ISSN 1415 - 8973

Corpus eletrônico de documentos históricos do sertão: as cartas de *inábeis*

Electronic corpus of historical documents of the sertão: the letters of awkwardness

Mariana Fagundes de Oliveira Lacerda[*]
Universidade Estadual de Feira de Santana, Feira de Santana, Bahia, Brasil

Zenaide de Oliveira Novais Carneiro[**]
Universidade Estadual de Feira de Santana, Feira de Santana, Bahia, Brasil

Huda da Silva Santiago[***]
Universidade Federal da Bahia, Salvador, Bahia, Brasil

Resumo: O projeto CE-DOHS – Corpus Eletrônico de Documentos Históricos do Sertão, que integra o Núcleo de Estudos Interdisciplinares em Humanidades Digitais (neiHD), da Universidade Estadual de Feira de Santana (UEFS), tem por objetivo realizar a edição digital de textos do Banco DOHS – Documentos Históricos do Sertão, do projeto Vozes do Sertão em Dados: história, povos e formação do português brasileiro, um dos projetos do Núcleo de Estudos em Língua Portuguesa (NELP), da UEFS, bem como sua anotação morfológica e sintática, elaborando um *corpus* diacrônico anotado que sirva como recurso eletrônico para o estudo linguístico do português brasileiro. A maior parte dos documentos do DOHS, datados e localizados – que hoje se encontram também em versão digital no CE-DOHS – são cartas manuscritas, dos séculos XIX e XX (1084 cartas, 422 remetentes), editadas sobretudo por Carneiro (2005), que investiu na busca e na organização de acervos documentais que pudessem contribuir para o processo de reconstrução sócio-histórica do português brasileiro, tanto da vertente popular como da vertente culta – especialmente do português no interior da Bahia. Além dos acervos constituídos por documentação epistolar, há também livros manuscritos, além de textos impressos e textos orais. O material disponível no Banco atende, entretanto, 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. Neste trabalho, a ênfase é para o acervo Cartas em Sisal, constituído por 91 cartas de inábeis, editadas por Santiago (2012), disponíveis no CE-DOHS, nas versões semidiplomática e modernizada; são cartas pessoais, escritas ao longo do século XX, por 43 sertanejos oriundos do semiárido baiano. Esse acervo tem especial relevância para a Linguística Histórica, por ser uma amostra representativa da escrita por *mãos inábeis* – termo consagrado pela tradição paleográfica –, considerando-se a dificuldade de encontrar textos que refletem a escrita cotidiana, vernacular, produtos de indivíduos com baixo nível de letramento.

Palavras chave: Edição digital. Português brasileiro. Corpus diacrônico. Cartas de inábeis.

Abstract: The CE-DOHS project - Electronic Corpus of Historical Documents of the Sertão, which integrates the Center for Interdisciplinary Studies in Digital Humanities (neiHD), of the State University of Feira de Santana (UEFS), aims to carry out the digital edition of texts Of the DOHS - Sertão Historical Documents, of the Vozes do Sertão project in Data: history, peoples and formation of Brazilian Portuguese, one of the projects of the Núcleo de Estudos em Língua Portuguesa (NELP), UEFS, as well as its morphological and syntactic annotation , Elaborating an annotated diachronic corpus that serves as an electronic resource for the linguistic study of Brazilian Portuguese. Most DOHS documents, dated and localized - which are now also in digital version in CE-DOHS - are handwritten letters from the 19th and 20th centuries (1084 letters, 422 senders), edited mainly by Carneiro (2005), Who invested in the search and organization of documentary collections that could contribute to the process of socio-historical reconstruction of Brazilian Portuguese, both popular and cultured - especially Portuguese in the interior of Bahia. In addition to collections consisting of epistolary documentation, there are also handwritten books, as well as printed texts and oral texts. The material available

[*] Universidade Estadual de Feira de Santana, Professora Adjunta do Departamento de Letras e Artes. E-mail: marianafag@gmail.com.
[**] Universidade Estadual de Feira de Santana, Professora Plena do Departamento de Letras e Artes. E-mail: zenaide.novais@gmail.com.
[***] Universidade Federal da Bahia, doutoranda pelo Programa de Pós-Graduação em Língua e Cultura. E-mail: huda.santiago@hotmail.com.

ISSN 1415 - 8973

at the Bank serves, however, not only researchers interested in analyzes of linguistic aspects, but in aspects of the diffusion of writing, reading, textual, historical, political, economic and social transmissions, among others. In this work, the emphasis is on the Letts in Sisal collection, consisting of 91 letters of awkwardness, edited by Santiago (2012), available in the CE-DOHS, in the semidiplomatic and modernized versions; Are personal letters, written throughout the 20th century, by 43 sertanejos from the semi-arid Bahia. This collection has a special relevance for Historical Linguistics, since it is a representative sample of the writing by unskilled hands - a term consecrated by the paleographic tradition -, considering the difficulty of finding texts that reflect everyday writing, vernacular, products of individuals with low level of literacy.

Keywords: Digital edition. Brazilian portuguese. Corpus diachronic. Letters of awkwardness.

1 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 se proceder a uma observação sistemática, e, na medida do 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).

Essa constituição tem em mente, por outro lado, a dificuldade, já destacada por Labov (1972), em relação aos dados para o estudo da língua no *tempo real*. Um obstáculo irrefutável, diante do qual só resta à ciência buscar contorná-lo, através da maximização dos recursos existentes, dos textos remanescentes escritos em fases pretéritas da língua.

Hoje, contando com melhores recursos tecnológicos, no universo das Humanidades Digitais, os bancos de textos disponibilizam não somente edições semidiplomáticas, em pdf, mas também edições digitais – a partir do estabelecimento de redes entre projetos que desenvolvem a Linguística de Corpus e a Linguística Computacional –, que servem como recurso eletrônico para estudos linguísticos, entre outros. 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)

Este trabalho discute o processo de constituição desses bancos de textos, no universo das Humanidades Digitais, apresentando a experiência na construção de *corpora* anotados e seus aspectos linguísticos e computacionais no âmbito do projeto CE-DOHS – Corpus Eletrônico de Documentos Históricos do Sertão (www.uefs.br/cedohs)[1].

[1] O projeto CE-DOHS – Corpus Eletrônico de Documentos Históricos do Sertão é financiado pela Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB)/5566-2010 e é desenvolvido atualmente no Núcleo de Estudos da Língua Portuguesa no Semiárido (NELP), no Departamento de Letras e Artes da Universidade Estadual de Feira de Santana (UEFS). Fazem parte da equipe de investigação Zenaide de Oliveira Novais Carneiro (coordenadora), Mariana Fagundes de Oliveira Lacerda (vice-coordenadora), Charlotte Galves-Chambelland, Huda da Silva Santiago. O projeto conta com inúmeros bolsistas de Iniciação Científica, de Mestrado e de Doutorado, que colaboram nas diferentes etapas do processo de constituição do banco de textos.

ISSN 1415 - 8973

O CE-DOHS, com o objetivo de contribuir com o Projeto Para a História do Português Brasileiro (PHPB), em diferentes perspectivas teóricas e por meio de parceria tecnológica com o projeto Corpus Histórico do Português Tycho Brahe (www.tycho.iel.unicamp.br), traz um conjunto de documentos originados sobretudo da grande área do semiárido baiano, editados em linguagem XML, com o uso do eDictor, desenvolvido por Paixão de Sousa, Kepler e Faria (2007; 2010a), um editor de textos especialmente voltado ao trabalho filológico e à análise linguística automática, o qual 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).

