando documentos editados em diferentes versões e seus metadados, além do perfil sociocultural dos escreventes. As fichas catalográficas – com essa caracterização sócio-histórica – trazem também informações sobre o processamento dos documentos. O CE-DOHS vem buscando, nesse trabalho com a dimensão externa da escrita, realizar diálogos interdisciplinares com diferentes campos, como o campo da História Social da Cultura Escrita, contribuindo com a discussão sobre o tratamento metodológico à constituição de *corpora*.

Palavras-chave: Português Brasileiro. Banco de dados. Caracterização sócio-histórica. História Social da Cultura Escrita. Constituição de *corpora*.

ABSTRACT

In this article, we present the socio-historical control that is done in the constitution of the Corpus Eletrônico de Documentos Históricos do Sertão (CE-DOHS), of the Departamento de Letras e Artes (DLA) from the State University of Feira de Santana (UEFS), allowing Brazilian Portuguese historical-diachronic analyzes. The CE-DOHS is a sociolinguistic database, which gathers more than a million words, making available documents edited in different versions and their metadata, in addition to the sociocultural profile of the writers. The catalogs - with this socio-historical description - also provide information on the processing of documents. CE-DOHS has been working on the external dimension of writing, seeking to carry out interdisciplinary exchange with different fields, such as the field of Social History of Written Culture, contributing to the discussion on the methodological treatment of the constitution of corpora.

Keywords: Brazilian Portuguese. Database. Socio-historical description. Social History of Written Culture. Corpora constitution.

Introdução

No âmbito do Programa para a História da Língua Portuguesa (PROHPOR) – UFBA, UEFS, UNEB, UnB –, fundado pela professora Rosa Virgínia Mattos e Silva, na década de 90 do século passado, sempre se defendeu a importância da constituição de *corpus* e dos dados empíricos, tendo em vista pesquisas solidamente fundamentadas. Para a grande pesquisadora na área da Linguística Histórica, os dados nunca foram menos importantes do que os quadros teóricos, que podem mudar, enquanto aqueles permanecem, podendo ser analisados segundo diferentes teorias, em qualquer tempo.

A formação de banco de dados, como o Corpus Eletrônico de Documentos Históricos do Sertão (CE-DOHS), do Núcleo de Estudos de Língua Portuguesa (NELP) da Universidade Estadual de Feira de Santana (UEFS), exige grande esforço e tempo dedicados à sua metodologia de organização: prospecção documental em fontes confiáveis; caracterização sócio-histórica de acervos; edições em diferentes formatos; processamento, armazenamento e disponibilização de dados. Como se vê, um trabalho de pesquisa que envolve diferentes etapas e uma equipe com formação especializada em campos diversos, como a Filologia, a História Social da Cultura Escrita, a Linguística Computacional, entre outros. Como afirma Mattos e Silva (2004, p. 120), "O trabalho com *corpus* não é 'commodo' – pelo contrário, é por vezes penoso, sempre curioso, às vezes divertido [...] Sem dúvida é um trabalho necessário, como base para a reconstrução do passado lingüístico do português que aqui se formou, o português brasileiro."

O CE-DOHS, há mais de 10 anos, investe na pesquisa documental, procurando disponibilizar, na rede mundial de computadores, dados sociolinguísticos que permitem análises histórico-diacrônicas do português brasileiro. Nem sempre, entretanto, tem sido possível controlar todas as variáveis, havendo lacunas na caracterização sócio-histórica dos acervos, afinal os dados históricos *stricto sensu*, diacrônicos, são "[...] os resultados de acidentes históricos para além do controle do investigador" (LABOV, 1982, p. 20). Mattos e Silva (2004, p. 110) lembra ao leitor que

> Todos sabemos que os dados históricos no sentido lato, sincrônicos, com que trabalham os lingüistas, sobretudo os sociolingüistas, que pesquisam sobre *corpus*, permitem, pelo menos, que se controlem o quando, o onde, o quem, o para quem, o tipo de texto dos dados sobre que se aplicam suas análises. Não é de esquecer que os *corpora* sincrônicos já são construídos com o objetivo de ter essas variáveis, além de outras, controláveis.

O estudo da mudança no tempo real, por outro lado, ressalta a autora, com base em c*orpora* diacrônicos, é visto como "[...] a arte de fazer o melhor uso de maus dados" (LABOV, 1982, p. 20).

Neste texto, apresentamos o controle sócio-histórico que é feito na constituição do CE-DOHS, um banco de dados sociolinguísticos, que reúne mais de um milhão de palavras, disponibilizando documentos editados em diferentes versões e seus metadados, além do perfil sociocultural dos escreventes. As fichas catalográficas – com essa caracterização sócio-histórica – trazem também informações sobre o processamento dos documentos. O CE-DOHS vem buscando, nesse trabalho com a dimensão externa da escrita, realizar diálogos interdisciplinares com diferentes campos, contribuindo com a discussão sobre o tratamento metodológico à constituição de *corpora*.

