DESENVOLVIMENTO DE UM MÓDULO COMPUTACIONAL PARA FINS LINGUÍSTICOS

Igor Leal Souza[1]; Zenaide de Oliveira Novais Carneiro[2]

1. Bolsista PIBIC/CNPq, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: igorengcomp@gmail.com

2. Orientador, Departamento de letras e artes, Universidade Estadual de Feira de Santana, e-mail: zenaide.novais@gmail.com

PALAVRAS-CHAVE: ferramenta de busca; linguística histórica; *corpus* eletrônico.

INTRODUÇÃO

É perceptível o crescimento de pesquisadores dedicados à linguística de corpus[1] com o objetivo de constituir amostras diacrônicas para o estudo da história do português brasileiro (PB), sobre a qual ainda há muito o que ser pesquisado. Criado na década de 1990, o Projeto para a História do Português *Brasileiro* (PHPB) conta com vários pesquisadores, em equipes regionais sediadas em universidades de treze estados brasileiros – Alagoas, Bahia, Ceará, Mato Grosso, Minas Gerais, Pará-Oeste, Paraíba, Paraná, Pernambuco, Rio Grande do Norte, Rio de Janeiro, Santa Catarina e São Paulo (https://sites.google.com/site/corporaphpb/).

O projeto *Corpus Eletrônico de Documentos Históricos do Sertão* (CE-DOHS) (www.uefs.br/cedohs), integrante do Núcleo de Estudos da Língua Portuguesa (NELP), tem o propósito de construir um banco de dados eletrônico e, além disso, desenvolver metodologias para formação de grandes bancos de dados eletrônicos para fins linguísticos, visando a otimização no acesso aos documentos históricos do sertão da Bahia através do uso da ferramenta computacional eDictor (PAIXÃO DE SOUSA; KEPLER, 2007) para a edição desses documentos em linguagem *XML (eXtensible Markup Language)*, a qual permite que sejam feitas edições de acordo com as necessidades das ferramentas para análise linguística, ao passo que garante a recuperação da versão original da edição, caso seja necessário. Nesse banco de dados, já são disponibilizados documentos em que os pesquisadores identifiquem as intervenções realizadas pelo editor. Isso garante que as informações linguísticas relevantes sejam mantidas, viabilizando a recuperação dessas informações de acordo com a modalidade de edição escolhida para a visualização do texto.

A importância desde tipo de edição para a construção de banco de dados eletrônico de documentos históricos para fins de pesquisa já é atestada e aplicada por alguns estudiosos:

> 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, 2006).

O banco de dados em questão já se encontra estruturado de forma a atender os requisitos necessários para a aplicação de tal sistema, uma vez que conforme Paixão de Souza, Kepler e Faria (2009):

[1] De acordo com Sardinha (2004, p.3) a linguística de corpus ocupa-se da "Coleta e da exploração de corpora, ou conjunto de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística".

> A especificação da estrutura XML para codificação no E-Dictor vai de encontro a dois objetivos principais:(i) ser o mais neutra possível (em relação ao conteúdo textual codificado) e (ii) atender a necessidades linguísticas e filológicas , em outras palavras, é preciso que a preparação de conteúdo para análises linguísticas seja simples e eficiente, sem que se percam informações relevantes para estudos filológicos. (PAIXÃO DE SOUSA; KEPLER; FARIA, 2009).

Desta forma, a proposta desenvolvida é a continuação do desenvolvimento de um sistema, denominado E-Corp, com o intuito de facilitar buscas de dados para fins linguísticos com confiabilidade e agilidade nas pesquisas. A aplicação do sistema se deu no banco CE-DOHS, o qual atende às especificações para a utilização da linguagem *XML* para a construção de banco de dados. Ainda, foram analisados os resultados obtidos a fim de comparar com os métodos tradicionais de levantamento de dados (manuais) *versus* eletrônico (automático).

Evidencia-se, desta maneira, a importância deste tipo de sistema para auxiliar no levantamento de dados para estudos, principalmente linguísticos. Observa-se que há uma tendência para a criação de bancos de dados eletrônicos, o que faz necessário o desenvolvimento desse tipo de sistema para facilitar a navegação dos pesquisadores interessados. Como já pode ser constatada a utilização de sistema semelhante no banco de dados Post Scriptum[2] (http://ps.clul.ul.pt/pt/index.php?action=home).