Como resultado da primeira fase de pesquisa, o projeto CE-DOHS já disponibiliza diversos acervos, sobretudo de cartas manuscritas, organizando-as por grau de escolaridade e por grau de habilidade com a escrita; são 1084 cartas particulares (1808-2000), num total de 350.850 palavras**,** escritas por 422 remetentes (nascidos entre 1724 e 1980), extraída a maior parte de Carneiro *et al* (2011).

O Quadro 1, a seguir, mostra a distribuição, por décadas, da documentação epistolar do CE-DOHS:

Quadro 1: Distribuição, por década, da documentação epistolar do CE-DOHS.

Década	***Subcorpus* de cartas**	**Quant.**
1800	Cartas para Vários Destinatários	1
1810	Cartas para Vários Destinatários	4
1820	Cartas para Vários Destinatários	9
1830	Cartas para Vários Destinatários	2
1840	Cartas para Vários Destinatários	2
1850	Cartas para Vários Destinatários	7
1860	Cartas para Vários Destinatários	102
1870	Cartas para Vários Destinatários	34
1880	Cartas para Vários Destinatários, Cartas para Cícero Dantas Martins, Barão de Jeremoabo	24
1890	Cartas para Vários Destinatários, Cartas para Cícero Dantas Martins, Barão de Jeremoabo	164
1900[2]	Cartas para Vários Destinatários, Cartas para Cícero Dantas Martins, Barão de Jeremoabo, Cartas para Severino Vieira, Governador da Bahia, Cartas do Acervo Dantas Jr.	158
Total de cartas do século XIX		**502**
1910	Cartas Baianas: o acervo de João da Costa Pinto Victoria, Cartas do Acervo Dantas Jr., Cartas em Sisal	20
1920	Cartas Baianas: o acervo de João da Costa Pinto Victoria, Cartas do Acervo Dantas Jr.	34
1930	Cartas Baianas: o acervo de João da Costa Pinto Victoria, Cartas do Acervo Dantas Jr., Acervo da família Freire	122
1940	Cartas Baianas: o acervo de João da Costa Pinto Victoria, Cartas do Acervo Dantas Jr., Acervo da família Freire, Acervo particular da Família Soledade	120
1950	Cartas Baianas: o acervo de João da Costa Pinto Victoria, Cartas do Acervo Dantas Jr., Cartas em Sisal, Acervo particular da Família Soledade	96
1960	Cartas do Acervo Dantas Jr., Cartas em Sisal, Acervo da família Oliveira	61
1970	Cartas em Sisal, Acervo da Família Oliveira	30

[2] 1900, data da última carta, considerada do século XIX, porque todos os remetentes escreveram no século XIX. No acervo Cartas para Vários Destinatários, são 18 cartas de 1809-1845, 105 cartas entre 1851-1870, 53 cartas de 1871-1889, e apenas 6 cartas de 1900-1904. No acervo Cartas para o Barão de Jeremoabo, são 149 cartas escritas de 1880-1899, e 38 cartas escritas de 1900-1903. E, no acervo Cartas para Severino, são 41 cartas de 1901 e 58 cartas de 1902. A essas foram acrescentadas 2 cartas do Dantas Jr., escritas na primeira década. Sua correspondência é toda do século XX.

ISSN 1415 - 8973

1980	Cartas em Sisal, Correspondências Amigas, o acervo de Valente, Bahia	69
1990	Cartas em Sisal, Correspondências Amigas, o acervo de Valente, Bahia	17
2000	Cartas em Sisal	1
Sem data no século XX	Cartas do Acervo Dantas Jr., Correspondências Amigas, o acervo de Valente, Bahia e do Acervo da família Freire	12
Total de cartas do século XX		**582**
Total Geral de Cartas		**1084**

Na segunda fase do projeto, que está em andamento, o número de documentos tem sido ampliado, tanto manuscritos[3] como impressos, com inserção, ainda, de amostras de fala, organizadas, no Banco, por comunidade, por tipo de contato linguístico e por vertente (popular e culta). Essa ampliação do *corpus*

> [...] 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. (BACELAR DO NASCIMENTO, 2004, p. 1).

Todo material do CE-DOHS – representativo de variedades diacrônicas do português brasileiro (PB), de diferentes regiões do país e de graus de escolaridade distintos – está sendo preparado para a anotação morfossintática, que manterá a maioria das características do padrão de anotação existente e permitirá a busca automática de dados, o que facilitará o estudo linguístico dos acervos, no que consiste o principal objetivo do CE-DOHS. O material disponível no Banco atende, entretanto, 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.

Neste texto, apresentamos o acervo Cartas em sisal, Riachão do Jacuípe, Conceição do Coité e Ichu (1906-2000), as cartas de inábeis, editadas por Santiago (2012), disponíveis nas versões semidiplomática e modernizada. A amostra é constituída por 91 cartas pessoais escritas ao longo do século XX, por 43 sertanejos oriundos do semiárido baiano. Esse é um *corpus* representativo, seja porque as cartas foram trocadas em relação de simetria entre redatores que fazem parte de um contexto sociocultural semelhante, seja pelo grau de transparência aos usos vernáculos que apresentam, pois são textos próximos de uma escrita cotidiana. Os redatores possuem pouca escolarização, já que a maioria teve contato com as primeiras letras em espaços extraescolares.

2 CONSTITUIÇÃO DO BANCO

O CE-DOHS apresenta a versão digital de documentos em edição semidiplomática que compõem o banco DOHS – Documentos Históricos do Sertão, do projeto Vozes do Sertão em Dados: história, povos e formação do português brasileiro[4], reunidos com o

[3] Em breve, será disponibilizado o acervo composto por 158 cartas particulares do Recôncavo da Bahia (1818-1886), escritas por brasileiros e portugueses radicados na Bahia, editadas em versão semidiplomática de forma primorosa, por Lobo (2001).

[4] O projeto Vozes do Sertão em Dados é financiado pelo Conselho Nacional de Pesquisa (CNPq)/ 401433-2009 e executado no Núcleo de Estudos em Língua Portuguesa (NELP), pensado e implementado por Norma Lucia Fernandes de Almeida e Zenaide de Oliveira Novais Carneiro, do Departamento de Letras e Artes da UEFS. Relaciona-se com o Programa para a História da Língua Portuguesa (PROHPOR) – fundado por Rosa Virgínia Mattos e Silva, na Universidade Federal da Bahia (UFBA) –, especificamente em seu arco temporal da história do PB, e resulta de desdobramentos de uma agenda de trabalho iniciada por Ilza Ribeiro, Norma Lucia de Almeida e Zenaide Carneiro, na UEFS, em 1997, na qual se previa a edição de documentos diversos, no âmbito do projeto Contribuições para a Constituição de um Banco de Textos

ISSN 1415 - 8973

objetivo de estudar o processo de formação histórica do PB, especialmente na região do semiárido baiano.