Está organizado o artigo da seguinte forma: na primeira seção, apresentamos, em síntese, a base documental do CE-DOHS; na seção 2, focamos na abordagem contextual dos *corpora* escritos do banco; na terceira seção, trazemos a possibilidade de montar sub*corpora* para estudo do português brasileiro, acessando a plataforma digital; por fim, há nossas considerações finais e as referências bibliográficas consultadas.

1. A base documental do CE-DOHS: seguindo os caminhos da Linguística histórica

O CE-DOHS consiste em um *corpus* documental seriado, representativo tanto das normas vernáculas como das normas cultas. Inicialmente composto por documentos dentro das fronteiras dos sertões, hoje o banco oferece materiais da maior parte do Brasil, apresentando-se, portanto, como uma plataforma de *corpora* para a história do português brasileiro, com acesso livre e gratuito, na rede mundial de computadores.

Figura 1 – Página inicial do site CE-DOHS

Fonte: Site CE-DOHS (http://www5.uefs.br/cedohs/).

O banco em questão encontra motivações na sócio-história do português brasileiro, conforme proposições de Rosa Virgínia Mattos e Silva, muito bem sintetizadas por Lobo (2015, p. 71):

> 1. A história linguística do Brasil não se restringe à história da língua portuguesa no Brasil, nem à história do português brasileiro.
> 2. O português brasileiro emerge em contexto multilíngue: o contato linguístico é, pois, elemento constitutivo da sua formação.
> 3. Na cena linguística do Brasil colonial, destacam-se três atores principais: o português europeu, as línguas gerais indígenas e o português geral brasileiro.
> 4. Africanos e afrodescendentes foram os principais difusores da língua portuguesa no Brasil e os principais formatadores do português brasileiro em sua variante social majoritária - o português popular brasileiro.
> 5. O passado sócio-histórico-linguístico do Brasil deverá ser interpretado para a compreensão do português brasileiro «heterogêneo e variável, plural e polarizado».

A base documental está organizada em dois conjuntos: conjunto 1 – composto por textos escritos entre 1823 e 2000, por indivíduos nascidos no Brasil, a partir de 1724, e por amostras de fala de brasileiros, gravadas na década de 90 do século XX, na Bahia; conjunto 2 – composto por manuscritos produzidos entre 1640 e 1822 por diferentes populações nascidas no Brasil, a partir de 1590, e em processo de edição no mesmo formato do conjunto 1, sendo paulatinamente liberado para acesso. Adicionalmente, há um conjunto de textos escritos no Brasil por portugueses, nos primeiros 150 anos da colonização.

1.1 Os *corpora* escritos

São 36 acervos (considerando os 2 conjuntos de documentos supracitados), de gêneros textuais diversos; a documentação epistolar se destaca no banco, sendo disponibilizadas mais de 1.000 cartas pessoais, em edição semidiplomática e, no campo da Nova Filologia, em edição modernizada, com uso do eDictor (PAIXÃO DE SOUZA; KEPLER; FARIA, 2009).

Figura 2 – *Corpus* Compartilhado diacrônico: Cartas brasileiras

Fonte: Site do CE-DOHS (http://www.tycho.iel.unicamp.br/cedohs/corpora.html)

