Revista RBBA

Revista Binacional Brasil Argentina

CORPUS ELETRÔNICO DE DOCUMENTOS HISTÓRICOS DO SERTÃO: ETAPA 1 (1750-2000)

CORPUS ELECTRÓNICO DE DOCUMENTOS HISTÓRICOS DE SERTÃO: ETAPA 1 (1750-2000)

Zenaide de Oliveira Novais Carneiro
Universidade Estadual de Feira de Santana – UEFS
zenaide.novais@gmail.com

Mariana Fagundes de Oliveira Lacerda
Universidade Estadual de Feira de Santana – UEFS
marianafag@gmail.com

Resumo
Este trabalho apresenta o Banco CE-DOHS – Corpus Eletrônico de Documentos Históricos do Sertão, na sua primeira etapa, que abrange o período que vai de 1750 a 2000, caracterizado pelo multilinguismo localizado. O CE-DOHS é a versão eletrônica – com textos editados em linguagem xml – do banco DOHS, do projeto *Vozes do Sertão em Dados: história, povos e formação do português brasileiro*, com textos em edição semidiplomática, além de amostras orais. A edição eletrônica é feita, no âmbito do CE-DOHS, usando o eDictor, desenvolvido por Paixão de Sousa, Kepler e Faria (2010); trata-se de um editor de textos especialmente voltado ao trabalho filológico e à análise linguística automática. Finalizada a primeira etapa, no ano de 2018, o banco CE-DOHS tem mais de um milhão de palavras, colaborando, de maneira muito significativa, com o Projeto Nacional para a História do Português Brasileiro (PHPB), do qual é parceiro.
Palavras-chave: Português Brasileiro. Banco de Dados. Edições Eletrônicas.

Resumen
Este documento presenta el Banco EC-DOHS - Corpus de Documentación Electrónica de Sertão, en su primera etapa, que abarca el período de 1750 a 2000, caracterizado por el multilingüismo localizado. CE-DOHS es la versión electrónica, con textos editados en lenguaje XML, del banco DOHS, del proyecto Vozes do Sertão en Datos: historia,

pueblos y formación del portugués brasileño, con textos en edición semidiplomática y muestras orales. La edición electrónica se realiza, en el marco de CE-DOHS, utilizando eDictor, desarrollado por Paixão de Sousa, Kepler y Faria (2010); Es un editor de texto enfocado especialmente en el trabajo filológico y el análisis lingüístico automático. Después de la primera fase, en 2018, el banco CE-DOHS tiene más de un millón de palabras, colaborando de manera muy significativa con el Proyecto Nacional de Historia del Portugués Brasileño (PHPB), del cual es socio.
Palabras clave: Portugués brasileño. Banco de datos. Ediciones electrónicas.

1- Introdução

O projeto CE-DOHS: Corpus Eletrônico de Documentos Históricos do Sertão[i], coordenado pelas professoras Zenaide de Oliveira Novais Carneiro e Mariana Fagundes de Oliveira Lacerda, faz parte do Núcleo de Estudos de Língua Portuguesa (NELP), do Departamento de Letras e Artes (DLA) da Universidade Estadual de Feira de Santana (UEFS).

O NELP trabalha com três agendas: (i) formação de banco de textos de língua portuguesa; (ii)estudo sociohistórico; (iii)estudo linguístico. O CE-DOHS destaca-se, oferecendo, por meio de parceria tecnológica com o projeto Corpus Histórico do Português Tycho Brahe[ii], da Universidade Estadual de Campinas e que está sob a coordenação da professora doutora Charlotte Galves, um banco eletrônico de mais de um milhão de palavras, para estudo da história do português brasileiro, numa parceria com o Projeto Nacional para a História do Português Brasileiro (PHPB). Essa constituição de banco de dados, segundo Bacelar do Nascimento (2004, p. 1),

> [...] favorece essencialmente uma Linguística descritiva, fortemente apoiada pelas 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.

As edições eletrônicas, no âmbito do CE-DOHS, são feitas a partir das edições semidiplomáticas do Banco DOHS, do projeto Vozes do Sertão em Dados: história, dados e formação do Português Brasileiro, também do NELP, e de edições semidiplomáticas

realizadas por pesquisadores do CE-DOHS. Usa-se, para a edição em linguagem xml ou eletrônica o eDictor, desenvolvido por Paixão de Sousa, Kepler e Faria (2010); trata-se de um editor de textos especialmente voltado ao trabalho filológico e à análise linguística automática.

O CE-DOHS está organizado em duas etapas: etapa 1, que abrange documentos de 1750 a 2000, período que se caracteriza pelo multilinguismo localizado, e etapa 2, que recua mais no tempo, com documentos de 1500 a 1750. Neste texto, trataremos da etapa 1 do projeto, finalizada em 2018.

O artigo está estruturado da seguinte forma: apresenta-se, para começar, o trabalho de edição eletrônica dos textos do banco DOHS, que fez nascer o CE-DOHS, na era das humanidades digitais; a etapa 1, seus objetivos e documentos são apresentados na parte 2; na sequência, considera-se a produtiva parceria entre o CE-DOHS e o PHPB.