Os documentos do Vozes – que mantém parceria com o PHPB, por meio da prospecção e edição de documentos, da formação de *corpora* representativos de demandas histórico-sociais da região semiárida baiana, com repercussões sobre o processo de formação histórica do PB, com amplo contato linguístico de populações de origem portuguesa, indígena e africana, bem como com projetos temáticos de análise linguística – vêm servindo de base para a composição de uma *Plataforma de Corpora do PHPB* (https://sites.google.com/site/corporaphpb), a cargo de Afrânio Barbosa, da Universidade Federal do Rio de Janeiro (UFRJ), e de Marcelo Módulo, da Universidade de São Paulo (USP).

A maior parte dos documentos do DOHS, datados e localizados – que hoje se encontram também em versão digital no CE-DOHS – são cartas manuscritas, dos séculos XIX e XX (1084 cartas, 422 remetentes), editadas sobretudo por Carneiro (2005), que investiu na busca e na organização de acervos documentais que pudessem contribuir para o processo de reconstrução sócio-histórica do PB, em um trabalho de investigação grandioso, percorrendo diversos arquivos, e publicadas em 2011, na obra, com três volumes, organizada pela mesma autora, intitulada *Cartas brasileiras: coletânea de fontes para o estudo do português*.

Em 98% dos acervos, é possível determinar, *onde, quando, por quem* e *para quem* as cartas foram escritas. Os acervos, desse modo, estão dentro do que se espera para um *corpus* 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, a saber:

> 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.

Os acervos de cartas presentes no CE-DOHS 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, formado via duas macro-origens, a culta e a popular, no período colonial. Nos termos colocados por Mattos e Silva (2001, p. 298-299),

> a) O português europeu na sua dialetação diatópica, diastrática, que teria ao longo do período colonial um contingente de 30% da população brasileira; seria esse português europeu, base histórica do português culto brasileiro que começaria a elaborar-se a partir da segunda metade do século XVIII;

e de um Banco de Dados para o Estudo da História do Português do Brasil, do séc. XVII ao XX. Integra o PHPB, coordenado por Ataliba de Castilho, da Universidade de São Paulo (USP), e da Universidade Estadual de Campinas (UNICAMP), via equipe baiana, coordenada por Tânia Lobo, da UFBA.

ISSN 1415 - 8973

b) As línguas gerais indígenas, que, plurais e dialetalizadas, poderiam até confundir-se com o português geral brasileiro nas áreas geográficas delimitáveis em que se difundiram;
c) O português geral brasileiro, antecedente histórico do português popular brasileiro que, adquirido na oralidade e em situações de aquisição imperfeita, é difundido pelo geral do Brasil, sobretudo, pela maciça presença africana e dos afro-descendentes que perfizeram uma média de mais de 60% da população por todo o período colonial.

Para além dos acervos de cartas, o CE-DOHS trabalha com livros de fazenda manuscritos, dos séculos XVIII e XIX, textos impressos, dos séculos XX e XXI[5], e amostras de fala do século XX[6]. Todo esse material foi selecionado tendo em vista os interesses de investigação do projeto: textos diacrônicos, manuscritos, impressos e orais, nas vertentes popular e culta[7], para constituição de banco de dados para estudo da história do PB.

A seguir, estão apresentados os acervos de cartas – que representam a parte mais significativa do Banco em questão –, organizados por local de nascimento dos remetentes e por local de escrita das cartas:

Acervo Cartas para Vários Destinatários (1809-1904[8]):

Trata-se de 208 cartas; 38 baianos dentre 114 remetentes (111 homens e 3 mulheres), oriundos, em sua maioria, da classe alta e letrada. Essas cartas foram dirigidas para vários destinatários. Essa amostra permite observar a formação de corpora com textos escritos por brasileiros cultos nascidos entre 1724 e 1880. Os remetentes pertencem, em sua maioria, à classe mais alta ou à classe imediatamente inferior, ou seja, constituem uma amostra de uma classe alta e letrada.

Dentre essas cartas, 124 foram escritas no Brasil, principalmente no Rio de Janeiro, possivelmente na capital (52 cartas) e mais 2 de Petrópolis, seguidas das cartas escritas na Bahia, 51, sendo 22 cartas originárias, provavelmente, da capital da província, Salvador. As demais cartas da Bahia provêm, em sua maioria, do interior. São basicamente as cartas enviadas por parentes, amigos e correligionários ao coronel Exupério Canguçu, provenientes da região da Chapada Diamantina e da Serra Geral, área de atuação desse coronel, assim distribuídas: além de uma carta do recôncavo, uma carta proveniente de uma localidade mineira onde o coronel possuía terras e uma carta da Bahia (Salvador), do seu sobrinho Marcolino de Moura e Albuquerque; as demais cartas, 39 no total, embora não tenham seus locais de origem especificados, parecem ter sido escritas no Brasil. As cartas do exterior são ao todo 45, quase todas provenientes de remetentes brasileiros em trânsito, ou em exercício de missões no exterior, ou então lá residindo temporariamente. São 22 cartas da Europa e mais 23 cartas de localidades envolvidas na Guerra do Paraguai.

Acervo Cartas para Cícero Dantas Martins, Barão de Jeremoabo (1880-1903):

São 190 cartas; 43 baianos/43 remetentes (42 homens e 1 mulher). São de uma espécie de elite rural sertaneja, pouco letrada. Essas cartas oferecem uma amostragem de textos escritos por baianos do interior da Bahia, parte maior do que designamos de semicultos e

[5] Os textos impressos encontram-se na obra *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)*, organizada por Carneiro e Lacerda (2012).

[6] Esse *corpus* oral, representativo do português popular brasileiro da década de 90 do século XX, é produto do projeto A Língua Portuguesa no Semiárido Baiano, do NELP (www.uefs.br/nelp), e foi publicado, em 2008, por Almeida e Carneiro, na coleção *Amostras da Língua Falada no Semiárido Baiano*. A escolha desse tipo de *corpus* se justifica pelo fato de que, como afirma Mattos e Silva (2001), o estudo vertical das variantes populares do PB é uma vertente importante de pesquisa para a recuperação do *português popular brasileiro*, cujo antecedente histórico é o *português geral brasileiro*, constituído do encontro multilíngue da população indígena, do português e da população de origem africana.

[7] Sobre as vertentes popular e culta do PB ver Lucchesi (1994; 2001).

[8] A distribuição é a seguinte: (i)1809-1845: 18 cartas; 1851-1870: 105 cartas; 1871-1889: 53 cartas; 1900-1904: 6 cartas (consideradas do século XIX, porque são de remetentes que escrevem no XIX); (ii) 1880 -1899: 149 cartas e 1900-1903: 38 cartas.

ISSN 1415 - 8973

uma pequena parte denominada de semipopulares, composta, basicamente por vaqueiros, administradores das fazendas do Barão de Jeremoabo.

A partir da data de nascimento e do cruzamento com a data de escrita da carta, foi possível identificar a idade de parte dos remetentes, nascidos entre fins do século XVIII e o terceiro quartel do século XIX, com nacionalidade brasileira identificada ou inferida. A idade média, quando da escrita da carta, variou entre 13 e 65 anos.

Acervo Cartas para Severino Vieira, Governador da Bahia (1901-1902):

Trata-se de 102 cartas; 8 baianos/60 remetentes[9] (57 homens e 3 mulheres), de maioria letrada e citadina, compondo-se basicamente por remetentes cultos ou semicultos. Nesse grupo, a situação se inverte, e a maioria pode-se dizer que é da elite imediatamente inferior à primeira, mas, ainda, letrada e, sobretudo, citadina. São em 1901 (41 cartas) e em 1902 (58 cartas).