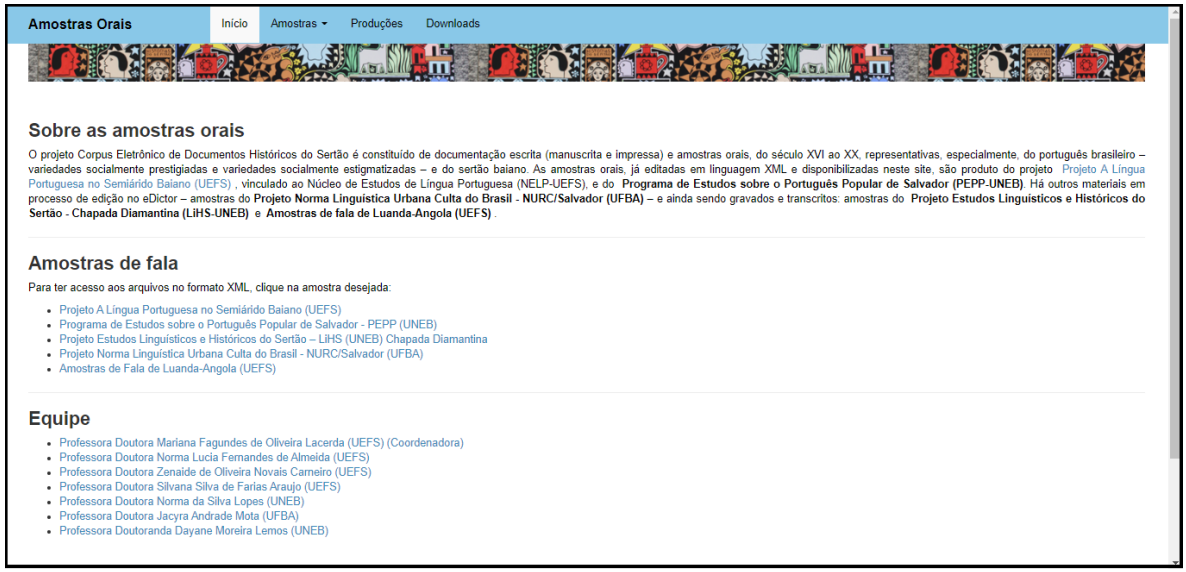
1.2 As amostras de fala

As amostras orais atualmente disponíveis no site CE-DOHS são produto do projeto A Língua Portuguesa no Semiárido Baiano, vinculado ao NELP-UEFS, e do Programa de Estudos sobre o Português Popular de Salvador (PEPP-UNEB). Encontram-se em processo de edição em linguagem XML, com uso do eDictor, as amostras do Projeto Norma Linguística Urbana Culta do Brasil (NURC/Salvador/UFBA)[1]. As amostras do Projeto Estudos Linguísticos e Históricos do Sertão -

[1] As amostras do NURC-SSA estão sendo editadas em linguagem XML pela estudante de Iniciação Científica Taine do Rosário (FAPESB).

Chapada Diamantina (LiHS-UNEB) estão sendo gravadas e transcritas[2], e as Amostras de fala de Luanda-Angola (UEFS) estão em processo de transcrição e revisão[3].

Figura 3 – Amostras Orais do CE-DOHS



Fonte: Site do CE-DOHS (http://www.uefs.br/cedohs/amostrasorais/index.html)

Na próxima seção, focaremos nos *corpora* escritos do banco, em sua caracterização sócio-histórica, tendo em vista análises histórico-diacrônicas.

2. Para uma abordagem contextual dos *corpora* escritos

No âmbito dos estudos em Sociolinguística Histórica, durante o processo de constituição de *corpora*, a busca pela identificação dos perfis socioculturais dos escreventes contribui para enfrentar o desafio de reunir fontes escritas, que sejam *representativas* e *significativas* (BARBOSA, 2006), de um grupo sociocultural de determinada época. Então, com o objetivo de tornar melhores os *maus dados*, a que se refere Labov (1982), as equipes do Projeto *Para a História do Português Brasileiro* (PHPB) têm enfrentado, desde os seus primeiros trabalhos[4], com vigor e paciência, como afirma Mattos e Silva (2002, p. 23), essa difícil tarefa de constituir *corpora*, buscando controlar *o quando*, *o onde*, *o quem*, *o*

[2] As amostras do projeto LiHS estão sendo gravadas e transcritas por Dayane Moreira Lemos (FAPESB), no âmbito de sua tese de doutorado, em desenvolvimento no Programa de Pós-Graduação em Estudos Linguísticos (PPGEL) da UEFS.

[3] As amostras de fala de Luanda-Angola estão em processo de revisão de transcrição, trabalho coordenado pela professora Silvana Silva de Farias Araujo (UEFS), no NELP.

[4] Ver, dentre outros, Barbosa (1999), Lobo (2001) e Carneiro (2005).

para quem e *o tipo de texto*. Na maioria dos casos, a dificuldade maior é com a recuperação do perfil biográfico, tanto do emissor (*quem*), quanto do receptor (*para quem*).

O projeto CE-DOHS tem tentado aproximar-se dessas questões no trabalho com a dimensão externa da escrita, a partir da garantia, na disponibilização dos acervos, de informações sobre os documentos e os escreventes. Na plataforma digital do projeto, na seção que disponibiliza as várias versões da edição, é possível encontrar os chamados "metadados" dos documentos, organizados em fichas catalográficas. Esses metadados são o resultado do trabalho investigativo de teses, dissertações e trabalhos de Iniciação Científica que se dedicaram/dedicam à constituição de cada acervo.

As informações tentam responder às questões metodológicas que perpassam a busca por material para o trabalho de análise linguística, porquanto, sem esse controle sócio-histórico, de acordo com Barbosa (2019), as análises seriam diacrônicas, mas não histórico-diacrônicas. Lima, Marcotulio e Rumeu (2019, p. 70-71) listam uma série dessas questões para as quais o pesquisador deve atentar-se na coleta dos documentos remanescentes. Destacam-se, aqui, algumas delas:

> 6. Como recuperar a informação não sistematizada? [...]
> 9. Como resgatar os perfis sociais dos nossos informantes? Que informações já estão disponíveis no próprio documento e/ou nos instrumentos do acervo?
> 10. Em função de nossas questões pessoais, tais informações são suficientes? Se não, como recuperar informações de natureza social? (LIMA; MARCOTULIO; RUMEU, 2019, p. 70-71)

Os trabalhos que se dedicam a enfrentar essas questões ou algumas delas, no âmbito do CE-DOHS e, de modo mais geral, do PHPB, têm em comum o necessário diálogo com outras áreas, como será comentado a seguir.

2.1 Diálogos interdisciplinares para a caracterização dos acervos

Mattos e Silva (2002) deixa clara a necessidade de um trabalho interdisciplinar, no texto *Reflexões e questionamentos sobre a constituição de corpora para o projeto Para a história do Português brasileiro*, uma conferência proferida no IV Seminário do PHPB, quando comenta sobre as três agendas do projeto, naquele momento: a constituição de um *corpus* diacrônico, o estudo linguístico e o estudo da história social-linguística do Brasil.