2- Do banco DOHS ao banco CE-DOHS: na era das humanidades digitais

De acordo com Gonçalves e Banza (2013, p. 4),

> 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.

O projeto CE-DOHS, criado em 2012, com financiamento da Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB)[iii] e, hoje, com 40 participantes – entre pesquisadores doutores, estudantes de Pós-Graduação e de Graduação –, aproxima o campo filológico e o campo computacional, promovendo a edição, em linguagem XML, dos textos editados tradicionalmente, segundo critérios de edição semidiplomática, pelos pesquisadores do projeto Vozes do Sertão em Dados, criado em 2009, com financiamento do Conselho Nacional de Desenvolvimento Científico e tecnológico (CNPq)[iv], e por pesquisadores do CE-DOHS, que vêm sempre buscando diversificar o banco, com textos representativos das vertentes popular, sobretudo, e culta do português brasileiro.

A aproximação entre o campo filológico e o campo computacional – observada desde a década de 1990 – encontra-se atualmente em plena expansão. O trabalho em ambiente digital no campo da Filologia e da Linguística Histórica tem sido cada vez mais significativo, fazendo surgir, segundo Crane *et al*. (2008), uma nova Filologia, a *e-philology*, ou determinando, de acordo com Schreibman *et al*. (2004), o nascimento das Humanidades Digitais.

2.1- O método Lapelinc

Como dito, pesquisadores do CE-DOHS, com o objetivo de diversificar o banco – que é constituído, especialmente, de documentação epistolar –, vêm editando textos de outros gêneros, como livros de fazenda, por exemplo.

Foi com o *Livro do Gado* e o *Livro de Razão* do arquivo do Sobrado do Brejo do Campo Seco, no sertão baiano[v], que o projeto iniciou, em 2012, a fotografia de documentos, de acordo com o método do Laboratório de Pesquisa em Linguística de Corpus (Lapelinc), da Universidade Estadual do Sudoeste da Bahia (UESB), coordenado pelo professor doutor Jorge Viana Santos e pela professora doutora Cristiane Namiuti.

O método em questão trata-se de uma técnica fotográfica, cientificamente controlada, específica para a captura de manuscritos históricos; são essas etapas:

> 1) Controle: etapa de captura de informação da fonte (por exemplo, catalogação de dados de um livro a ser fotografado); 2) Captura fotográfica da imagem do original: fotografia sequenciada dos documentos utilizando equipamentos adequados, inseridos na imagem a quantidade necessária de dados que garanta a sua relação com o objeto que a originou. Ou seja: fotografa-se o DF para se formar o DD; 3) Catalogação no *Database* Dovic das folhas-imagens componentes do documento; 4) Edição; 5) Criação de imagens de uso co-indexadas à imagem-original (SANTOS; BRITO, 2014, p. 424).

A mesa cartesiana é um instrumento fundamental desse método, garantindo a qualidade dos aspectos do documento, durante a captura digital. Segue a imagem da mesa cartesiana:

Figura 1- Mesa cartesiana (*Layout*)
Fonte: Santos e Brito (2014, p. 425)

Os elementos que a compõem, sinalizados por números, são assim descritos pelos autores:

> a) Escala de tom (1) e escala de cores (2): sendo escalas científicas elaboradas para o controle fotográfico, possui amostras de tons e cores com parâmetros que, podem ser interpretados por programas e *softwares* de edição e leitura de imagem, capazes por isso de, por exemplo, recuperar numa tela de computador as tom/cores originais de um documento, independente da leitura que o olho humano faça. b) Instrumentos de medição (3, 4, 5): sendo escalas científicas elaboradas para controle milimétrico, do modo como estão dispostas, formam um perfeito plano cartesiano, capaz de matematicamente permitir o cálculo preciso das medidas de quaisquer documentos (livros, folhas...), independente da sua posição. c) Informações catalográficas (6), paginação (7) 4, sequenciação (8) 5: garantem um vínculo permanente entre o DF e o DD (SANTOS; BRITO, 2014, p. 425).

Como se vê, o método conserva as características físicas do texto, como, por exemplo, a cor, o tamanho, a paginação, etc., de forma próxima ao original. Entre as vantagens do método, está também a facilidade de aumentar o texto original na tela do computador, para verificar os detalhes ou tirar dúvidas em relação à escrita.

