

A ferramenta de busca E-CORP aplicada ao Corpus Eletrônico de Documentos Históricos do Sertão

The E-CORP search tool applied to the Corpus Eletrônico de Documentos Históricos do Sertão

Igor Leal Souza*

*Universidade Estadual de Feira de Santana
Feira de Santana, Bahia, Brasil*

Gabriela Ribeiro Peixoto Rezende Pinto**

*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*

Pablo Faria****

*Universidade Estadual de Campinas
Campinas, São Paulo, Brasil*

Mariana Fagundes de Oliveira Lacerda*****

*Universidade Estadual de Feira de Santana
Feira de Santana, Bahia, Brasil*

Resumo: Com o surgimento das Humanidades Digitais, cada vez mais a tecnologia está sendo aceita como parceira no desenvolvimento de pesquisas linguísticas. Essa parceria traz uma contribuição para ambas as áreas: para a linguística, novas possibilidades de estudo, com mais velocidade e confiabilidade; para a computação, essa interdisciplinaridade a enriquece com novos conceitos e aumenta sua área de atuação. Este trabalho, nessa interface entre a computação e a linguística, tem como objetivo apresentar o desenvolvimento do *E-Corp* – uma ferramenta de busca de dados para fins linguísticos, e sua aplicação no CE-DOHS – *Corpus Eletrônico de Documentos Históricos do Sertão* (www.uefs.br/cedohs), da Universidade Estadual de Feira de Santana (UEFS). O desenvolvimento dessa ferramenta visa a auxiliar os pesquisadores da área da linguística a fazer exploração de *corpora* de maneira mais rápida e confiável.

Palavras-chave: Humanidades Digitais. *Corpora* Eletrônicos. Busca de Dados. XML. Estudos Linguísticos.

*Discente da Universidade Estadual de Feira de Santana, igorengcomp@gmail.com.

**Professora Adjunta do Departamento Ciências da Computação da Universidade Estadual de Feira de Santana/UEFS, Bahia/Brasil, gabrielarprp@gmail.com.

***Professora Plena do Departamento de Letras e Artes da Universidade Estadual de Feira de Santana/UEFS, Bahia/Brasil zenaide.novais@gmail.com.

****Professor Doutor I do Instituto da Linguagem da Universidade Estadual de Campinas/UNICAMP, São Paulo/Brasil, pablofaria@gmail.com.

*****Professora Adjunto B do Departamento de Letras e Artes da Universidade Estadual de Feira de Santana/UEFS, Bahia/Brasil marianafag@gmail.com.

Abstract With the emergence of Digital Humanities, the technology is being increasingly applied in linguistic research. This partnership contributes with speed and reliability in data manipulation and retrieval, and also triggers new study opportunities. For computing, this interdisciplinary work with linguistics increases its application area, enriching it with new concepts. In this regard, the present paper introduces the E-Corp tool and its application to the electronic database of the CE-DOHS – Corpus Eletrônico de Documentos Históricos do Sertão, the State University of Feira de Santana (UEFS). This tool is meant to help linguistic researchers in making faster and more reliable searches in corpora, allowing for faster development of research and corpus construction.

Keywords: Digital humanities; Electronic *corpora*; Data extraction; XML; Linguistic studies.

1 INTRODUÇÃO

Este artigo é fruto de uma parceria entre a engenharia de computação e a linguística. A ferramenta *E-Corp*, aqui apresentada, foi criada para colaborar com o Projeto CE-DOHS – *Corpus* Eletrônico de Documentos Históricos do Sertão (www.uefs.br/cedohs), e com outros projetos de *corpora* eletrônicos, na busca de ocorrências de palavras em documentos no formato *XML*. O desenvolvimento de ferramentas de busca como o E-Corp otimiza o acesso aos dados contidos em documentos históricos, o que pode contribuir para o avanço das pesquisas linguísticas.

O texto está organizado da seguinte forma: o item 1 é composto por uma breve introdução sobre a Linguística de *Corpus* e a tendência de construção de *corpora* digitais, além de apresentar, de forma sucinta, as Humanidades Digitais. No item 2 são apresentados o banco de dados CE-DOHS e as ferramentas computacionais eDictor e *E-Corp*. Nesse item, é possível conferir, ainda, como se deu o desenvolvimento do programa E-Corp, sua estrutura interna e exemplos de sua aplicação. Por fim, no item 3, são apresentadas as considerações finais.

2 LINGUÍSTICA DE *CORPUS*

Segundo Sardinha (2004, p. 325), a linguística de *corpus* 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”. Sardinha (2004) também alerta para o fato de que a linguística de *corpus* tem sofrido mudanças desde o desenvolvimento do computador. Ao final da década de 50, as críticas ao processamento manual de grandes *corpora* levavam à desconfiança dos dados levantados, já que esse tipo de procedimento é mais suscetível a erros. Já no começo da década de 60, a mudança passa a acontecer nos grandes centros universitários, onde pesquisas desse tipo são realizadas com a utilização de computadores para auxiliar no levantamento e na coleta de dados.