> [...] vê-se que o nosso *Projeto* é interdisciplinar: engloba estudos propriamente linguísticos, com interpretações que não priorizam uma teoria única e seu conseqüente método; engloba o trabalho filológico, propriamente dito, ou seja, a edição de textos para uso em análise linguística e, por fim, deverá interrelacionar fatos sócio-históricos da história brasileira, como embasamento essenciais para a reconstrução da sócio-história linguística do português brasileiro, na sua formação nesses passados quinhentos anos. (MATTOS E SILVA, 2002, p. 18)

Então, segundo a autora, os estudos sobre a história do português buscam integrar as pesquisas pelos campos múltiplos da Linguística, da Filologia, da História social, cultural, demográfica e econômica do Brasil.

Um dos diálogos que tem sido bastante produtivo é com o campo da História Social da Cultura Escrita, motivado pelo que propôs Houaiss (1985), quando indicou, como uma das vias para a tarefa de reconstrução da complexa formação sócio-histórica linguística do português brasileiro, *a penetração da língua escrita no Brasil, das origens aos nossos dias*. Essa via é um espaço privilegiado, segundo Tânia Lobo e Klebson Oliveira (inédito), para o encontro entre historiadores da língua e historiadores da cultura escrita.

Essa aproximação à História Social da Cultura Escrita, além de permitir novos olhares em torno da exploração dos arquivos pessoais, das práticas sociais de escrita cotidiana, de pessoas comuns, não ilustres, contribui para um melhor tratamento metodológico à constituição dos *corpora*. Ao fazer referência a uma autêntica *história da cultura escrita*, o paleógrafo Petrucci (2003, p. 7-8) indica o uso do método indiciário, ao caracterizá-la como uma disciplina que deve ocupar-se da história da produção, das características formais e dos usos sociais da escrita em uma determinada sociedade, independentemente das técnicas e materiais utilizados.

Então, com os desdobramentos de uma paleografia renovada, Petrucci (2003) apresenta, como problemas a serem enfrentados pelo investigador, além daqueles já característicos da paleografia de leitura (*o quê?*) e da paleografia de análise (*quando? onde?* e *como?*), os problemas que se referem a uma História da Cultura Escrita, propondo os questionamentos sobre a difusão social da escrita (*quem?*) e a sua função social (*para quê?*). São questões que contribuem diretamente no trabalho de constituição de *corpora* e de investigação em torno da história social linguística.

2.2 Organização dos metadados dos acervos

A contextualização sócio-histórica de cada *corpus* que compõe o CE-DOHS é discutida e detalhada em teses e dissertações que se dedicaram à constituição, edição e estudo linguístico de cada

acervo[5]. A partir desses trabalhos, nas fichas catalográficas dos documentos disponíveis na plataforma CE-DOHS, apresentam-se as informações possíveis sobre os escreventes – como nome, data de nascimento, local de nascimento, gênero – e sobre os documentos, como destinatário, data, local, gênero do documento e conteúdo. As fichas catalográficas apresentam também informações sobre o processamento do documento, com o nome do pesquisador que trabalhou em cada etapa de edição e de revisão. Na fase atual do projeto, estão sendo incluídas outras informações em alguns acervos: escolaridade, nível de habilidade com a escrita e profissão do escrevente[6].

Figura 4 – Exemplo de ficha catalográfica disponibilizada na plataforma CE-DOHS

Ficha catalográfica

Dados do documento	
acervo	CS
autor	Antonio Fortunato da Silva.
destinatário	João Carneiro de Oliveira.
data de nascimento	06 de setembro de 1936.
genero do autor	Masculino.
nascido(N)/radicado(R)	Fazenda Varjota, município de Riachão do Jacuípe, BA.
data do documento	9 de julho de 1962
genero do documento	Cartas pessoais.
conteúdo	Informa sobre envio de dinheiro para pagamento de suas dívidas.
referencia	SANTIAGO, H. S. (Org.) . CD-ROM 1. (Vol. 3) Cartas em Sisal: Riachão do Jacuípe, Conceição do Coité e Ichu (1906-2000): edição fac-similada. In. SANTIAGO, H. S. (Org.); CARNEIRO, Z. O. N. (Org.); OLIVEIRA, K. (Org.). Volume 3 de Cartas brasileiras (1809-2000): coletânea de fontes para o estudo do português.. 1. ed. Feira de Santana: UEFS Editora, 2011.
fonte	Depoimentos concedidos por João Carneiro de Oliveira e Almerinda Maria Oliveira nos dias 05 de setembro de 2010 e 30 de janeiro de 2011. Depoimento concedido por Antonio Fortunato da Silva e Maura Ribeiro da Silva no dia 12 de março de 2011. Documento de identidade de Antonio Fortunato, RG.
Processamento	
edição semi-diplomática	Huda da Silva Santiago.
edição XML	Janaina de Oliveira Costa Mascarenhas.
revisão da edição semi-diplomática	Huda da Silva Santiago.
revisão final	Huda da Silva Santiago.
1ª revisão da edição XML	Janaina de Oliveira Costa Mascarenhas e Huda da Silva Santiago.
número de palavras	210

Fonte: Site do CE-DOHS (http://www.tycho.iel.unicamp.br/cedohs/corpora/CS/03-AFS-09-07-1962.xml)

[5] A exemplo das dissertações de Batista (2017), Santos (2019), Brito (2020), Cardoso (2020), Souza (em construção) e Santos (em construção) e das teses de Santiago (2019), Silva (em construção) e Santos (em construção). Alguns desses trabalhos disponibilizam ainda os *índices analíticos* dos documentos editados, colaborando, com isso, para a caracterização externa dos materiais.