Para a captura dos *fac-símiles* do *Livro do Gado* e do *Livro de Razão*, pelo método Lapelinc, em 2012, não foi utilizada a mesa cartesiana tal como demonstrada na figura 1, mas a placa preta (primeira versão da mesa), haja vista que aquela só foi aperfeiçoada em 2014,

como descrito na Dissertação de Mestrado de Brito (2015). A seguir, figura que ilustra a aplicação do método Lapelinc no *Livro do Gado*[vi]:

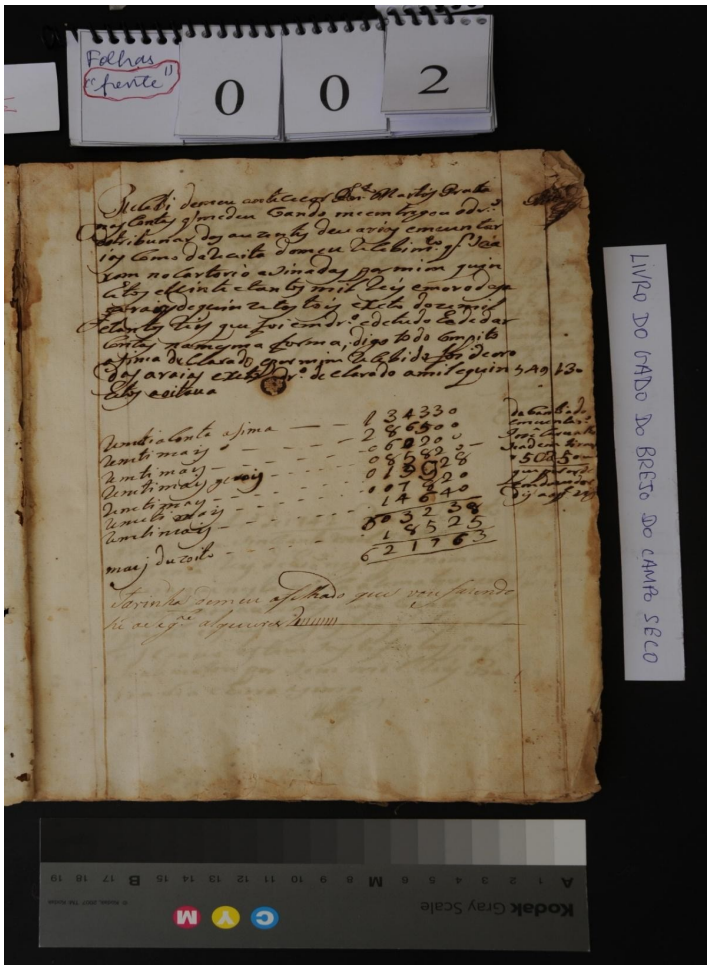

Figura 2- Aplicação do método Lapelinc no *Livro do Gado*
Fonte: CE-DOHS/Foto: Jorge Viana (UESB/Lapelinc)

2.2- O eDictor

As edições eletrônicas, em linguagem *xml*, do banco CE-DOHS, são feitas com uso do eDictor, que 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).

As edições filológicas (ou textos-fontes) recebem, nos *corpora* digitais, uma versão modernizada; são padronizadas a grafia, a acentuação, desenvolvidas abreviaturas, ficando

todas essas alterações feitas com uso do eDictor visíveis ao leitor, o que possibilita o controle e mapeamento das intervenções realizadas nos textos, garantindo a recuperabilidade das formas originais. São mantidas na edição modernizada mudanças de parágrafo, de linha, as correções do autor, os acidentes do suporte, a orientação da escrita etc., sendo, desta maneira, oferecida uma versão eletrônica de textos, sem perder o rigor filológico.

No âmbito do CE-DOHS, procura-se seguir os mesmos critérios de edição digital e de anotação morfossintática que seguem outros projetos de *corpora* eletrônicos, como o projeto Tycho Brahe, já referido na introdução deste texto, o projeto Labor Histórico, da Universidade Federal do Rio de Janeiro (UFRJ), o projeto Post Scriptum: archivo digital de escritura cotidiana em Portugal e Espanha na Época Moderna, do Centro Linguístico da Universidade de Lisboa (CLUL), o que garante maior praticidade no trabalho e nas consultas e maior integração entre os pesquisadores[vii].

Tomando como exemplo uma carta do acervo Cartas para Vários Destinatários (1809-1904), apresenta-se, a seguir, na Figura 3, o processo de edição no eDictor, e, na Figura 4, está a edição modernizada final[viii]:

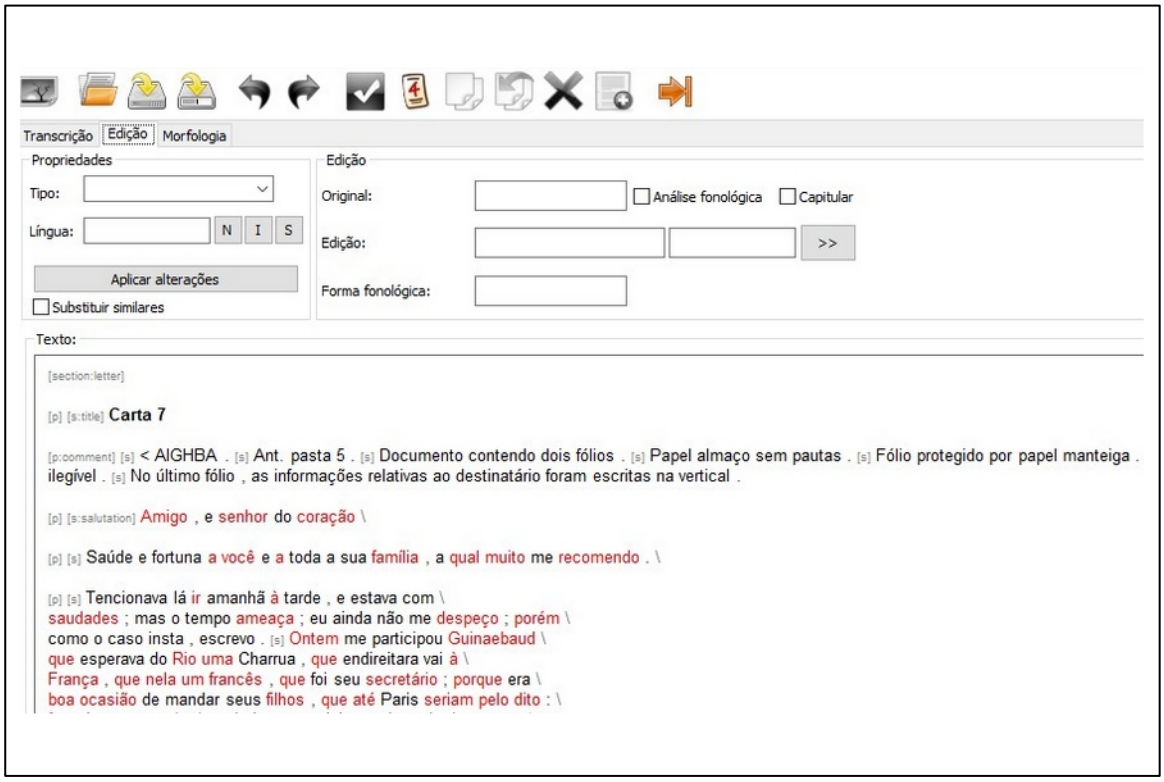

Figura 3- Processo de edição no eDictor
Fonte: CE-DOHS

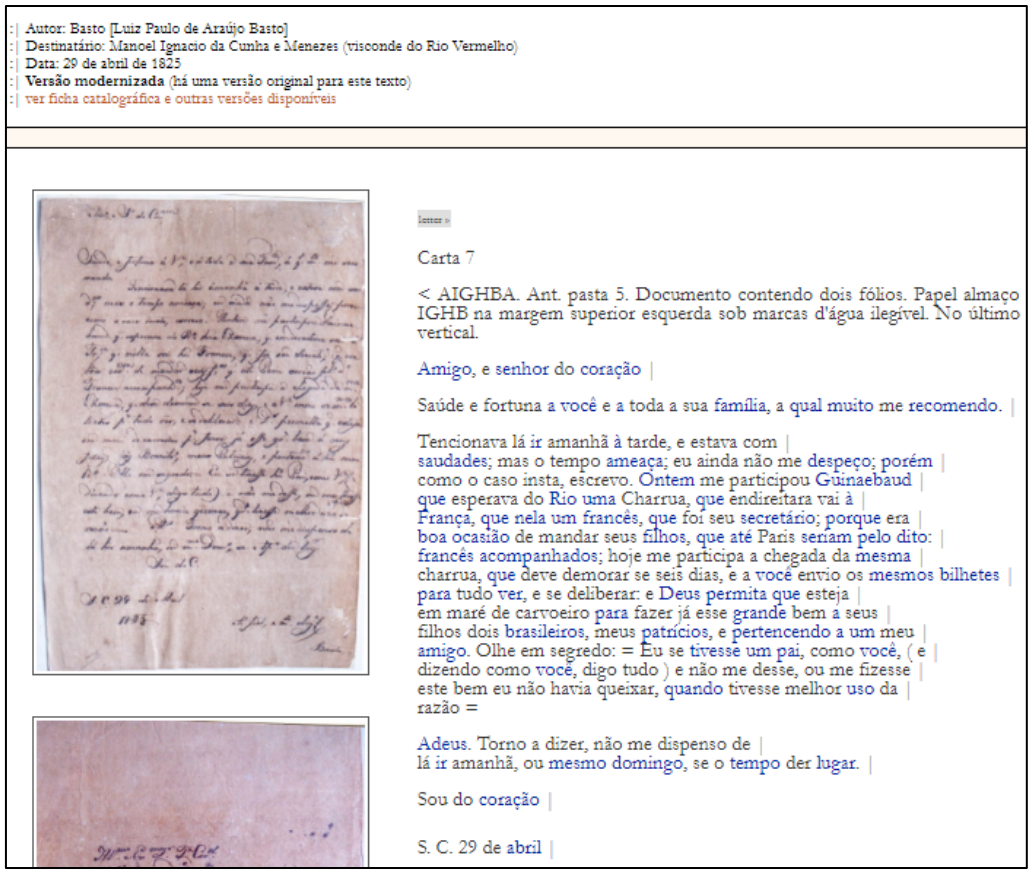

Figura 4- Edição modernizada final
Fonte: CE-DOHS

Para cada texto editado, são processados os metadados, que reúnem informações diversas, tanto sobre dados do documento (autor, conteúdo, destinatário, local, data do documento, referência, fonte, gênero do autor, gênero do documento) como sobre dados do processamento (revisão final da edição XML, primeira revisão da edição XML, edição semidiplomática, revisão da edição semidiplomática, edição XML, número de palavras).