MATERIAL E MÉTODOS OU METODOLOGIA (ou equivalente)

O desenvolvimento deste projeto se dará em três etapas:

(i) o levantamento bibliográfico e dos dados a serem estudados e utilizados;
(ii) revalidação e desenvolvimento do sistema;
(iii) análise do sistema.

Devido à interdisciplinaridade, se fez necessária uma bibliografia diferenciada, contendo textos das duas áreas de interesse: a linguística e a computacional. A finalidade disso foi de evidenciar a necessidade de utilização de ferramentas computacionais em estudos linguísticos.

Após o estudo bibliográfico, foi realizada a verificação da estrutura *XML* do banco de dados do CE-DOHS, visto que o sistema, até então, foi focado no *corpus* manuscrito e agora, voltamo-nos às demais modalidades de documentos disponibilizados por ele. Atualmente, o banco contém três acervos com diferentes especificidades – acervos manuscritos, orais e impressos -, e a partir deles daremos continuidade ao estudo proposto.

Foi primordial o levantamento e a revalidação dos termos que foram buscados na estrutura a partir do estudo da estrutura do *XML*, uma vez que estes termos foram os norteadores do desenvolvimento, baseados nas necessidades linguísticas. Deste modo, os termos foram discutidos e acompanhados por pesquisadores da área da Linguística.

Uma vez concluída a validação dos termos a serem utilizados como base no sistema de busca, a próxima etapa consistiu no desenvolvimento e na adequação do sistema. Para isso, foi utilizada a linguagem de computação Java desde do início do desenvolvimento até sua conclusão.

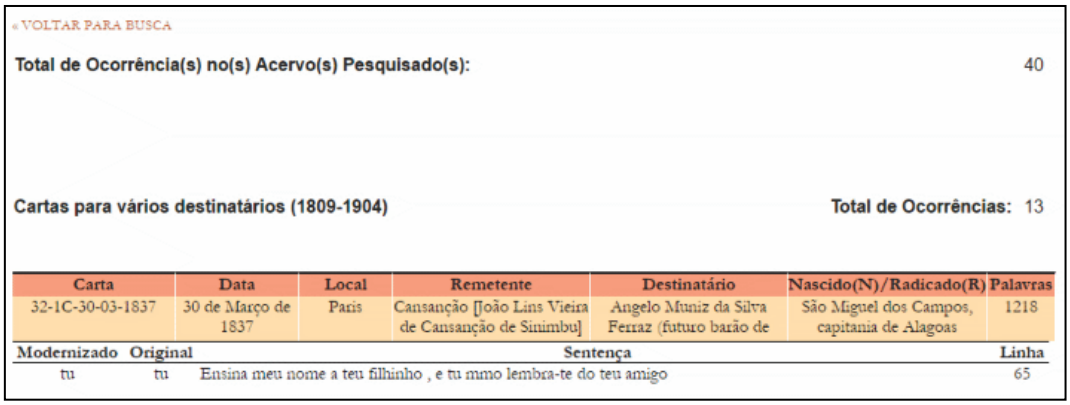
Após o levantamento de dados e o desenvolvimento do módulo -, realizamos um estudo comparativo a respeito do uso do sistema de buscas automáticas, o desenvolvido durante este estudo, e o levantamento de dados manual, a forma tradicional de coleta de dados. Para isso, contamos com a colaboração dos pesquisadores do projeto CE-DOHS, que utilizaram o novo módulo (o sistema de busca de dados automática) e fizeram contribuições

[2] Na apresentação do site, é possível verificar a seguinte descrição: "No projeto P.S. (Post Scriptum), desenvolve-se pesquisa sistemática, edição e estudo histórico-linguístico de cartas privadas escritas durante a Idade Moderna em Portugal e em Espanha. Estes documentos são escritos epistolares inéditos, feitos por autores de diferentes proveniências sociais".

enquanto usuários da ferramenta. Pudemos, assim, proporcionar uma aproximação entre as duas áreas de conhecimento.

RESULTADOS E/OU DISCUSSÃO (ou Análise e discussão dos resultados)

A partir da Figura 1 é possível observar o resultado da busca da palavra *Tu*, uma das palavras utilizadas nos testes. Os dados são exibidos para o usuário a partir desta interface. Além de mostrar os dados extralinguísticos da carta, é exibido a quantidade de vezes que a palavra buscada foi encontrada no documento respectivo, e é, também, disponibilizado o resultado com a quantidade de vezes que essa mesma palavra aparece no acervo.