[6] O processo de revisão e ampliação dos metadados dos acervos do CE-DOHS tem sido desenvolvido pelas estudantes de Iniciação Científica Larissa de Souza (PEVIC/UEFS) e Ticianny da Silva (PROBIC/UEFS), com a colaboração de Igor Leal, Lara Cardoso, Priscila Batista e Shirley Guedes.

2.2.1 O perfil biográfico do escrevente

O questionamento sobre a difusão social da escrita, sobre o ambiente sociocultural a que pertencia o escrevente, na caracterização de *quem* escreveu, é fundamental, considerando-se que, segundo Castro (2004, p. 91-92), o que o documento oferece é um resultado (modificado pela passagem do tempo sobre o manuscrito) do percurso da mão, "[...] condicionado por diversos factores – dialecto natal, dialecto do local de produção do documento, aprendizado da escrita, modelos de documentos em que se inspira".

O pesquisador pode deparar-se, na reconstrução do perfil sociocultural, com escreventes que são socialmente/historicamente conhecidos ou desconhecidos. Como propõem Lopes e Rumeu (2018), para os conhecidos é possível levantar, além de informações circunscritas ao próprio documento e da identificação de suas redes sociais e de escrita, informações arquivísticas e enciclopédicas. Estas últimas não são recursos possíveis para os escreventes desconhecidos. No âmbito do CE-DOHS, essas estratégias têm sido produtivas para a caracterização da etnia, do nível de escolarização e habilidade/inabilidade com a escrita, do sexo, da profissão, da estratificação social, da data e local de nascimento (naturalidade e nacionalidade), por exemplo.

Para os escreventes desconhecidos, as dificuldades para reconstituição do perfil sociocultural são maiores. As pistas legadas pelos próprios escritos ainda são úteis, mas nem sempre fornecem informações completas e precisas. Alguns trabalhos desenvolvidos pela equipe do CE-DOHS investigam o grau de habilidade com a escrita, o que pode trazer pistas sobre a escolaridade e contribuir para a caracterização de redatores pouco ilustres, como recomenda Barbosa (2006), uma estratégia que se mostrou produtiva também para Lopes e Rumeu (2018). Nessa linha, situa-se a metodologia proposta em Santiago (2019), para identificação do nível de domínio da escrita, que pode ser aplicada na busca de pistas para uma melhor caracterização, seja de escreventes conhecidos – como os que constituem o *corpus* produzido por redatores não ilustres utilizado pela pesquisadora –, seja de escreventes desconhecidos.

Dentre os redatores do CE-DOHS, há escreventes nascidos no século XX que estão vivos ou que têm familiares vivos. Então, uma possibilidade tem sido investir na metodologia da História Oral (PORTELI, 2016; PRINS, 1992), como também ilustra o trabalho de Santiago (2019), por meio da produção de entrevistas-narrativas com remetentes e destinatários do acervo *Cartas em Sisal*. Outros trabalhos, ainda que não assumam a metodologia da História Oral, desenvolveram, nesse caso, contatos pessoais com os escreventes ou seus familiares, para conversar e saber sobre os perfis biográficos, como os estudos de Batista (2017) e Brito (2020).

2.2.2 O texto: para que e para quem foi escrito

O diálogo com os estudos da História Social da Cultura Escrita, além de contribuir para o pesquisador em sócio-história linguística na tarefa de investigar *quem* escreveu, também é necessário na busca de saber *para que* o texto foi escrito, o que pressupõe a incursão na história social dos escritos. As pistas deixadas no texto convertem-se em indícios para que se alcance alguma notícia sobre o lugar social da escrita, os usos que se faziam dessa técnica e o que eles representaram em uma dada situação. Constitui interesse do pesquisador também situar o sujeito que escreve nesse contexto, refletindo sobre seu papel, o alcance social da escrita no período e o que representava ter o domínio dessa prática em determinada conjuntura social. O que está em questão, então, não é somente a apropriação da escrita enquanto saber institucionalizado, mas as variadas formas de participação em atividades e práticas de escrita. Analogamente, o interesse recai tanto na mão que escreve, quanto no indivíduo a quem o texto é destinado.