Nos bancos digitais, as edições modernizadas vêm ganhando anotações morfológica e sintática. A anotação morfossintática e a anotação sintática – feitas com o objetivo principal de possibilitar, de maneira ampla, a recuperação de informações filológicas e linguísticas dos documentos – são realizadas, de forma semiautomática: o programa computacional devolve ao pesquisador, de forma automática, o texto etiquetado, que pode apresentar erros de anotação, os quais devem ser corrigidos pelo linguista, de modo manual. Hoje, o CE-DOHS disponibiliza seis acervos anotados, possibilitando a busca automática de dados, para estudo da história do português brasileiro; são eles: Acervo Cartas para Vários Destinatários; Acervo Cartas para

Severino Vieira, Governador da Bahia; Acervo Cartas para Cícero Dantas Martins, Barão de Jeremoabo; Acervo Correspondências Amigas; Acervo Cartas em Sisal. Há outros trabalhos de anotação em andamento, devendo ser disponibilizados dentro de alguns meses.

3- O CE-DOHS em etapas

O projeto CE-DOHS tem 2 etapas: etapa 1, de 1750 a 2000, caracterizada pelo multilinguismo localizado, cujos documentos permitem estudar a história do português brasileiro culto, semiculto e popular nesse contexto, e etapa 2, de 1500 a 1750, caracterizada pelo multilinguismo generalizado, cujos documentos permitem estudar a gestação do português brasileiro culto e do português brasileiro popular (português geral brasileiro, com história de contato com línguas africanas e indígenas) (MATTOS E SILVA, 2004; LUCCHESI, 2017).

São esses os subprojetos da etapa 1, que começou a ser executada em 2012 e foi encerrada em 2018:

a) Elaboração de ferramentas computacionais para construção e uso do CE-DOHS.
b) Aplicação de técnicas de anotação linguística e web-semântica no CE-DOHS.
c) Acervos de cartas de português brasileiro culto, semiculto e popular (séculos XIX-XX).
d) Cartas escritas por mãos "cândidas": o caso dos inábeis (século XX).
e) Corpora orais de português brasileiro culto e popular (Século XX).

A etapa 2, que começa a ser executada em 2019, enfrenta a raridade das fontes: são raros os textos escritos por grupos nascidos no Brasil, sobretudo de índios e negros, etnias que não tiveram acesso à escola (as fontes para o estudo da história linguística das classes dominantes são mais generosas); o projeto, todavia, tem pequenos acervos desse período e bastante significativos, em breve disponibilizados na Plataforma. São esses os subprojetos da fase 2:

a) Um *corpus* para os seiscentos (a partir de 1617). Documentos escritos por brasileiros: família Vieira Ravasco e outros contemporâneos.
b) Um *corpus* raro: escrito por indígenas integrados, mamelucos, pretos, pardos e brancos pobres (anos finais do século XVII).
c) Documentos da Feira do Capuame (1729-1830).
d) Inserção do indígena no mundo da escrita: dos aldeados, dos tempos de guerras e no âmbito da Reforma Pombalina (séculos XVII e XVIII).
e) Cartas e Atas produzidas por homens bons da Câmara de Salvador (a partir do século XVII).
f) Recuando ao Século XVIII: os livros do Sobrado da Fazenda do Campo Seco (1755-1800).

A metodologia utilizada no controle de aspectos sócio-históricos é a Teoria da Variação Linguística (LABOV, 1994), com aplicação para textos escritos, na chamada Linguística Histórica Sócio-Histórica (MATTOS E SILVA, 2008). Consideram-se as causas que apresentam impacto no processo de mudança do ponto de vista da Linguística Diacrônica, na visão da Gramática Gerativa Chomskiana (CHOMSKY, 1986).

3.1- Os documentos da Etapa 1

Nesses oito anos de execução, o projeto CE-DOHS formou um banco de mais de um milhão de palavras, com textos manuscritos, sobretudo, mas também impressos e orais; trata-se de um material representativo de variedades diacrônicas do português brasileiro, de diferentes regiões do país e de graus de escolaridade distintos, que atende não somente a pesquisadores interessados em análises de aspectos linguísticos, mas em aspectos da difusão da escrita, da leitura, das transmissões textuais, históricos, políticos, econômico-sociais, entre outros.

A documentação do banco representa, em sua maior parte, a região semiárida da Bahia. São diversos acervos de cartas manuscritas; 1084 cartas particulares (1808-2000), escritas por 422 remetentes (nascidos entre 1724 e 1980), extraída a maior parte de Carneiro (2005)[ix].

Esses acervos representam, conforme sugestão de Mattos e Silva (2001), as normas vernáculas e as normas cultas, de forma seriada, oferecendo um painel dos modelos de escrita e uma amostra da língua do período, em um *continuum*: documentos que expressam mais claramente a fala, em que há praticamente uma transposição da fala para a escrita (os mais populares) e documentos em que um modelo de escrita bloqueia a língua falada (os mais formais, produzidos pelos cultos). Essa documentação oferece indícios para o acesso às origens do PB, popular e culto, no período colonial.