Paixão de Sousa e Kepler (2007) atestam sobre a importância desse tipo de edição para a construção de uma base de dados eletrônica de documentos históricos:

Os estudos históricos realizados com base em textos antigos dependem, antes de tudo, da garantia da fidelidade às formas originais dos textos – sendo este o pilar de sustentação que qualquer estudo linguístico, em qualquer quadro teórico, deve pressupor. Entretanto, no caso dos corpora eletrônicos, esse pressuposto fundamental precisa ser integrado com requerimentos impostos pela vertente computacional e linguística dos estudos – tais sejam: a necessidade de quantidade, agilidade e automação no trabalho estatístico de seleção de dados. (PAIXÃO DE SOUSA; KEPLER, 2007).

Com essa demanda crescente, a construção de base de dados eletrônica se faz cada dia mais essencial, uma vez que com esse tipo de banco, a agilidade e a confiabilidade nos dados se tornam maiores.

2.1 CORPUS ELETRÔNICO

Com o passar dos anos, a tecnologia tem se tornado cada vez mais acessível e, como resultado dessa acessibilidade, há uma maior quantidade de documentos digitais. Diante desse cenário, surge as “Humanidades Digitais”, para a qual, conforme alerta Paixão de Sousa (2011), é possível atribuir mais de uma definição na tentativa de explicar o que descreve essa área de estudo, visto que é um tema que está em constante discussão e apresenta pontos de vista variados. Entre algumas dessas definições, acredita-se que as que melhor se aplicam ao contexto do trabalho aqui proposto são algumas das elencadas em Paixão de Sousa (2011):

P. O'Donnell, por exemplo, define “Humanidades Digitais” como uma “*atividade interdisciplinar que transfere para os meios digitais o trabalho tradicional com textos, objetos culturais e outros dados, com isso estendendo radicalmente seus usos potenciais.* [...] James Cummings, para quem “Humanidades Digitais” designa um campo de estudos cujo objeto de (*auto-reflexão*) é a própria aplicação da tecnologia digital nas investigações em humanidades (PAIXÃO DE SOUSA, 2011).

Tais definições colocam como pontos comuns atividades que buscam, de alguma maneira, utilizar meios digitais em função de atividades antigas, anteriores aos computadores, ou anteriores a determinadas tecnologias. Seguindo essa tendência, torna-se cada vez mais recorrente a construção de *corpora eletrônicos de textos* – conjuntos de textos digitais para fins linguísticos com o intuito de possibilitar estudos nos diversos campos, como, por exemplo o léxico, o semântico, o sintático e o discurso (cf. Paixão de Sousa, 2014).

Ainda tratando de estudo linguístico, podemos citar a construção de banco de árvores a partir de *corpora* sintaticamente anotados. A partir dessas anotações, o texto possui diversas marcações, podendo ser utilizado para aplicações diversas dentro da linguística computacional, seja para extração de informação que não seriam possíveis, de forma rápida aos seres humanos, até a aplicação de traduções (cf. FARIA; GALVES 2016). Como destaque desse tipo de construção no Brasil, podemos citar o projeto Corpus Histórico do Português Tycho Brahe, que disponibiliza 76 textos (3.302.696 palavras) para pesquisa livre, com um sistema de anotação linguística em duas

etapas: anotação morfológica aplicada em 45 textos, num total de 2.012.798 palavras, e anotação sintática, aplicada em 27 textos, num total de 1.234.323 palavras (cf. GALVES; ANDRADE; FARIA 2017).

Faria e Galves (2016) apontam a importância da anotação e da utilização da computação para as demais análises linguísticas que podem ser feitas:

Corpora anotados são importantes em todos os ramos da linguística, uma vez que constituem bases de dados perenes sobre as quais se podem efetuar análises qualitativas e quantitativas de vários tipos, que complementam outras abordagens como o recurso à intuição dos falantes ou ainda estudos baseados em experimentos, prática corrente em aquisição da linguagem e cada vez mais em análises sintáticas. Em linguística histórica, uma vez que não há falantes nativos disponíveis, os corpora são indispensáveis. Eles podem até abranger a totalidade dos dados disponíveis, quando se consideram os períodos mais antigos das línguas. A anotação morfossintática permite explorar de maneira consistente e reproduzível quantidades de dados inacessíveis ao trabalho manual, permitindo um acesso cada vez mais completo e confiável aos dados do passado.