Seguindo essa linha, as pesquisas associadas ao CE-DOHS atentam para a história da difusão e da função social dos escritos, entendendo que essa é, além do mais, uma das vias possíveis para a reconstrução mais apropriada do português brasileiro. Os trabalhos desenvolvidos até o momento se aproximam em graus variados desse objetivo. Há trabalhos que, mesmo indiretamente, apresentam dados sobre os usos que eram feitos da escrita e o alcance dessa prática; outros estudos propõem-se a observar, mais de perto, essas questões em um verdadeiro esforço de reconstituição da história social da escrita na dimensão espaço-temporal em que os documentos foram produzidos e circularam.

2.2.3 Datação e localização

As informações sobre data e local de produção dos textos, em alguns casos, estão dadas nos próprios documentos, de modo especial no gênero carta, no qual, em geral, informa-se, no cabeçalho, a data e o local de onde a missiva está sendo remetida. E essa é uma das motivações para o interesse particular dos estudos linguísticos históricos por esse gênero[7]. A datação e a localização dos documentos podem ser estabelecidas também por inferência. Nesse caso, a partir de informações sobre o perfil biográfico do escrevente, do conteúdo do próprio escrito ou, ainda, de informações contidas em outros materiais do mesmo acervo, é possível, muitas vezes, estipular a data e o local de produção.

[7] O conjunto 1 da base documental do CE-DOHS (composto por textos escritos entre 1823 e 2000, por indivíduos nascidos no Brasil) é constituído, majoritariamente, por cartas.

No entanto, o estabelecimento dessas informações é apenas a etapa inicial de um processo que envolve a contextualização sócio-histórica mais ampla dos *corpora* constituídos no âmbito do Projeto CE-DOHS. Entende-se, pois, que tão importante quanto a edição dos materiais e as análises linguísticas é a investigação das redes de relações sociais do momento e do espaço em que os textos foram produzidos. É assim que as teses e dissertações, vinculadas ao CE-DOHS, que se dedicam à preparação e disponibilização de *corpora* para investigação linguística, reservam parte do trabalho para a caracterização sócio-histórica dos acervos, o que envolve considerar, dentre outros pontos, fatos históricos, aspectos políticos, sociais, demográficos e econômicos da região e, ainda, dados do acesso à escolarização e às práticas de leitura e escrita.

Considerem-se, para citar apenas dois exemplos, os trabalhos de Santos (2019) e de Silva (em construção), que apresentam dados da distribuição dos diferentes grupos étnicos no sertão baiano, sugerindo a incidência pluriétnica da região e, consequentemente, a interação de diferentes matizes linguísticas no contexto em que o *Livro do Gado* e o *Livro de Razão*, respectivamente utilizados como *corpora* dessas pesquisas, foram produzidos, entre fins do século XVIII e inícios do século XIX, dando ênfase ao processo de fundação do povoado de Bom Jesus dos Meiras, no alto sertão baiano, onde se situava a fazenda de Criação do Sobrado do Brejo do Campo Seco. O *Livro de Razão* fornece um número significativo de dados sobre as principais atividades econômicas do período (agricultura, pecuária e comércio) e os perfis que compunham a estratificação social da região em que os fazendeiros desempenhavam funções de grande relevância.

Esse cuidado metodológico adotado pela equipe do CE-DOHS amplia o potencial analítico dos trabalhos, conjecturando três aspectos basilares apontados por Lopes e outros (2010) para uma caracterização sociocultural do indivíduo no cenário maior de elaboração e circulação dos textos: o percurso de vida dos sujeitos, o contexto de produção dos textos e o mapeamento e descrição das redes de escrita.

3. Montando sub*corpora*: um banco de dados sociolinguísticos à disposição do pesquisador

O CE-DOHS resulta, como vimos, de uma intensa prospecção em fontes confiáveis, com textos editados de forma fidedigna e com autoria bem controlada. O pesquisador pode, acessando a plataforma, personalizar o *corpus*, de acordo com seu interesse; pode optar por separar os materiais, no que concerne ao autor, por: *etnia* (indígenas, brancos, negros do Brasil, mulatos, mamelucos e pardos); nível de escolarização e habilidade/inabilidade com a escrita; sexo; profissão; estratificação social; data e local de nascimento do autor (naturalidade e nacionalidade); no que concerne ao documento, pode

separar materiais por: data e local de escrita, meio urbano e meio rural, para quê foi escrito e a quem foi destinado.

No âmbito do NELP-UEFS, no campo da Linguística de *Corpus*, que, segundo Sardinha (2004, p. 325), "[...] ocupa-se da coleta e da exploração de *corpora*, ou do conjunto de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística", foi desenvolvido o *E-Corp*, com aplicação inicial no CE-DOHS.

> Com o crescente desenvolvimento da construção de banco de dados eletrônicos, surge a necessidade de criação de ferramentas que auxiliem na exploração de documentos em formato XML. Dessa maneira, na tentativa de otimizar o contato inicial do pesquisador com os *corpora*, foi desenvolvida a ferramenta *E-Corp*, que torna a busca nos bancos de dados mais rápida e confiável, além de permitir a exploração dos acervos, ajudando na construção de *subcorpora*, já que é possível filtrar as informações sobre o documento a partir dos metadados de cada documento (SOUZA *et al*, 2018, p. 14).