O corpus em questão consiste em um material seguro, atendendo à proposta de Petrucci (2003, p. 7-8), para quem, para qualquer tempo histórico, quem trabalha com a Cultura Escrita deve responder a um conjunto mínimo de questões:

i. ¿*Qué*? En qué consiste el texto escrito, qué hace falta transferir al código gráfico habitual para nosotros, mediante la doble operación de lectura y transcripción;
ii. ¿*Cuándo*? Época en que el texto en sí fue escrito en el testimonio que estamos estudiando;
iii. ¿*Dónde*? Zona o lugar en que se llevó a cabo la obra de transcripción;
iv. ¿*Cómo*? Com qué técnicas, com qué instrumentos, sobre qué materiales, según qué modelos fue escrito ese texto;
v. ¿*Quién lo realizo*? A qué ambiente sociocultural pertenecía el ejecutor y cuál era en su tiempo y ambiente la difusión social de la escritura.
vi. ¿*Para qué fue escrito ese texto?* Cuál era la finalidad específica de ese testimonio en particular y, además, cuál podia ser en su época y en su lugar de producción la finalidad ideológica y social de escritura.

São os seguintes os acervos de cartas disponíveis no banco CE-DOHS, em edição semidiplomática, segundo as normas de transcrição do PHPB, definidas no II Seminário para a História do Português Brasileiro, em Campos do Jordão, em 1998:

a) Acervo Cartas para Vários Destinatários (1809-1904).
b) Acervo Cartas para Cícero Dantas Martins, Barão de Jeremoabo (1880-1903).
c) Acervo Cartas para Severino Vieira, Governador da Bahia (1901-1902).
d) Acervo Cartas para Dantas Jr. (1902-1962).

e) Acervo Cartas em Sisal, Riachão do Jacuípe, Conceição do Coité e Ichu (1906-2000)[x].
f) Acervo Cartas Baianas (1911-1958).
g) Acervo Cartas Particulares da Família Freire (1937-1942).
h) Acervo Cartas Particulares da Família Soledade (1948-1951).
i) Acervo da Família Oliveira (1962-1973).
j) Acervo Correspondências Amigas (1980-1993).

Há, ainda, editados anúncios e cartas impressas, de leitores e redatores (CARNEIRO; OLIVEIRA, 2012). Em linguagem xml, encontram-se também amostras orais.

4- Contribuições ao PHPB: contando a história com um milhão de palavras

O PHPB, já referido na Introdução deste trabalho, é coordenado pelo professor doutor Ataliba de Castilho, da Universidade de São Paulo (USP), desde 1997, organizando-se em equipes regionais pelo país, uma delas a equipe baiana.

Participam da equipe baiana do PHPB a Universidade Federal da Bahia (UFBA), a Universidade Estadual do Sudoeste da Bahia (UESB) e a Universidade Estadual de Feira de Santana (UEFS).

As contribuições da UEFS ao projeto nacional são muito significativas, nas três frentes de investigação propostas (LOBO; CARNEIRO, 2019):

a) um *campo histórico-filológico*, visando à constituição de *corpora* diacrônicos de documentos de natureza vária, escritos no Brasil, a partir do século XVI;
b) um *campo gramatical*, visando ao estudo de mudanças linguísticas depreendidas na análise dos *corpora* constituídos, e
c) um *campo de história social linguística*, visando à reconstrução mais ampla da história social linguística do Brasil e, em particular, do português brasileiro.

No que diz respeito à constituição de *corpora* diacrônicos, os projetos A Língua Portuguesa no Semiárido Baiano, Vozes do Sertão em dados e CE-DOHS, da UEFS, destacam-se no PHPB-Bahia, apresentando material criteriosamente transcrito e editado,

somando mais de um milhão de palavras. É uma grande contribuição aos estudos do português brasileiro, em sua realidade plural e polarizada entre normas vernáculas e normas cultas (LUCCHESI, 2015); de forma mais específica, ao estudo do processo de formação da língua portuguesa no espaço do semiárido baiano. Esse material está também disponível na Plataforma do *Corpora* Bahia[xi], coordenada pelas professoras Zenaide de Oliveira Novais Carneiro e Mariana Fagundes de Oliveira Lacerda.

5- Considerações finais

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.

O banco CE-DOHS veio somar-se aos *corpora* eletrônicos constituídos fundamentalmente para análises linguísticas, dos quais o Brown Corpus é o primeiro representante. Trata-se de um trabalho valioso essa formação de banco de dados nas plataformas digitais, para os estudos linguísticos de maneira geral, especialmente, no que diz respeito ao CE-DOHS – considerando as perguntas sociohistóricas que embasaram sua constituição –, para os estudos da formação do português brasileiro, na área da Linguística histórica.