Assim, a construção de uma base de dados eletrônicos vem se mostrando uma tendência mundial, tanto que projetos de grande importância para os estudos linguísticos já lançaram mão dessa tecnologia, como o pioneiro Penn Helsinki Parsed Corpus of Middle English, (<http://www.ling.upenn.edu/hist-corpora>) coordenado por Anthony Kroch na Universidade da Pensilvânia e os seus afiliados; o York Helsinki Parsed Corpus of Old English Poetry, por Susan Pintzuk e Leendert Plug; o York Toronto Helsinki Parsed Corpus of Old English Prose, por Ann Taylor, Anthony Warner, Susan Pintzuk, Frank Beths, ambos na Universidade de York; e o Parsed Corpus of Early English Correspondence, por Ann Taylor, Anthony Warner, Susan Pintzuk na Universidade de York, e por Terttu Nevalainen e Arja Nurmi na Universidade de Helsinki, além de outros como o projeto de Corpus annoté syntaxiquement de textes de français (9^e au 17^e siècle), por F. Martineau e Paul Hirschbuhler na Universidade de Ottawa, e o projeto Corpus Dialetal Sintático (CordialSin), por Ana Maria Martins, na Universidade de Lisboa.

3 O CE-DOHS E AS FERRAMENTAS COMPUTACIONAIS: eDICTIONARY E E-CORP

O CE-DOHS – *Corpus* Eletrônico de Documentos Históricos do Sertão, sediado na Universidade Estadual de Feira de Santana (UEFS), é uma base de dados eletrônica composta por documentos do banco DOHS – Documentos Históricos do Sertão. Trata-se de um conjunto de documentos escritos (manuscritos e impressos) e de textos orais, dos séculos XIX e XX, representativos das vertentes culta e popular do PB.

Os textos-fonte do banco CE-DOHS são disponibilizados 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 a ferramenta eDICTOR¹ (PAIXÃO DE SOUSA, KEPLER E FARIA, 2007)), a qual possibilita a conversão dos textos para diferentes formatos (TXT, XML, HTML) e evita problemas de processamento eletrônico.

O objetivo dos pesquisadores responsáveis pelo banco em questão é possibilitar o acesso a uma quantidade significativa de documentos, para que possam ser aplicadas teorias que “propugnem uma apreensão globalizante do objeto através de sua estrutura interna (linguística)”, além das teorias que “propõem a apreensão dos fatos através da interação sistemática de relações linguísticas com as disposições e relações nas quais esse sistema se atualiza (as relações sociolinguísticas)” (LACERDA; CARNEIRO; SANTIAGO, 2017, p.128).

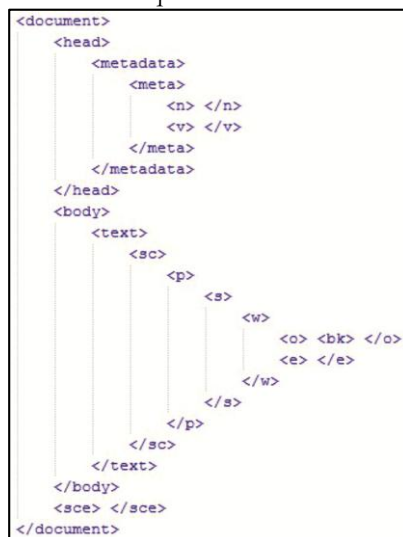
De modo geral, esse tipo de edição eletrônica em linguagem XML e seus produtos abrem uma grande janela de oportunidades para edições filológicas e pesquisas em linguística histórica, por possibilitar diferentes camadas de informação em um único documento, ao invés do uso não manipulável da digitação em Word (PAIXÃO DE SOUZA, 2006) e por atender a “dois objetivos principais: (i) ser o mais neutra possível (em relação ao conteúdo textual codificado) e (ii) atender a necessidades linguísticas e filológicas” (PAIXÃO DE SOUSA; KEPLER; FARIA, 2009). Assim, o uso dessa tecnologia poderá dar um novo impulso à Linguística Histórica.

3.1 O eDICTOR

Como apontado anteriormente, os documentos do CE-DOHS encontram-se editados eletronicamente a partir do eDICTOR, desse modo, os documentos seguem o padrão gerado por essa ferramenta. A partir da observação dos arquivos que foram gerados, foi possível obter (pode ser “foi possível capturar”) a estrutura do XML, apresentada na *Figura 1* (LEAL, 2016, p.31):

¹ Por meio de parceria tecnológica com o projeto Corpus Histórico do Português Tycho Brahe (www.tycho.iel.unicamp.br).

Figura 1: Estrutura do arquivo XML utilizado no CE-DOHS



Fonte: Extraído de Leal, 2016, p. 31.

Como pode ser visto na *Figura 1*, a etiqueta *document* é a etiqueta pai e é a partir dela que todas as outras etiquetas são geradas. Todo o conteúdo do documento está contido nessa etiqueta. Logo em seguida, é gerada a etiqueta *head*, que é o cabeçalho da carta, onde estão dispostos os metadados, e, para isso, são geradas as etiquetas *metadata*. Cada conteúdo do metadados está contido na etiqueta *meta*, que contém duas outras etiquetas *n* e *v*, nas quais os valores podem ser recuperados. A etiqueta *n* contém o nome e a etiqueta *v* contém o valor referente do nome de *n*. Tais etiquetas compõem o cabeçalho. O texto do documento é gerado logo depois do cabeçalho, sendo gerada a etiqueta *body* e logo em seguida, a etiqueta *text*.