É possível, utilizando a referida ferramenta, fazer buscas personalizadas – estando disponível na plataforma uma riqueza de metadados (como visto na seção 2) –, montando sub*corpora*, de acordo com o interesse do pesquisador.

Considerações finais

No presente artigo, demonstramos por que caminhos o CE-DOHS vem constituindo-se como um banco de dados sociolinguísticos para a história do português brasileiro, controlando, sempre que possível, as variáveis sociais, no campo da Linguística de base empírica, cujos fundamentos, segundo Callou (2009, p. 126), são trazidos à tona, no século XVI, por Fernão de Oliveira.

Considerando as relações sociolinguísticas, isto é, a interação entre o sistema de relações linguísticas e as disposições nas quais ele se atualiza, o CE-DOHS é hoje visto como um dos *corpora* mais significativos – disponibilizando sobretudo documentação epistolar – para análises histórico-diacrônicas do PB.

Um conjunto de estudiosos vem enfrentando um rico leque de desafios na Linguística Histórica, para atingir o objetivo de reconstrução de uma história do português brasileiro. Essa reconstrução do passado, com base em textos escritos remanescentes, e o estudo sociolinguístico de formas linguísticas históricas exigem um alinhamento teórico e metodológico como o proposto pela Sociolinguística Histórica (ROMAINE, 1982; HERNÀNDEX-CAMPOY; CONDE-SILVESTRE, 2012). Os pesquisadores do CE-DOHS, "Como quixotes ou como loucos, ou apenas como brasileiros

interessados em compreender um aspecto fundamental da sua história pregressa" (MATTOS E SILVA, 2004, p. 67), juntam-se a esse conjunto, buscando colaborar com essa agenda desafiadora, oferecendo um *corpus* documental seriado e um banco de dados sociolinguísticos.

Referências bibliográficas

BARBOSA, Afrânio Gonçalves. **Para uma história do português colonial**: aspectos linguísticos em cartas do comércio. Tese (Doutorado em Língua Portuguesa) – Faculdade de Letras, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1999.

BARBOSA, Afrânio Gonçalves. Tratamento dos corpora de sincronias passadas da língua portuguesa no Brasil: recortes grafológicos e linguísticos. *In*: LOBO, Tânia et al. (org.). **Para a história do português brasileiro**: novos dados, novas análises. v. 6, t. 2. Salvador: EDUFBA, 2006. p. 761-780.

BARBOSA, Afrânio Gonçalves. A Plataforma de *corpora* do PHPB: uma apresentação *ad infinitum*. *In*: Ataliba Teixeira de Castilho. (org.). **História do português brasileiro**: corpus diacrônico do português brasileiro. v. 2. São Paulo: Contexto, 2019. p. 16-52.

BATISTA, Priscila Starline Estrela Tuy. **A variação no uso de você/tu em cartas particulares baianas do século XX em relações de simetria**. Dissertação (Mestrado em Estudos Linguísticos) – Programa de Pós-Graduação em Estudos Linguísticos, Universidade Estadual de Feira de Santana, Feira de Santana, 2017.

BRITO, Patrícia Santos de Jesus. **Cartas marienses (séc. XX)**: edição fac-similar e semidiplomática e estudo da concordância nominal. Dissertação (Mestrado em Estudos Linguísticos) – Programa de Pós-Graduação em Estudos Linguísticos, Universidade Estadual de Feira de Santana, Feira de Santana, 2020.

CALLOU, Dinah. De Fernão de Oliveira e da (Socio)lingüística. *In*: ABAURRE, Maria Bernadete; PFEIFFER, Claudia; AVELAR, Juanito. (org.). **Fernão de Oliveira**: um gramático na história. Campinas: Pontes, 2009.

CARDOSO, Lara da Silva. **A gramática dos pronomes clíticos no Brasil Colônia**: o português clássico na história do português brasileiro. Dissertação (Mestrado em Estudos Linguísticos) – Programa de Pós-Graduação em Estudos Linguísticos, Universidade Estadual de Feira de Santana, Feira de Santana, 2020.

CARNEIRO, Zenaide de Oliveira Novais. **Cartas brasileiras (1809-1904)**: um estudo linguístico-filológico. 4 v. Tese (Doutorado em Linguística) – Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, São Paulo, 2005.

CORPUS CE-DOHS. **Corpus Eletrônico de Documentos Históricos do Sertão** (FAPESB 5566/2010 - Consepe UEFS 202/2010). Coordenado por Zenaide de Oliveira Novais Carneiro e Mariana Fagundes de Oliveira Lacerda (UEFS). [Projeto Vozes do Sertão em Dados: história, povos e formação do português brasileiro (CNPq. 401433/2009-9 - Consepe UEFS 102/2009). (CNPq. Processo 401433/2009-9/Consepe: 102/2009) Disponível em: http://www5.uefs.br/cedohs/. Acesso em 21 jan. 2021.

CASTRO, Ivo. **Introdução à história do português –** Geografia da língua, português antigo. Lisboa: Edições Colibri, 2004.