Vencida a primeira etapa do projeto, aqui descrita em linhas breves, o banco CE-DOHS será ampliado, nos próximos 4 anos, alcançando, aproximadamente, três milhões de palavras, com documentos do século XVI a meados do século XVIII. Cada vez um *corpus* melhor para estudo da história da língua portuguesa na América.

6- 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.

BARBOSA, A. G.. A plataforma de *corpora* do PHPB: uma apresentação *ad infinitum*. In: CARNEIRO, Z. de O. N. (Org.). *Cartas brasileiras (1809-2000)*: coletânea de fontes para o estudo do português. Feira de Santana: UEFS, 2011.

BRITO, G. S. *Do texto ao documento digital*: transposição fotográfica de documentos manuscritos históricos para formação de corpora linguísticos eletrônicos. Dissertação (Mestrado em Linguística) – Programa de Pós Graduação Linguística da Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2015.

CARNEIRO, Z. O. N. *Cartas Brasileiras*: um estudo linguístico-filológico. Tese (Doutorado em Estudos Linguísticos) – Programa de Pós-Graduação em Linguística, Universidade Estadual de Campinas, Campinas, 2005.

CARNEIRO, Z. de O. N., OLIVEIRA M. F. de. *Publica-se em Feira de Santana*: das cartas de leitores e redatores e dos anúncios em O Progresso e Na Folha do Norte (1901-2006). Feira de Santana: UEFS, 2012.

CE-DOHS: Corpus eletrônico de documentos históricos do sertão. Disponível em: [www.uefs.br/cedohs]. 2011.

CHOMSKY, N. *Knowledge of Language*: its nature origin, and use. New York: Praeger, 1986.

CORPUS Histórico do Português Tycho Brahe. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/>

CRANE, G. (et al.). *ePhilology*: when the boocks talk to their readers. Blackwell Campanion to Digital Literary Studies. Oxford: Blackwell, 2008.

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.

LABOV, W.. *Principles of Linguistic Change: internal factors*. Oxford: Blackwell, 1994.

LACERDA, M. F. O; CARNEIRO, Z. O. N. Edição filológica e edição digital do Livro do Gado e do Livro de Razão do Arquivo do Sobrado do Brejo (Bahia setecentista e oitocentista). In: Revista Labor Histórico. 2016, n. 2, p. 151-163. Disponível em: <https://revistas.ufrj.br/index.php/lh/article/download/4814/3522>. Acesso em: 3 mai 2019.

LOBO, T.; CARNEIRO, Z. N.. Reflexões sobre a constituição e análise de *corpora* linguísticos históricos e sobre a identificação de perfis sociais de redatores do passado. In: CASTILHO, A. T. de. (Coord.). *História do português brasileiro*: *corpus* diacrônico do português brasileiro. São Paulo: Contexto, 2019.

LUCCHESI, D. *Língua e sociedade partidas*: a polarização sociolinguística do Brasil. São Paulo: Contexto, 2015.

LUCCHESI, D.. *A periodização da história sociolinguística do Brasil***.** In: Revista DELTA. 2017, vol.33, n.2, p.347-382. Disponível em: <http://dx.doi.org/10.1590/0102-445067529349614964>. Acesso em 4 mai 2019.

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

MATTOS E SILVA, R. V.. *Caminhos da Linguística Histórica: ouvir o inaudível*. São Paulo: Parábola Editorial, 2008. p.7-26.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. P. F. E-Dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. In: SHEPHERD T.; SARDINHA T. B.; PINTO M. V. (Org.). *Caminhos da linguística de corpus*. Campinas: Mercado de Letras, 2010.

Penn Helsinki Parsed Corpus of Middle English. Disponível em: <http://www.ling.upenn.edu/hist-corpora/>

PETRUCCI, A. *La ciencia de la escritura: primera lección de paleografía*. Buenos Aires: Fondo de Cultura Económica de Argentina, 2003.

Plataforma de Corpora do PHPB. Disponível em: <https://sites.google.com/site/corporaphpb>

Post Scriptum: archivo digital de escritura cotidiana em Portugal e Espanha na Época Moderna. Disponível em: <http://www.clul.ul.pt/pt/recursos/462-post-scriptum-home>

SANTIAGO. H. da S.. *A escrita por mãos inábeis*: uma proposta de caracterização. 2019. 2v. 722 f. Tese (Doutorado em Língua e Cultura) – Programa de Pós-graduação em Língua e Cultura, Universidade Federal da Bahia, Salvador, 2019.

SANTIAGO. H. da S.. *Um estudo do português popular brasileiro em cartas pessoais de "mãos cândidas" do sertão baiano*. 2012. 2v. 421 f. Dissertação (Mestrado em Estudos Linguísticos) – Programa de Pós-graduação em Estudos Linguísticos, Universidade Estadual de Feira de Santana, Feira de Santana, 2012.

SANTOS, E. B. *O Livro do Gado do Brejo do Campo Seco (Bahia)*: edição semidiplomática e descrição de aspectos grafo-fonéticos. Dissertação (Mestrado em Estudos Linguísticos) - Programa de Pós-graduação em Estudos Linguísticos, Universidade Estadual de Feira de Santana, Feira de Santana, 2019.