No século XIX, são 500 cartas manuscritas escritas entre 1809-1904 por indivíduos nascidos dos fins do século XVIII até o terceiro quartel do século XIX, cuja nacionalidade brasileira identificada ou inferida permite opor duas variantes distintas (Carneiro, 2005), as variantes 1 e 2.

> Variante 1: textos escritos por brasileiros cultos nascidos e/ou educados em regiões urbanas;
> Variante 2: textos escritos por brasileiros semi-cultos e não cultos nascidos/radicados no interior, especificamente da Bahia.

Acervo Cartas para Dantas Jr. (1902-1962):

São 242 cartas; 64 baianos/113 remetentes[10], constando apenas 7 mulheres. Parte dos remetentes desse acervo do século XX apresenta perfil próximo ao perfil dos remetentes das cartas do século XIX. A maior parte dos remetentes é formada por Bacharéis em Direito, com altos cargos no período republicano em que viveram.

Acervo Cartas em Sisal, Riachão do Jacuípe, Conceição do Coité e Ichu (1906-2000):

É um acervo composto por 91 cartas, escritas por 43 sertanejos do interior da Bahia, editadas por Santiago (2012). O Acervo Cartas em Sisal tem especial relevância para a Linguística Histórica, por ser uma amostra representativa da escrita por *mãos inábeis* – expressão usada por Marquilhas (2000), para designar os redatores estacionados em fase incipiente de aquisição da escrita –, considerando-se a dificuldade de se encontrar textos que refletem a escrita cotidiana, vernacular, produtos de indivíduos com baixo nível de letramento.

Acervo Cartas Baianas (1911-1958):

Compõe-se de 102 cartas; 5 baianas/5 remetentes. São cartas escritas por cinco mulheres semicultas e cultas, oriundas de Salvador (1), de Santo Amaro (3) e do Rio de Janeiro (1)[11]. As cartas provêm de diversas regiões da Bahia, sendo a maioria da Capital baiana, Salvador, e outras de vilarejos e fazendas. As cartas desse acervo foram escritas por

[9] Os demais são: 1 goiano, 1 mineiro, 1 paraibano, 1paraense, 1 pernambucano, 1 piauense, 8 cariocas, 1 potiguar e 1 sergipano, 2 brasileiros sem especificação de naturalidade e 16 brasileiros por inferência. E, ainda, 2 estrangeiros: John T. Lewis (carta, nº. 280) e M. Wicks (cartas nº. 292 e nº. 293). Há, ainda, outros 16 remetentes não identificados. São em 1901 (41 cartas) e em 1902 (58 cartas).

[10] Dentre os 113 remetentes identificados, apenas um apresenta outra nacionalidade, a portuguesa, o Padre Antonio da Costa Gaito, Bacharel em Ciências Sociais e Jurídicas (Universidade de Coimbra).

[11] Essas pertencentes a famílias com representatividade no Brasil Colônia, no Brasil Império e com significação também no Brasil República, das famílias Araújo Pinho, Argolo, Carvalho, Costa Pinto, Ferreira de Moura e Wanderley. As cartas foram trocadas entre familiares e tratam de assuntos pessoais e cotidianos.

ISSN 1415 - 8973

descendentes dos remetentes das Cartas para Vários Destinatários, acervo do século XIX. Representam uma amostra da escrita culta.

Acervo Cartas Particulares da Família Freire (1937-1942)[12]:

Trata-se de um pequeno conjunto formado por 17 correspondências (cartas, bilhetes e cartões manuscritos), trocadas, sobretudo, pelo casal Carlos Ribeiro Freire e Iracema Batista Chéquer Freire, nascidos em Jussiape e Araíba, cidades baianas. São Cartas representativas de normas de escolarizados do PB, editadas por Gandra (2010).

Acervo Cartas Particulares da Família Soledade (1948-1951)[13]:

Esse acervo, também editado por Gandra (2010), é formado por 100 cartas manuscritas (apenas uma é datilografada), trocadas entre o casal Otto Soledade Júnior e Renée da Silva Barros Soledade, nascidos em Salvador e em Ilheús. São documentos representativos do português culto.

Acervo Correspondências Amigas (1980-1993):

São 79 cartas; 31 baianos/38 remetentes[14]. Há, no conjunto de cartas manuscritas, 2 cartas datilografadas. São 38 remetentes, 31 do sexo feminino e 7 do sexo masculino, sendo 77 cartas e os 25 cartões destinados a Adelmário Carneiro Araújo, nascido em Valente, em 17 de outubro de 1959 e mais 2 cartas de Adelmário Carneiro Araújo para a Eliana de Oliveira Lima, mais tarde sua esposa, e a Regina Célia Siqueira dos Santos. O conjunto das correspondências, formado por cartas e cartões, foi numerado progressivamente de 1 a 104. A maior parte dos remetentes é nascida ou radicada no interior da Bahia, sendo jovens, com idade entre 20 e 30 anos, estudantes do Ensino Fundamental ou do Ensino Médio, falantes de um português semipopular ou semiculto. São pessoas comuns, amigos e/ou familiares de Adelmário Carneiro Araújo, que escreveram cartas e cartões informais. O material permitirá a análise de fatos linguísticos diversos e representa o português escrito na segunda metade do século XX por remetentes pouco escolarizados ou com média escolaridade.

Acervo da Família Oliveira (1962-1973):

São 23 cartas, escritas por 2 remetentes[15]. As cartas parecem ser de brasileiros "representantes" de um português semipopular.

Uma descrição completa do local de nascimento e da ficha biográfica de cada remetente pode ser visto em Carneiro *et al*, 2011 e também em www.uefs.br/cedohs.

3 EDIÇÃO FILOLÓGICA E EDIÇÃO DIGITAL

[12] Das 17 correspondências, 12 foram escritas por Carlos, sendo 11 destas endereçadas a Iracema, e 1, endereçada ao pai dela, pedindo-lhe a mão da jovem em casamento. 3 cartas foram escritas por Iracema para Carlos, e 2 cartas têm como autora intelectual Maria Laurinda, mãe de Carlos, redigidas e assinadas por outra pessoa e endereçadas a Iracema.

[13] Otto Soledade Júnior nasceu em Salvador-BA, em 1925. Formou-se como 'modelador de fundição' na antiga Escola Técnica de Salvador. Não teve diploma universitário, mas sempre gostou de ler e estudava por conta própria. Renée nasceu em 1932, em Ilhéus-BA. Estudou em um colégio de freiras, onde fez o primário, o ginásio e o curso normal, que oferecia a formação de professora. Nunca exerceu a profissão. Mas foi autodidata nos estudos de cinco línguas, religião comparada, mitologia, filosofia, psicologia, genética, história, física quântica. .

[14] A amostra, depositada em Valente-BA, é constituída por 79 cartas, 25 cartões e 80 envelopes, havendo seis cartas e dois cartões no formato aerograma. Sendo 17 baianos nascidos nas cidades: Salvador, Valente, Capim Grosso, Quijingue, Antas, Jacobina, Retirolândia, Santa Rita de Cássia, Feira de Santana, São Domingos, Barra Mansa, Lauro de Freitas, Conceição do Coité, Cícero Dantas, Salgadália, Euclides da Cunha e Riachão do Jacuípe. Além de 1 do Rio de Janeiro (RJ), 1 de São Luís (MA), 1 do Rio Grande do Norte (RN), 1 de Fortaleza (CE), 1 de Recife (PE) e 1 de Guarulhos (SP).