A etiqueta *sce*, localizada abaixo da etiqueta *text*, é a etiqueta que divide o texto em seções, caso necessário. Dentro das seções existem parágrafos, sentenças e palavras; as etiquetas para essas seções são *p*, *s* e *w*, respectivamente. O conteúdo só é encontrado dentro das etiquetas *o* e *e*, onde as palavras originais e editadas são encontradas. Caso aconteça uma quebra de linha, é gerada uma etiqueta *bk* dentro da etiqueta *o*, indicando que houve a quebra. O mesmo acontece se a quebra de linha acontecer no meio da palavra.

Ainda é possível observar que existe a etiqueta *sce*, responsável pelo controle de notas de rodapé. Essa etiqueta pode aparecer tanto na palavra, quanto no final do texto, variando apenas quanto à posição em que é usada. Só após o uso dela é que o documento é fechado.

Além da marcação da estrutura, alguns itens contêm propriedades, o que permite criar subtipos, como por exemplo, capítulos, várias seções, identificação da quantidade de palavras, páginas, saudação, título, entre outros. Esses subtipos são criados de acordo a necessidade de cada usuário e do documento que está sendo editado.

3.2 O E-CORP

Com o crescente desenvolvimento da construção de banco de dados eletrônico, 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 *corpus*, já que será possível filtrar as informações sobre o documento a partir dos dados do metadados de cada documento (LEAL, 2016).

A verificação da ferramenta, enquanto objeto de uso da linguística², foi feita desde o início de seu desenvolvimento em parceria com o projeto CE-DOHS. O objetivo dessa verificação foi obter as informações necessárias para adequar a ferramenta às necessidades linguísticas do projeto, tanto nas buscas, quanto nos resultados exibidos. Ainda no âmbito do CE-DOHS, foi feita a validação com os testes de exibição dos dados linguísticos e extralinguísticos retornados pela busca realizada com a ferramenta.

Quanto às funcionalidades iniciais do E-Corp, essas foram discutidas entre os pesquisadores do CE-DOHS a fim de atender a uma agenda interna projeto e algumas foram escolhidas para o desenvolvimento inicial da ferramenta (LEAL, 2016)³:

- i. Consulta em todos os corpora contidos no CE-DOHS;
- ii. Os filtros utilizados seriam os dados extralinguísticos presentes em cada um dos documentos;
- iii. Permitir a inserção de novos dados para serem utilizados como filtro;
- iv. Os dados seriam retornados mostrando a ocorrência apenas do *corpus* que foi selecionado;
- v. Retorno da:
 - a) linha;
 - b) sentença;
 - c) palavra modernizada;
 - d) palavra original;
 - e) quantidade de vezes que a palavra aparece dentro do acervo pesquisado.

Como exemplo de uso da ferramenta, realizou-se a busca pela palavra *Excelência* no *corpus Cartas para vários destinatários*⁴, um dos acervos do CE-DOHS, composto por 208 cartas escritas por 114 remetentes. Como resultado, para essa busca foram 164 ocorrências, ainda retornando todas as formas de escritas e abreviaturas contidas no *corpus*, como apresentado na *Figura 2*. Utilizando o método de busca tradicional em um banco de texto com esse volume de cartas, a tarefa se tornaria árdua, pouco ágil e

² Um dos exemplos de uso em pesquisas pode ser visto em Tuy Batista (2016), disponível em: < <http://www.tycho.iel.unicamp.br/~tycho/pesquisa/monografias/DISSERT-PRISCILA-2017.pdf> >

³ Para acessar o trabalho na íntegra, consultar Leal (2016).

⁴ A descrição completa do acervo pode ser consultada em Carneiro (2005).

propicia a erros. Utilizando o E-Corp, a busca pela palavra foi feita em pouquíssimo tempo e a confiabilidade dos dados retornados é função apenas da qualidade do corpus em si.