HERNÁNDEX-CAMPOY, Juan Manuel; CONDE SILVESTRE, Juan Camilo. **The Handbook of Historical Sociolinguistics**. Oxford: Wiley-Blackwell, 2012.

HOUAISS, Antônio. **O português no Brasil**. Rio de Janeiro: UNIBRADE, 1985.

LABOV, William. Building on Empirical Foundations. *In*: Lehmann, W. &Malkiel, Y. (org.). **Perspectives on Historical Linguistics**. Amsterdam/Philadelphia: John Benjamins, 1982. p. 17-92.

LIMA, Alexandre Xavier; MARCOTULIO, Leonardo Lennertz; RUMEU, Márcia Cristina de Brito. Experiências metodológicas em constituição de corpora: pistas para um pesquisador iniciante. *In*: CASTILHO Ataliba Teixeira de. (org.). **História do português brasileiro**: corpus diacrônico do português brasileiro. v. 2. São Paulo: Contexto, 2019. p. 68-91.

LOBO, Tânia Conceição Freire; OLIVEIRA, Klebson. **História da cultura escrita no Brasil**: um programa de investigação. (inédito).

LOBO, Tânia Conceição Freire. **Para uma sociolinguística histórica do português no Brasil**. Edição filológica e análise linguística de cartas particulares do Recôncavo da Bahia, século XIX. Volume II. Tese (Doutorado em Filologia e Língua Portuguesa) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2001.

LOBO, Tânia Conceição Freire. Rosa Virgínia Mattos e Silva e a história social linguística do Brasil, **Estudos de Linguística Galega**, v. 7, Santiago de Compostela, p. 69-82, 2015.

LOPES, Célia Regina dos Santos; RUMEU, Márcia Cristina de Brito. A identificação dos perfis socioculturais dos redatores de *corpora* históricos: encaminhamentos metodológicos, **Diadorim**, Rio de Janeiro, vol. 20 – Especial, p. 147-168, 2018.

LOPES, Célia Regina dos Santos *et al*. Reflexões metodológicas para a análise sociocultural de redatores em corpora históricos, **Gragoatá**, Niterói, n. 29, p. 239-253, 2010.

MATTOS E SILVA, Rosa Virgínia. Reflexões e questionamentos sobre a constituição de corpora para o projeto Para a história do Português brasileiro. *In*: DUARTE, Maria Eugênia Lamoglia; CALLOU, Dinah. (org.). **Para a história do português brasileiro**: notícias de corpora e outros estudos. v. 4. Rio de Janeiro: Faculdade de Letras da UFRJ/FAPERJ, 2002. p. 17-28.

MATTOS E SILVA, Rosa Virgínia. **Ensaios para uma sócio-história do português brasileiro**. São Paulo: Parábola, 2004.

PAIXÃO DE SOUSA, Maria Clara; KEPLER, Fábio; FARIA, Pablo. E-dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. *In*: VIII Encontro de Linguística de Corpus, 2009. Rio de Janeiro, **Anais do VIII Encontro de Linguística de Corpus**. Rio de Janeiro: UERJ, 2009. p. 69-105.

PETRUCCI, Armando. **La ciencia de la escritura**: primeralección de paleografía. Buenos Aires: Fondo de Cultura Económica de Argentina, 2003.

PLATAFORMA de corpora do PHPB. Disponível em: https://sites.google.com/site/corporaphpb/home/plataforma-de-corpora-phpb. Acesso em: 21 jan. 2021.

PRINS, Gwyn. História oral. *In*: BURKE, Peter (org.). **A escrita da história**: novas perspectivas. Tradução de Magda Lopes. São Paulo: Editora da Universidade Estadual Paulista, 1992. p. 163-198.

PORTELLI, Alessandro. **História oral como arte da escuta**. São Paulo: Letra e Voz, 2016.

ROMAINE, Suzzane. **Socio-historical linguistics**: its status and methodology. Cambridge: Cambridge University Press, 1982.

SANTIAGO, Huda da Silva. **A escrita por "mãos inábeis"**: uma proposta de caracterização. Tese (Doutorado – Língua e Cultura) – Programa de Pós-Graduação em Língua e Cultura, Universidade Federal da Bahia, Salvador, 2019.

SANTOS, Elaine Brandão. **O Livro do Gado do Brejo do Campo Seco (Bahia)**: edição semidiplomática e descrição de índices grafo-fonéticos. Dissertação (Mestrado em Estudos Linguísticos) – Programa de Pós-Graduação em Estudos Linguísticos, Universidade Estadual de Feira de Santana, Feira de Santana, 2019.

SARDINHA, Tony Berber. Lingüística de corpus: histórico e problemática. **D.E.L.T.A.**, São Paulo, v. 16. n. 2. p. 323-367, 2000. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005. Acesso em: 11 mar 2020.

SOUZA, Igor Leal *et al*. A ferramenta de busca E-CORP aplicada ao Corpus Eletrônico de Documentos Históricos do Sertão, **A Cor das Letras**, v. 19, n. 2, Feira de Santana, p. 8-21, 2018.