[15] São 21 apógrafas do recém-casado Arnaldo Andrade Dias, e 1 autógrafa, dirigidas a sua esposa Lourdinha [Maria de Lourdes Lima de Oliveira], na Bahia, e uma autógrafa de João Carvalho de Matos, "compadre" de Lourdinha. As cartas foram editadas por Bruna Trindade (I.C.Pibic/Projeto Vozes do Sertão em Dados). (CNPq. Processo 401433/2009-9/Consepe: 102/2009)].

ISSN 1415 - 8973

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.

Considerando as etapas do processo de constituição dos bancos de textos eletrônicos, vencida a etapa de localização e seleção de documentos, eles são transcritos segundo normas filológicas conservadoras e, a partir dessa transcrição, realiza-se, na etapa seguinte, a edição digital, em linguagem XML, finalizando com o preenchimento dos metadados.

No projeto CE-DOHS, os textos-fonte são apresentados em edição semidiplomática, segundo as normas de transcrição do PHPB, sendo oferecidas informações sobre os documentos, sua descrição extrínseca e intrínseca, e, sempre que possível, dados biográficos sobre os autores ou, no caso das cartas, sobre os remetentes e os destinatários, como nome, origem, idade, nível de escolaridade, profissão, estado civil etc. A codificação dos dados, textuais e extratextuais (ou metadados), é feita com o uso da ferramenta eDictor, o que possibilita a conversão dos textos para diferentes formatos (TXT, XML, HTML) e evita problemas de processamento eletrônico.

As edições filológicas, fidedignas ao texto original, realizadas segundo critérios de transcrição bem definidos, ganham, nos *corpora* digitais, uma versão modernizada, com padronização da grafia, da acentuação e desenvolvimento de abreviaturas, todas as alterações ficando visíveis ao leitor, o que possibilita o controle e mapeamento das intervenções realizadas nos textos, garantindo a recuperabilidade das formas originais. Respeitam-se, entretanto, na edição digital, as mudanças de parágrafo, de linha, as correções do autor, os acidentes do suporte, a orientação da escrita etc. Com isso, oferece-se uma versão eletrônica de textos sem perder o rigor filológico.

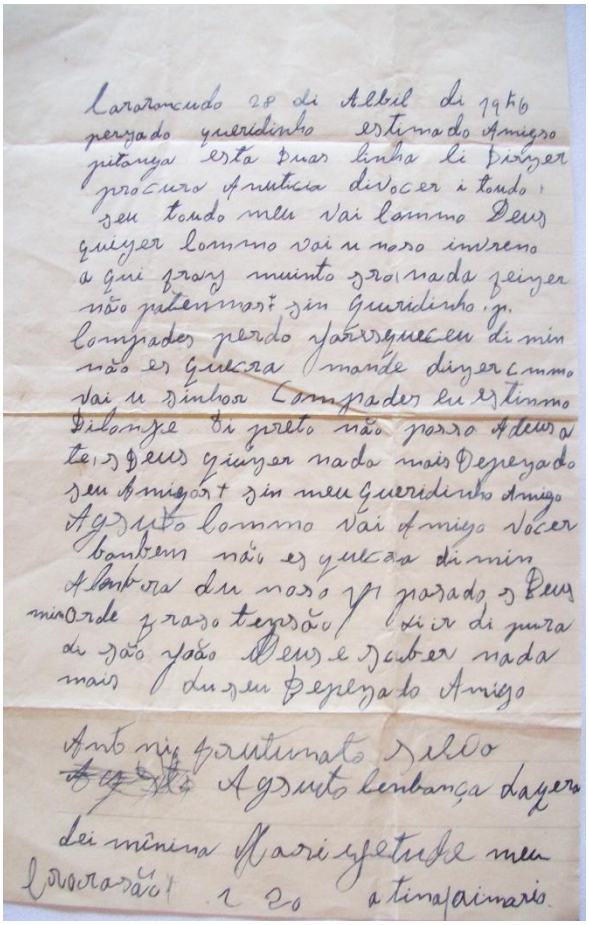
Procura-se, no âmbito do CE-DOHS, seguir os mesmos critérios de edição digital e de anotação morfossintática que seguem outros projetos de *corpora* eletrônicos, como o projeto Corpus Histórico do Português Tycho Brahe (UNICAMP), o projeto Labor Histórico, da Universidade Federal do Rio de Janeiro (UFRJ), 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.[16]

A seguir, será apresentado o passo-a-passo da edição digital de textos, segundo a metodologia utilizada no CE-DOHS, tomando, como exemplo, uma carta escrita por *mão inábil*, do acervo Cartas em Sisal (1906-2000):

[16] 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.

ISSN 1415 - 8973

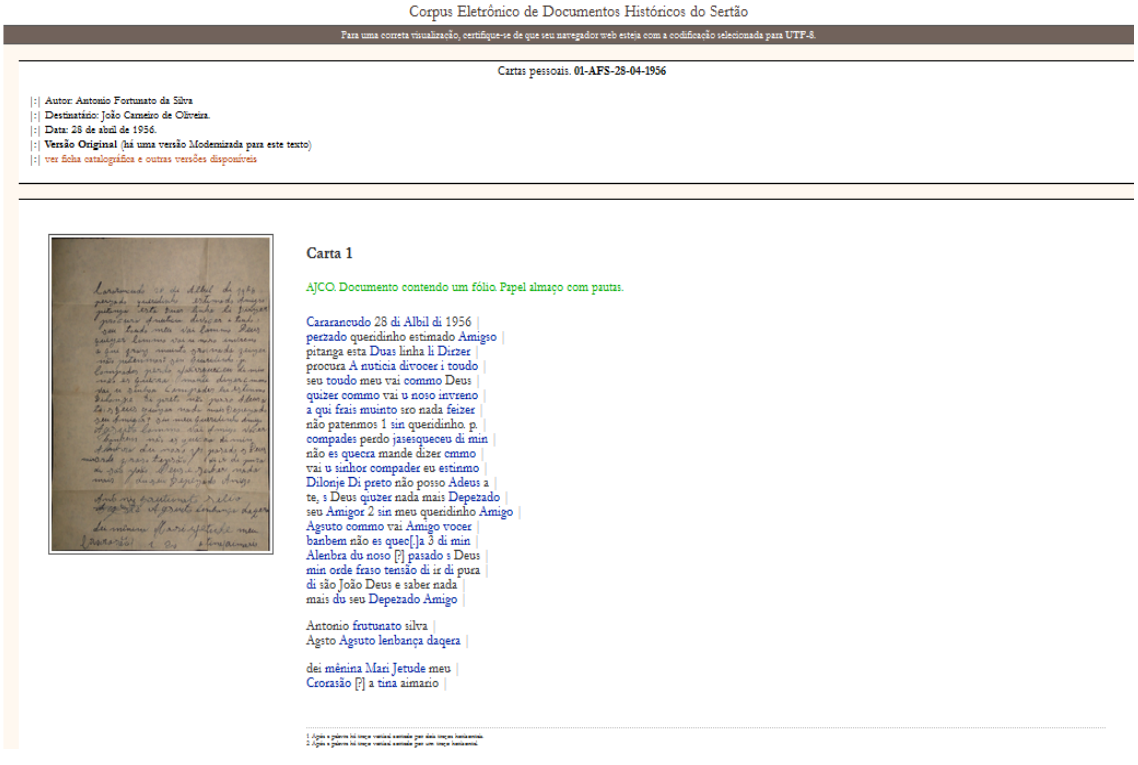
Figura 1 - Edição semidiplomática com fac-símile (SANTIAGO, 2012).