SANTOS, J. V.; BRITO, Giovane Santos. Fotografia técnica de documentos para a formação de corpora digitais eletrônicos: o método desenvolvido no Lapelinc. *Letras & Letras*, v. 30, n. 2, p. 421, 30 jul./dez. 2014.

SCHREIBMAN, S. (et al.). *A Companion to Digital Humanities*. Oxford: Blackwell, 2004.

Vozes do sertão em dados: história, povos e formação do português brasileiro. Disponível em: <www.uefs.br/nelp>. 2011

Sobre as autoras

Zenaide de Oliveira Novais Carneiro. Possui Graduação em Letras com Inglês pela Universidade Estadual de Feira de Santana (1988), Mestrado em Letras e Linguística (1996) pela Universidade Federal da Bahia, Doutorado em Linguística (2005) e Pós-Doutorado em Linguística de *Corpus* (2010) pela Universidade Estadual de Campinas. Atualmente é Professora Plena da Universidade Estadual de Feira de Santana, onde coordena o projeto CE-DOHS – Corpus Eletrônico de Documentos Históricos do Sertão (FAPESB), disponível em <www.uefs.br/cedohs>, e atua como Membro Permanente no Programa de Pós-Graduação em Estudos Linguísticos (PPGEL) e no Mestrado Profissional em Letras (PROFLETRAS). É Membro Colaborador do Programa de Pós-Graduação em Língua e Cultura da Universidade Federal da Bahia (PPGLinC), atuando como co-coordenadora do Banco Informatizado de Textos do Programa para a História da Língua Portuguesa (BIT-PROHPOR/UFBA), disponível em <http://www.prohpor.org/bit-prohpor>. Integra também a equipe de pesquisadores do Projeto Nacional para a História do Português Brasileiro (PHPB), onde é co-coordenadora da Plataforma de Corpora Bahia, disponível em <https://sites.google.com/site/corporaphpbba/?pli=1>

Mariana Fagundes de Oliveira Lacerda. Possui Graduação em Letras Vernáculas (2002) pela Universidade Federal da Bahia (UFBA), Mestrado (2005) e Doutorado (2009) em Linguística pela mesma instituição, com estágio de doutoramento no Centro Linguístico da Universidade de Lisboa, financiado pela CAPES. Na Universidade Estadual de Feira de Santana (UEFS), onde é Professora Titular da subárea de

Linguística Histórica e Membro Permanente do Programa de Pós-Graduação em Estudos Linguísticos (PPGEL) e do Mestrado Profissional em Letras (PROFLETRAS), coordena o Núcleo de Estudos de Língua Portuguesa (NELP) e é co-coordenadora do projeto CE-DOHS - Corpus Eletrônico de documentos Históricos do Sertão (FAPESB), disponível em <www.uefs.br/cedohs>. Integra, ainda, a equipe de pesquisadores do Projeto Nacional para a História do Português Brasileiro (PHPB), co-coordenando a Plataforma de Corpora Bahia, disponível em <https://sites.google.com/site/corporaphpbba/?pli=1>, e a equipe do Programa para a História da Língua Portuguesa (PROHPOR-UFBA), onde é co-coordenadora do Banco Informatizado de textos (BIT), disponível em <http://www.prohpor.org/bit-prohpor>.

Notas

[i] Página oficial: Página oficial: <http://www.uefs.br/cedohs/>.

[ii] Página oficial: < http://www.tycho.iel.unicamp.br/corpus/>.

[iii] Processo FAPESB 5566/2010. CONSEPE 202/2010.

[iv] Processo CNPq. 401433/2009-9. CONSEPE 102/2009).

[v] Ver Lacerda e Carneiro (2016).

[vi]Para maiores detalhes sobre a aplicação do método Lapelinc no *Livro do Gado* e a edição semidiplomática do documento, ver dissertação de Santos (2019). A edição do *Livro de Razão*, com detalhes sobre a aplicação do método Lapelinc ao documento, está sendo realizada por Silva, no âmbito do seu curso de doutorado, na Universidade Federal da Bahia (UFBA).

[vii] Por ocasião do Workshop Construction and use of large annotated corpora, realizado na UNICAMP, em 2013, pela equipe do projeto Corpus Histórico do Português Tycho Brahe, do qual pesquisadores de diversos projetos de *corpora* eletrônicos participaram – entre eles o CE-DOHS –, reafirmou-se a importância de esses projetos seguirem os mesmos padrões de edição digital e de anotação morfossintática, tendo em vista a praticidade do trabalho e a integração dos pesquisadores.

[viii] Para maiores detalhes sobre o processo de edição, usando o eDictor, a partir de textos-fontes em edição semidiplomática, consultar a página do banco: <http://www.uefs.br/cedohs/>.

[ix] Conferir também a Coleção Cartas Brasileiras (CARNEIRO, 2011).

[x] Conferir Santiago (2012; 2019).

[xi] Página oficial: <https://sites.google.com/site/corporaphpb/>.