Figura 2: Retorno da palavra *Excelência* no documento

Cartas para vários destinatários (1809-1904)						Total de Ocorrências: 164
Carta	Data	Local	Remetente	Destinatário	Nascido(N)/Radicado(R)	Palavras
01-ARAB-13-12-1829	13 de Dezembro de 1829	Rio de Janeiro	Antonio Rodriguez de Araujo Basto	Senhor Manoel Ignacio da Cunha e Menezes	undefined	271
Modernizado	Original	Sentença				Linha
Excelência	Exa	A sua carta de 6 do mes p pssso me deo grande saptisfação pr trazer-me não só a noticia da sua feliz viagem , como a de ter achado com saúde toda a sua Familia , á qm rendo os meus respeitos , q igualmte são derigi= dos pr ma mulher , a ql agradece os cumprimtos deV Exa , dando-lhe o paraben de se – achar restituído ao seio da sua cara Familia , sendo n'estes sentimentos acompanhada pr meu sogro , e sogra l , q mto se – recomendão .				9
Excelencia	Exa	Dezejando á V Exa saúde , e ventu= ras passo á sollicitar com instancia q me- empregue no seu serviço , pois sempre me – acha prompto pr ser				13
Excelência	Exa	Rogo á V Exa me – recomende aos Exmos Snres Telles , e Anto e Augusto .				17
Excelencia	Exa	De V Exa				20

Fonte: próprios autores.

A busca feita pode ser personalizada usando os seguintes filtros:

1. Acervo;
2. Tipo de Documento;
3. Autor;
4. Sexo;
5. Nascido/Radicado;
6. Destinatário;
7. Local de Escrita;
8. Ano do Documento.

Esses filtros permitem a criação de um *corpus* baseado na necessidade de cada pesquisador. Por exemplo, o pesquisador poderia querer levantar todas as ocorrências de *Excelência* produzidos por *mulheres* em todos os acervos do CE-DOHS. Realizando uma busca refinada ao utilizar os filtros, o pesquisador conseguiria encontrar a quantidade de ocorrências exata por acervo e no geral, como será detalhado adiante.

É inegável a existência de diversas ferramentas que auxiliam nos estudos linguísticos, desde a edição até a etiquetagem morfológica e sintática de todo o texto. Como é mostrado por Costa (2015), as diversas funcionalidades dessas ferramentas podem ser divididas em: (i) contadores de frequência – que são as ferramentas que contam a quantidade de vezes que determinada palavra aparece; (ii) buscadores de concordância – que disponibilizam uma forma de busca em que o usuário procura uma determinada palavra dentro de um *corpus*; (iii) buscadores de colocação – a ferramenta retorna as organizações de determinadas palavras seguindo um padrão. Podemos acrescentar que isso vale não apenas para palavras, mas para outros fenômenos

linguísticos (p.e., classes de palavras, construções sintáticas, papéis semânticos etc.), a depender do grau de anotação do corpus.

Dessa forma, as ferramentas que mais se assemelham a ferramenta apresentada são as ferramentas que fazem parte dos grupos (i) e (ii), pois são ferramentas que além de retornarem à quantidade de ocorrências, retornam um recorte do documento de onde a palavra foi encontrada.

A ferramenta⁵ do projeto *Post Scriptum* se assemelha a ferramenta proposta, porém, não atende à demanda interna do projeto CE-DOHS. O projeto utiliza um recurso web para que quaisquer usuários possam fazer buscas dentro de seus acervos, filtrando de acordo com a sua necessidade e também conta com um acervo de documentos editados no formato XML.

3.2.1 Uso da ferramenta *E-Corp*

Para facilitar o primeiro contato do usuário com a ferramenta, ao abrir a janela de busca é apresentada uma imagem indicando os passos necessários para que a busca possa ser feita. A imagem indica três passos: (1) seleção de filtro; (2) palavra a ser pesquisada e (3) pesquisar.

Na *Figura 3*, é apresentada a janela de busca do E-Corp no site do CE-DOHS, sendo possível conferir todos os campos que podem ser preenchidos para filtrar a busca, inclusive o campo exclusivo para uma palavra (um dado) específica a ser buscada. Para facilitar o preenchimento pelo usuário, alguns campos oferecem a lista de opções disponíveis na base de dados do acervo.

Figura 3: Janela de busca

Feira de Santana - BA, Brasil
Corpus Eletrônico de Documentos Históricos do Sertão [CE-DOHS]

Corpora Projeto
Cartas brasileiras Outros manuscritos Impresso Oral Apresentação Participantes Produção Contato Créditos Links Busca

Dados para Busca

Acervo: Todos Nascido(N)/Radicado(R):
Tipo do Documento: Todos Destinatário:
Autor: Local de Escrita:
Sexo: Ambos Ano do Documento:
Palavra a ser buscada: Pesquisar

Patrocínio e apoio: UFFS fapesb 10 anos GOVERNO DA Bahia Secretaria de Cultura, Saberes e Memória CNPq 60 ANOS

* Ferramenta desenvolvida por Igor Leal (email).

Universidade Estadual de Feira de Santana
Departamento de Letras
e Artes
Núcleo de Estudos em Língua Portuguesa (NELP)
Av. Transoceanista, s/n -
Nova Heliópolis
Feira de Santana-BA, Brasil
Casa Postal: 252 e 204
CEP: 44.056-900
TEL: (75)3161-8265
Projeto Vozes do Sertão em Dados: história, poesia e jornalismo de parangolé brasileiro
CNPq, Projeto: 40143/2008-9 - Conselho 102/2009 (processo)

Fonte: próprios autores.