Carta 1 AJCO. Documento contendo um fólio. Papel almaço com pautas.

Cararancudo 28 di Albil di 1956| perzado queridinho estimado Amigso| pitanga esta Duas linha li Dirzer| procura A nuticia divocer i toudo| seu toudo meu vai commo Deus| quizer commo vai u noso invreno| a qui frais muinto sro. nada feizer| não patenmos sin queridinho. p.| compades perdo jasesqueceu di min| não es quecra mande dizer cmmo| vai u sinhor compader eu estinmo| Dilonje Di preto não posso Adeus a| te, s*e* Deus qiuzer nada mais Depezado| seu Amigor sin meu queridinho Amigo| Agsuto commo vai Amigo vocer| banbem não es quec[.] a di min| Alenbra du noso [?] pasado s*e* Deus| min orde fraso tensão di ir di pura| di são João Deus e saber nada| mais du seu Depezado Amigo| Antonio frutunato silva| ~~Agsto~~ Agsuto lenbança daqera| dei mênina Mari Jetude meu | Crorasão [?] a tina aimario|

Figura 2 - Edição diplomática com fac-símile, no eDictor.



ISSN 1415 - 8973

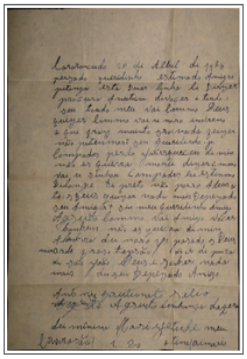
Figura 3 - Edição modernizada final.

Corpus Eletrônico de Documentos Históricos do Sertão

Para uma correta visualização, certifique-se de que seu navegador web esteja com a codificação selecionada para UTF-8.

Cartas pessoais. **01-AFS-28-04-1956**

|:| Autor: Antonio Fortunato da Silva
|:| Destinatário: João Carneiro de Oliveira.
|:| Data: 28 de abril de 1956.
|:| **Versão modernizada** (há uma versão original para este texto)
|:| ver ficha catalográfica e outras versões disponíveis



Carta 1

AJCO. Documento contendo um fólio. Papel almaço com pautas.

Carrancudo 28 de Abril de 1956 |
prezado queridinho estimado amigo |
pitanga esta duas linha lhe dizer |
procura a notícia de você e todo |
seu todo meu vai como Deus |
quiser como vai o nosso inverno |
aqui faz muito seo nada fazer |
não patenmos 1 sem queridinho. p. |
compadres perdo já esqueceu de mim |
não esqueça mande dizer como |
vai o senhor compadre eu estimo |
de longe de perto não posso adeus a |
te, se Deus quiser nada mais de prezado |
seu amigo 2 sem meu queridinho amigo |
Augusto como vai amigo você |
também não esqueç[.]a 3 de mim |
lembra do nosso [?] passado se Deus |
mim ordem faço intenção de ir de pura |
de são João Deus e saber nada |
mais do seu prezado amigo |

Antonio Fortunato silva |
Agsto Augusto lembrança daquela |

dei menina Maria Jertrude meu |
Coração [?] a tinha aimario |

Figura 4 – Lista de dados alterados.

Cartas pessoais. **01-AFS-28-04-1956**

|:| Autor: Antonio Fortunato da Silva
|:| Destinatário: João Carneiro de Oliveira.
|:| Data: 28 de abril de 1956.
|:| **Lista dos itens editados** (por ordem alfabética)
|:| ver ficha catalográfica e outras versões disponíveis

Item Original	Edições	Fonologia
A	pad: a,	
Adeus	pad: adeus,	
Agsuto	pad: Augusto,	
Agsuto	pad: Augusto,	
Albil	pad: Abril,	
Alenbra	pad: lembra,	
Amigo	pad: amigo,	
Amigo	pad: amigo,	
Amigo	pad: amigo,	
Amigor	pad: amigo,	
Amigso	pad: amigo,	
a qui	jun: aqui,	
banbem	pad: também,	
Cararancudo	pad: Carrancudo,	
cmmo	pad: como,	
commo	pad: como,	
commo	pad: como,	
commo	pad: como,	
compader	pad: compadre,	
compades	pad: compadres,	
Crorasão	pad: Coração,	
daqera	pad: daquela,	
Depezado	pad: de prezado, seg: De pezado,	
Depezado	pad: prezado, seg: De pezado,	
di	pad: de,	
di	pad: de,	
di	pad: de,	
di	pad: de,	
di	pad: de,	
di	pad: de,	

ISSN 1415 - 8973

3.1 PADRONIZAÇÃO ORTOGRÁFICA DOS TEXTOS

A padronização ortográfica dos textos originais – que apresentam muitas variações ortográficas – é necessária na edição modernizada, que será anotada linguisticamente, porquanto dessa padronização depende uma maior eficiência de programas de etiquetação automática. Não se interveio, todavia, nos regionalismos, arcaísmos lexicais e neologismos.[17]

Foram padronizadas palavras como: cete (sete); podião (podiam); acentarmos (acertarmos); emcomodos (incômodos); cabeça (cabeça); rezulvida (resolvida); intendo (entendo); avizarei (avisarei); instruçons (instruções); lógar (lugar); prencipio (princípio); fis (fiz); conclusão (conclusão); comferir (conferir); conversou (conversou); pudia (podia); piriodos (períodos); poderse (pudesse); munto (muito).

Os seguintes são exemplos de padronização da acentuação: debito (débito); tirár (tirar); nos (nós); más (mas); credito (crédito); contavamos (contávamos); ultima (última); negocio (negócio).

Com a ferramenta eDictor, pode-se selecionar a palavra original, que apresenta variação, e editá-la segundo a grafia padrão, ficando disponível a lista de alterações realizadas na edição modernizada. Seleciona-se, manualmente, palavra por palavra que se deseja alterar (substituir, separar, juntar, expandir etc.)[18].

No projeto CE-DOHS, no período inicial das edições, os pesquisadores enfrentaram dificuldades para decidir o que padronizar; por exemplo, casos em que o pronome está ligado ao verbo, em posição enclítica, sem hífen, o que contraria a norma-padrão atual (athe que eu poça pençar arespeito de *poderse* cuidar neste negocio; *tendome* medicado veio o Jezuino para me dar conta, e eu poder *dalla*; V*ossa* Exc*elenci*a esta rezulvida a *intregarlhe* a obra etc.). Nesses casos, decidiu-se manter a ortografia original, porque isso faz diferença no estudo da sintaxe dos clíticos no PB.

Pensou-se inicialmente em desenvolver apenas as abreviaturas desconhecidas do leitor atual; depois os pesquisadores decidiram expandir todas as abreviaturas, em razão de sua expansão evitar erros na aplicação do anotador automático, que é incapaz de reconhecer a forma correspondente. Como se vê, há uma relação dinâmica nas etapas de constituição de *corpora* digitais: uma decisão tomada em determinado nível (por exemplo, na padronização dos textos) tem consequências no nível posterior (por exemplo, na anotação linguística). As formas abreviadas, que se mantêm na edição semidiplomática, aparecem expandidas na edição XML, usando-se a etiqueta <expan>.