« 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	Cansanção [João Lins Vieira de Cansanção de Sinimbu]	Angelo Muniz da Silva Ferraz (futuro barão 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

Figura 1 – Exemplo de uma tela de exibição dos dados

Atualmente os resultados obtidos pela ferramenta são:

- busca e contagem da palavra buscada;
- retorno da quantidade de palavras por acervo;
- retorno da quantidade da palavra em todos os acervos pesquisados;
- retorno da sentença em que a palavra aparece;
- retorno do número da linha em que a palavra aparece;
- retorno da palavra "original" (como ela foi escrita pelo autor);
- retorno da palavra "modernizada" (como ela é escrita atualmente);
- retorno dos dados do metadados (dados extralinguísticos) da carta com ocorrência.

Para a validação da ferramenta em campo, foram utilizados dois questionários. Os questionários aplicados foram qualitativos e quantitativos, o que permitiu uma análise mais completa sobre a ferramenta desenvolvida. Os questionários foram aplicados com a finalidade de verificar a utilidade da ferramenta enquanto possibilidade de uso no campo da linguística e obter sugestões para melhoria da interface e da usabilidade, no ponto de vista do usuário. A aplicação aconteceu na sala do CE-DOHS com usuários que já tinham feito pesquisas linguísticas em diferentes ocasiões, desde a iniciação científica até o doutorado. Na figura 2 é possível observar o exemplo de uma das perguntas de um dos questionários.

Figura 2 – Exemplo de uma das perguntas contidas em um dos questionários

CONSIDERAÇÕES FINAIS (ou Conclusão)

A partir dos resultados obtidos neste trabalho, pode-se concluir que a ferramenta E-Corp é uma solução viável para a otimização do tempo para pesquisas realizadas a partir de um banco de dados eletrônico, ou seja, que use esse tipo de base de dados em linguagem XML. Além disso, proporciona uma confiabilidade nas coletas dos dados para pesquisas linguísticas uma vez que garante a varredura completa nos documentos editados. Pode-se, também, perceber que após a análise dos resultados obtidos a partir da aplicação dos questionários, a ferramenta facilitará a exploração de *corpora*, pois a seleção dos filtros utilizados para a varredura do acervo ficará a critério do pesquisador, permitindo a construção ou análise do *corpus*. O uso da ferramenta E-Corp pode, ainda, ser expandido para todos os *corpora* eletrônicos que utilizem do padrão de linguagem XML gerado pelo eDictor.

REFERÊNCIAS

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.

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 jan. 2016.

CORPUS CE-DOHS. **Corpus Eletrônico de Documentos Históricos do Sertão**. Disponível em: <www.uefs.br/cedohs>. Acesso em: 2 dez 2015.

GALVES, C.; FARIA, P. ***Corpus* Histórico do Português Tycho Brahe.** Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/corpus/>. Acesso em: 10 ago 2016.

KROCH, A. Morphosyntactic variation. In: BEALS, K et al. (Ed.). **Papers from the 30th Regional Meeting of the Chicago Linguistics Society**, v.2, 1994, p. 180-201.

KROCH, A. Syntactic change. In: BALTIN, M; COLLINS, C. (Ed.). **The handbook of contemporary syntactic theory**. Oxford: Blackwell Publishers, 2001, p. 699-729.

LIGHTFOOT, D. **The development of language:** Acquisition, change, and evolution. Maryland lectures in language and cognition. Malden.Blackwell, 1999.

PAIXÃO DE SOUSA, M. C. A Filologia Digital em Língua Portuguesa: Alguns caminhos. In: BANZA, A. P.; GONÇALVES, M. F. **Património 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: 2 dez 2015.

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 Lingüística de Corpus**, São Paulo, 2007.

ROBERTS, I. **Verbs and diachronic syntax**. Dordrecht: Kluwer, 1993.

PRESSMAN, R. S. **Engenharia de Software:** Uma Abordagem Profissional. Porto Alegre: Makron Books, 2011.

SOMMERVILLE, I. **Engenharia de Software**. 8ª ed. São Paulo: Pearson Addison Wesley, 2011.