⁵ URL de acesso: <http://ps.clul.ul.pt/index.php?action=cqp&act=advanced>

Após fazer a pesquisa, o usuário é redirecionado para uma nova janela, onde o resultado da busca é exibido, como pode ser visto nas *Figuras 4* e 5. Os dados são exibidos para o usuário por documento, nesse caso, por carta, conforme apresentado na *Figura 4*. Além de indicar a quantidade de vezes que a palavra buscada foi encontrada no documento, é disponibilizado, também, o resultado com a quantidade total de vezes que essa mesma palavra aparece em todo o acervo.



Figura 4: Dados retornados da busca

» VOLTAR PARA BUSCA						
Total de Ocorrência(s) no(s) Acervo(s) Pesquisado(s):						40
Cartas para vários destinatários (1809-1904)						Total de Ocorrências: 13
Carta	Data	Local	Remetente	Destinatário	Nascido(N)/Radicado(R)	Palavras
32-1C-30-03-1837	30 de Março de 1837	Paris	Cansação [João Lins Vieira de Cansação de Sinimbu]	Angelo Muniz da Silva Ferraz (futuro baía de	São Miguel dos Campos, capitania de Alagoas	1218
Modernizado	Original	Sentença				Linha
tu	tu	Ensina meu nome a teu filhinho , e tu mmo lembra-te do teu amigo				65

Fonte: próprios autores.

Nos casos em que o pesquisador deseja fazer uma busca em todos os acervos, é exibida a quantidade total de retornos discriminados por acervo e por documento, como exposto na *Figura 5*:

Figura 5: Janela de exibição dos dados

 Feira de Santana - BA, Brasil Corpus Eletrônico de Documentos Históricos do Sertão [CE-DOHS] Composição: <input type="text"/> Projeto: <input type="text"/> Cartas brasileiras Outros manuscritos Impresso Oral Apresentação Participantes Produção Contato Créditos Links							 Universidade Estadual de Feira de Santana Departamento de Letras e Artes Núcleo de Estudos de Língua Portuguesa (NELP) Av. Transodestina, s/n - Novo Horizonte Feira de Santana-BA, Brasil Caixa Postal: 252 e 294 CEP: 44.036-900 +55 75 31610265 Projeto Vozes do Sertão em Dados: história, poesia e Arquivo de pesquisa literária (CNPq, Processo 401423/2008-9 / Conselho 102/2009) (ANEXOS)
» VOLTAR PARA BUSCA							
Total de Ocorrência(s) no(s) Acervo(s) Pesquisado(s):						40	
Cartas para vários destinatários (1809-1904)						Total de Ocorrências: 13	
Carta	Data	Local	Remetente	Destinatário	Nascido(N)/Radicado(R)	Palavras	
32-1C-30-03-1837	30 de Março de 1837	Paris	Cansação [João Lins Vieira de Cansação de Sinimbu]	Angelo Muniz da Silva Ferraz (futuro baía de	São Miguel dos Campos, capitania de Alagoas	1218	
Modernizado	Original	Sentença				Linha	
tu	tu	Ensina meu nome a teu filhinho , e tu mmo lembra-te do teu amigo				65	
57-1FOAR-15-02-1866	15 de fevereiro [18]66	Buenos Ayres	F. Octaviano [Francisco Octaviano de Almeida	Ferraz	Rio de Janeiro	687	
Modernizado	Original	Sentença				Linha	
tu	tu	A parte que eu tenho nelles , tu comprehendes q he a da inspiração .				87	
61-1FOAR-09-08-1866	9 de agosto [18]66	Corrientes	F. Octaviano [Francisco Octaviano de Almeida	Ferraz	Rio de Janeiro	219	
Modernizado	Original	Sentença				Linha	
tu	tu	Quem he buioso , como tu , respeita os buios dos amigos e não deixa que paire sobre elles a menor sussepta .				6	

Fonte: próprios autores.

Outras informações também são disponibilizadas, como os dados da carta e a palavra buscada na sua forma modernizada e original, a sentença original em que a palavra buscada aparece e a sua respectiva linha.

3.2.2 Desenvolvimento das Funções

Para ter acesso a todos os elementos do texto, o processo de desenvolvimento foi baseado em duas etiquetas do *XML* utilizado no projeto CE-DOHS. As etiquetas utilizadas para a busca foram `<o>` e `<e>`, onde a primeira indica a palavra original, conforme o remetente escreve e a segunda, a palavra editada – seja uma modernização ou uma simples junção, conforme edições previstas. Isso permite que todas as palavras sejam encontradas no texto, como apresentado anteriormente na *Figura 5*, independente da edição realizada sobre ela.