3.2 AS DIFICULDADES QUE O *CORPUS* DE *MÃOS INÁBEIS* APRESENTA NA EDIÇAO PARA USO ELETRÔNICO

As cartas de inábeis, editadas por Santiago (2012), estão disponíveis no CE-DOHS, nas versões semidiplomática e modernizada. A amostra é constituída por 91 cartas pessoais, escritas ao longo do século XX, como já mencionado, a maioria nas décadas de 50, 60 e 70, por 43 remetentes oriundos da zona rural dos municípios de Riachão do Jacuípe, Conceição do Coité e Ichu, localizados no semiárido baiano.

Trata-se de um *corpus* significativo, seja porque as cartas foram trocadas em relação de simetria entre sertanejos que fazem parte de um contexto sociocultural semelhante, seja pelo grau de transparência aos usos vernáculos que apresentam, pois são textos próximos de uma escrita cotidiana, de caráter afetivo, demonstrando considerável grau de intimidade entre os

[17] No caso das amostras de fala, quase não se fizeram alterações.

[18] Os criadores do eDictor, Paixão de Souza, Kepler e Faria (2010a), têm trabalhado para sofisticar a ferramenta, tornando-a mais inteligente, a fim de que os processos de edição digital dos textos e sua anotação morfossintática e sintática sejam mais automatizados, facilitando a constituição dos bancos eletrônicos.

ISSN 1415 - 8973

remetentes e destinatários. Os redatores são lavradores, trabalham com agricultura e criação de animais, possuem baixas condições financeiras e pouca escolarização; a maioria teve contato com as primeiras letras em espaços extraescolares, como a própria casa ou a de parentes, considerando-se a ausência e/ou precariedade das escolas e seu funcionamento irregular, na zona rural dessa região da Bahia.

Um conjunto de propriedades presente nas cartas fornece algumas pistas para perceber que os redatores são indivíduos pouco familiarizados com a língua escrita e, por isso, a amostra apresenta indícios da variedade popular do PB. O afastamento das normas gramaticais e ortográficas indica que os redatores tiveram pouco contato com os modelos normativos prescritos pela escola e que os textos são, então, produtos de *mãos inábeis*. Essa pouca habilidade com a escrita é evidenciada em produtos gráficos que apresentam marcas de inabilidade em vários planos. Para identificar essas marcas, realizou-se a descrição (SANTIAGO, 2012), a partir dos trabalhos de Marquilhas (2000), Barbosa (1999) e Oliveira (2006), das seguintes propriedades:

> a) aspectos supragráficos e paleográficos, como ausência de *cursus*, módulo grande, ausência de regramento, traçado inseguro e letras com aparência desenquadrada;
>
> b) segmentação gráfica, em dados de hipossegmentação e hipersegmentação;
>
> c) aspectos relacionados à *escriptualidade*, como grafia de sílabas complexas, representação da nasalidade e representação de dígrafos;
>
> d) escrita fonética, como os casos de elevação de vogais médias (em posição pretônica, postônica e em monossílabos), abaixamento de vogais altas, anteriorização, e posteriorização de vogais, redução de ditongos, ditongação, nasalização, palatalização, rotacismo, lambdacismo, prótese, paragoge, aférese, síncope, apócope e metátese;
>
> e) repetição lexical.

Essas marcas de inabilidade oferecem desafios para a edição. Considerando que há, nos documentos, muitas peculiaridades que requerem um cuidado especial, o trabalho filológico com manuscritos desse tipo exige que se assegure, ainda mais, a confiabilidade da edição, afinal, a investigação acerca de alguns aspectos presentes nesses textos necessita de informações precisas, como por exemplo, em relação à grafia utilizada. Isso exige uma interferência mínima do editor, já que a presença e a frequência de formas não convencionais fornecerão indícios da maior ou menor familiaridade do redator com a escrita.

A grande quantidade de variação ortográfica que os manuscritos apresentam implica, na etapa da edição modernizada, em decidir sobre o que padronizar. Alguns aspectos de aquisição da escrita podem se confundir com uma escrita fonética, com regionalismos, em que não se deve intervir.

Em relação às propostas de segmentações não convencionais apresentadas nos textos, com grafias hipersegmentadas e hipossegmentadas, nota-se a dificuldade dos redatores em interpretar as fronteiras das palavras, de modo que a inserção ou não do espaço em branco parece ser baseada na percepção da fala ou nas próprias experiências anteriores com o código escrito. Observa-se, ainda, que alguns aspectos da linguagem escrita oferecem maior dificuldade para aqueles que estão nos estágios iniciais da aquisição, como a grafia de sílabas complexas. Há, no *corpus,* um elevado número de dados com grafias irregulares em sílabas complexas, de modo que as ocorrências envolvendo o /r/ são as que demonstram uma maior variação; mas há, também, ocorrências com o /l/ e o /s/ (como em *lenbarnsa* por *lembrança, ato* por *alto, eteve* por *esteve)*. A representação gráfica da nasalidade também apresenta irregularidades, e predominam tanto representações exageradas, com a repetição da consoante nasal, como casos em que essa consoante é omitida (como em *vanmos* por *vamos, mado* por *mando)*. Outro aspecto identificado é a representação de dígrafos, realizada sob o princípio da relação monogâmica entre letra e som (como em *pigenno* por *pequeno*).

ISSN 1415 - 8973

Sobre a transferência de traços próprios da oralidade para a escrita, os fenômenos fônicos, constatou-se sua presença nas mãos de todos os redatores. Alguns mais gerais, que não são específicos dos inábeis, como a apócope de /R/ em final de verbos no infinitivo e a elevação das vogais médias pretônicas e postônicas, e outros mais raros, estigmatizados, como a aférese, a prótese e o rotacismo (como em *chora* por *chorar, sigundo* por *segundo, duentada* por *adoentada, avoar* por *voar).*

Dada a especificidade do *corpus*, a opção por apresentar o texto em edição semidiplomática, acompanhada do fac-símile, facilita o trabalho do pesquisador no que se refere à necessidade de que sejam obtidas informações gráficas precisas, principalmente quando o que se apresenta é uma escrita que surpreende. Além disso, garante a avaliação de alguns aspectos, como os dados supragráficos e paleográficos, importantes na caracterização dos manuscritos, que se torna mais segura com a visualização das imagens.

3.3 ANOTAÇÃO LINGUÍSTICA

Atualmente, os textos que compõem o CE-DOHS estão na forma não-anotada, ou seja, sem informações linguísticas. Essa tarefa será desempenhada na próxima fase do projeto, depois de realizadas oficinas, junto ao projeto Corpus Histórico do Português Tycho Brahe, para treinamento de bolsistas de mestrado e doutorado que irão colaborar com a equipe de pesquisadores do CE-DOHS nessa tarefa de etiquetação morfossintática e anotação 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, na edição modernizada dos textos: 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. Essa revisão da anotação exige conhecimento de morfossintaxe, atenção e cuidado.

Até o presente, a equipe de pesquisadores do CE-DOHS ocupou-se com a edição filológica e modernizada dos textos (sobretudo com esta, já que a maior parte dos textos que integram o CE-DOHS já se encontravam filologicamente editados no Banco DOHS, no âmbito do projeto Vozes do Sertão em Dados), o que demandou bastante tempo.

A revisão das edições modernizadas – 1084 cartas manuscritas, mais livros manuscritos, textos impressos, amostras de fala – é feita manualmente, exigindo bastante dedicação da equipe. Todos os textos do banco já passaram pela primeira revisão. Atualmente, passam pela revisão final, antes de receberem a anotação linguística.

O CE-DOHS, no processo de anotação linguística dos textos, seguirá os mesmos padrões utilizados por outros projetos de *corpora* eletrônicos, a exemplo do projeto Corpus Histórico do Português Tycho Brahe, que é o maior *corpus* eletrônico anotado de textos históricos em português.

As etapas básicas que as equipes de pesquisadores de projetos de *corpora* anotados seguem são essas: anotação de edição; anotação morfossintática e anotação sintática.

Na primeira etapa, é utilizado o eDictor; trata-se da codificação de informações sobre o texto original, ou sobre decisões editoriais, ou sobre a estrutura do texto. Essa etapa, semiautomática, já foi vencida pelo CE-DOHS.

A próxima etapa, automatizada, é a de anotação morfossintática, com uso do programa desenvolvido por Kepler (2007; 2010b), um analisador morfossintático automático, com taxa

ISSN 1415 - 8973

de acerto de 95%, acoplado ao eDictor. Os erros possíveis de etiquetação, como já dito, devem ser corrigidos manualmente pelo linguista[19].

A anotação sintática, também automatizada, é a terceira e última etapa na constituição de *corpora* anotados; ela diz respeito à identificação e codificação da estrutura sintagmática do texto. É uma tarefa complexa, mais do que a etiquetação morfossintática, e exige um *parser*, ou, na forma aportuguesada, um parseador, que realiza a análise sintática, reconhecendo identidades em sequências linearmente dispostas e padrões de agrupamentos hierárquicos.

Para alcançar seu objetivo final – que é a busca automática de dados para estudo da história do PB – a equipe de pesquisadores do CE-DOHS deve cumprir essas duas últimas etapas: a anotação morfossintática e a anotação sintática do *corpus*. O trabalho, até o presente, foi imenso! E ainda há muito a ser feito!

4 CONSIDERAÇÕES FINAIS

Hoje, no banco do projeto CE-DOHS – que teve início em 2011 e deverá ser finalizado em 2018 –, estão à disposição dos pesquisadores e demais interessados a edição semidiplomática dos documentos, com fac-símile; as transcrições das amostras de fala; a edição modernizada dos textos; resumos com a contextualização sócio-histórica dos materiais e com informações sobre os autores e, no caso das cartas, sobre os destinatários, além da ficha técnica, com nome dos editores, revisores etc.

Nos próximos anos, será possível consultar o *corpus* anotado, fazendo buscas automáticas. O CE-DOHS possui um material extenso e rico, que oferece à comunidade científica diferentes possibilidades de pesquisa; e, para a história do PB – especialmente do português no interior da Bahia, através de um contínuo, do mais escolarizado para o menos escolarizado, *os inábeis* –, trata-se de um *corpus* extremamente significativo.

REFERÊNCIAS

ALMEIDA, N. L. F. de; CARNEIRO, Z. de O. N. (Org.). *Coleção amostras da língua falada no semiárido baiano*. Feira de Santana: UEFS, 2008.

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.. *Para uma história do português colonial*: aspectos linguísticos em cartas do comércio. 1999. 484f. Tese (Doutorado em Língua Portuguesa) – Faculdade de Letras, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1999.

CARNEIRO, Z. de O. N. *Cartas brasileiras* (1809-1907): um estudo filológico-linguístico. Tese de Doutorado. Campinas: UNICAMP, 2005.

CARNEIRO, Z. de O. N. (Org.). *Cartas brasileiras (1809-2000)*: coletânea de fontes para o estudo do português. Feira de Santana: UEFS, 2011.

CARNEIRO, Z. de O. N., LACERDA, M. F. de O. *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.

[19] O código de etiquetas do eDictor baseia-se no sistema de anotação manual dos Penn Corpora of Historical English (KROCH; SANTORINI; DIERTANI, 2010), da Universidade da Pensilvânia, Estados Unidos. Esse sistema, para adequar-se às peculiaridades da gramática do português, sofreu pequenas alterações.

ISSN 1415 - 8973

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

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.

FARIA, P.; KEPLER, F. N.; PAIXÃO DE SOUSA, M. C. An Integrated Tool for Annotating Historical Corpora. In: *Fourth Linguistic Annotation Workshop* (LAW IV), 48th Annual Meeting of the ACL, 2010, Uppsala, Sweden. Proceedings of the Fourth Linguistic Annotation Workshop, 2010b. p. 217-221.

GALVES, C., and H. Britto. 2002. *The Tycho Brahe Corpus of Historical Portuguese*. Department of Linguistics, University of Campinas. Online publication, _rst edition.

GANDRA, A. A. *Cartas de amor na Bahia do século XX*: normas linguísticas, práticas de letramento e tradições do discurso epistolar. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal da Bahia, Salvador, 2010.

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.

KROCH, A.; SANTORINI, B.; DIERTANI, A. *Penn Parsed Corpus of Modern British English*. 2010. Disponível em: [http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html].

LABOV, W. *Sociolinguistic Patterns*. Pennsylvania: University of Pennsylvania Press, 1972.

LOBO, T. C. F. *Para uma sócio-linguística histórica do português no Brasil*: edição filológica e análise linguística de cartas particulares do Recôncavo da Bahia, século XIX. Tese (Doutorado em Letras) – Instituto de Letras, Universidade Federal da Bahia, Salvador, 2001.

LUCCHESI, D. *Variação e norma*: elementos para uma caracterização sociolinguística do português do Brasil. Revista internacional de língua portuguesa, n. 12, p. 17-28, 1994.

LUCCHESI, D. *As duas grandes vertentes da história sociolinguística do Brasil*. DELTA, São Paulo, v.17, n.1, p. 97-130, 2001.

MARQUILHAS, R. *A faculdade das letras*: leitura e escrita em Portugal no séc. XVII. Lisboa: Imprensa Nacional-Casa da Moeda, 2000.

MATTOS E SILVA, R. V. De fontes sócio-históricas para a história social linguística do Brasil: em busca de indícios. In: MATTOS E SILVA, R. V. (Org.). *Para a história do português brasileiro*: primeiros estudos. V. II, tomos I e II. São Paulo: Humanitas/FFCHL/USP:FAPESP, 2001, v.2, t. 2, p. 275-302.

ISSN 1415 - 8973

OLIVEIRA, K. *Negros e escrita no Brasil do século XIX*: sócio-história, edição filológica de documentos e estudo linguístico. 2006. 3v. 1144f. Tese (Doutorado em Letras) – Instituto de Letras, Universidade Federal da Bahia, Salvador, 2006.

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, 2009. p. 69-105.

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, 2010a.

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

SANTOS, J. V.; BRITO, G. S. *Fotografia técnica de documentos para formação de corpora digitais eletrônicos*: o método Lapelinc. In: Revista LETRAS & LETRAS, v. 30, n. 2, 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.