Nesse exemplo, a palavra buscada foi “Exa” e o retorno foi sua forma original, modernizada, a sentença em que aparece, a respectiva linha e o total de ocorrências no acervo. Concluído o mecanismo de busca da ferramenta em um documento, o passo seguinte foi desenvolver o método para realizar a busca em todos os documentos. Foram utilizados os arquivos em *XML* responsáveis por conter todos os acervos que estão disponíveis no CE-DOHS, são os catálogos para o *corpus de impressos*, de *manuscritos*, *oral* e de *outros manuscritos*. Os arquivos, respectivamente, são *corpora_i*; *corpora*; *corpora_o* e *corpora_ot*, e, pode ser visto na *Figura 6*, um exemplo do conteúdo dos arquivos.

Figura 6: Recorte do catálogo do *corpus* manuscrito

```

<!--
  Catálogo dos corpora disponíveis no CE-DOHS.
  Atualizado em 20.06.2012
-->
<catalog>
  <!-- Registro de acervo -->
  <corpus id="CAV" title="Cartas para vários des
Novais. Cartas Brasileiras (1809-1904): um est
Edição fac-similar e semidiplomática compos
Instituto Geográfico e Histórico da Bahia (I
</corpus>

```

Fonte: www.uefs.br/cedohs

Na *Figura 6*, é possível ver a etiqueta *corpus*, que contém duas propriedades *id* e *title*. A quantidade de etiquetas *corpus* é a quantidade de *corpus* pertencentes ao catálogo. Na propriedade *id* uma sigla do acervo é criada e na propriedade *title* é descrito qual o acervo equivalente. Sendo assim, para varrer todos os documentos, inicialmente é necessário varrer esse arquivo para poder ter acesso aos *corpora*. Após o acesso aos *corpora*, com a identificação do acervo, se tem acesso aos catálogos, arquivos que contém as cartas que estão disponíveis online. Cada *corpus* tem seu catálogo de cartas, assim como cada acervo (oral, manuscrito, outros manuscritos e impresso) tem seus *corpora*, sendo possível observar o conteúdo desses catálogos na *Figura 7*.

Figura 7: Recorte de um catálogo de cartas

```
<catalog id="CAV">
  <corpusfile filename="01-ARAB-13-12-1829.xml"/>
  <corpusfile filename="02-1J5L-19-09-1809.xml"/>
  <corpusfile filename="03-1J5L-09-04-1810.xml"/>
  <corpusfile filename="04-1J5L-09-07-1810.xml"/>
  <corpusfile filename="05-1J5L-15-09-1810.xml"/>
  <corpusfile filename="06-1J5L-16-03-1812.xml"/>
  <corpusfile filename="07-1AB-29-04-1825.xml"/>
  <corpusfile filename="08-1AB-01-04-1828.xml"/>
  <corpusfile filename="09-1AB-21-10-1828.xml"/>
  <corpusfile filename="100-1LPCF-02-02-1878.xml"/>
  <corpusfile filename="10-1AB-30-11-1828.xml"/>
  <corpusfile filename="101-D-09-06-1874.xml"/>
```

Fonte: www.uefs.br/cedohs

Para ser possível varrer todos os documentos, os seguintes passos devem ser seguidos:

1. Ter acesso aos quatro arquivos: *corpora_i*, *corpora*, *corpora_o* e *corpora_ot*;
2. Acessar os identificadores de cada *corpus* e abrir o catálogo de cartas referente a cada acervo;
3. Percorrer todo o arquivo acessando o valor da propriedade *filename*, que é o nome da carta disponível.

A busca é realizada através da varredura de todas as cartas, uma por uma, em todos os acervos e comparando palavra por palavra.

4 CONSIDERAÇÕES FINAIS

A partir dos resultados obtidos, conclui-se que a ferramenta E-Corp é uma alternativa viável para a otimização do tempo de pesquisa nesse tipo de banco de documentos eletrônicos, por proporcionar uma busca ágil e garantir confiabilidade nas coletas dos dados.

Finalizada a ferramenta, outros testes foram realizados no âmbito do CE-DOHS para obter o *feedback* dos usuários. O resultado dos testes apontou que a ferramenta pode facilitar a exploração de *corpora*, uma vez que a seleção dos filtros utilizados para a varredura do acervo ficará a critério do pesquisador, permitindo a construção ou análise de um *corpus*. O uso da ferramenta E-Corp pode, ainda, ser expandido para todos os *corpora* eletrônico que utilizem do padrão de linguagem XML gerado pelo eDicator.

Foi possível concluir, ainda, que a ferramenta pode ser incrementada, com o acréscimo de novas funcionalidades e de melhorias na interface. Além da detecção de algumas melhorias para o banco de dados em que o E-Corp foi testado. A saber, as novas funcionalidades e melhorias levantadas foram:

- i. Melhorias no site do CE-DOHS, para que a ferramenta possa ser desenvolvida de forma responsiva, mantendo uma mesma estrutura para todo o site;
- ii. Melhorias na interface, deixando-a mais amigável com os usuários;
- iii. Criação de um tutorial indicando as possibilidades das buscas;
- iv. Possibilidade de gerar gráficos com a quantidade de ocorrências encontradas;
- v. Verificar a confiança dos dados em todos os acervos do CE-DOHS, a partir de uma amostragem.
- vi. Criação de uma opção para gerar os resultados no formato PDF, para que possa ser usado como apêndice;
- vii. Permitir a busca de mais de uma palavra ao mesmo tempo, fazendo com que seja possível observar variações, como, por exemplo, no uso de Tu e Você;
- viii. Permitir a busca de sentenças utilizando expressões regulares, para que possa pesquisar, por exemplo, palavra seguidas ou precedidas de outras;
- ix. Colocar um questionário para que seja possível fazer avaliações a qualquer momento;

Com a conclusão deste estudo, ficou evidente que a computação pode oferecer recursos que contribuam com as pesquisas em diversas áreas, como as pesquisas realizadas no âmbito da linguística, possibilitando a inclusão de diversas áreas no campo das humanidades digitais.

REFERÊNCIAS BIBLIOGRÁFICAS

- CARNEIRO, Z. de O. N. *Cartas brasileiras (1808-1904): um estudo linguístico-filológico*. 2005. 4v. 2.329f. Tese (Doutorado em Linguística) – Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, São Paulo, 2005.
- COSTA, A. S. *WebSinC: Uma Ferramenta Web para buscas sintáticas e morfossintáticas em corpora anotados - Estudo de Caso do Corpus DOViC – Bahia*. 2015. 1v. 190f. Dissertação (Mestrado em Linguística) - Programa de Pós-graduação em Linguística, Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2015.
- CLUL (Ed.). 2014. P.S. *Post Scriptum*. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna. Disponível em: <URL: <http://ps.clul.ul.pt>>. Acesso em: 31 fev. 2018.
- CORPUS CE-DOHS. *Corpus Eletrônico de Documentos Históricos do Sertão*. Disponível em: <www.uefs.br/cedohs>. Acesso em: 10 mar 2018.
- FARIA, Pablo; GALVES, Charlotte. Criando “Bancos de Árvores”: O Sistema de Anotação e o Processo Automático. *Cadernos de Estudos Linguísticos*. Campinas: v. 58, n. 2 p. 299-315, maio/ago./2016. Disponível em

<http://revistas.iel.unicamp.br/index.php/cel/article/view/5133>. Acesso em 25 mar. 2018.

GALVES, Charlotte; ANDRADE, Aroldo Leal de; and FARIA, Pablo (2017, December). *Tycho Brahe Parsed Corpus of Historical Portuguese*. Disponível em: <<http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip>>. Acesso em: 14 mar 2018.

LACERDA, Mariana Fagundes de Oliveira; CARNEIRO, Zenaide de Oliveira Novais; SANTIAGO, H. S. *Corpus eletrônico de documentos históricos do sertão: as cartas de inábeis*. A COR DAS LETRAS (UEFS), v. 17, p. 127, 2016.

LEAL, Igor. *E-Corp - uma ferramenta de busca de dados para fins linguísticos: aplicação em banco de dados de corpus eletrônico*. Monografia (Graduação em Engenharia de Computação). Universidade Estadual de Feira de Santana, 2016.

PAIXÃO DE SOUSA, M. C. A Filologia Digital em Língua Portuguesa: Alguns caminhos. In: BANZA, A. P.; GONÇALVES, M. F. *Patrimônio textual e humanidades digitais: da antiga à nova Filologia*. Évora: Centro Interdisciplinar de História, Culturas e Sociedades da Universidade de Évora (CIDEHUS)/ Fundação para a Ciência e a Tecnologia (FCT).

PAIXÃO DE SOUSA, M. C. *Memórias do Texto*. Texto Digital (UERJ), 2006. v. 1. p. 10. Disponível em: <<http://www.periodicos.ufsc.br/index.php/textodigital/>>. Acesso em: 10 mar 2018.

PAIXÃO DE SOUSA, M. C. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa*, 2014. v. 16. p. 53-93.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. 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.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N. E-Dictor: Uma ferramenta integrada para a anotação de edição e classe de palavras. In: *VI Encontro de Linguística de Corpus*, São Paulo, 2007.

SARDINHA, T. B. Linguística de corpus: histórico e problemática. *D.E.L.T.A.*, São Paulo, 2000. v. 16. n. 2. p. 323-367. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005>. Acesso em: 15 mar 2018.

TUY BATISTA, P, S, E. *O uso de tu e você em cartas baianas pessoais no século xx em reações de simetria*. (a sair). Dissertação (Mestrado em Estudos Linguísticos) – Programa de Pós-Graduação em Estudos Linguísticos, Universidade Estadual de Feira de Santana, Feira de Santana, 2016.

Recebido em: 05/06/2018

Aprovado em: 23/07/2018

Publicado em: 31/12/2018