NOVAS MATERIALIDADES DAS FONTES DOCUMENTAIS ANTIGAS: A IMPORTÂNCIA DO TEMPO NO MUNDO DE GRANDES VOLUMES DE DADOS

Aline Silva Costa
Instituto Federal da Bahia (IFBA), Brasil
Endereço eletrônico: alinesilvacosta10@gmail.com

Cristiane Namiuti
Universidade Estadual do Sudoeste da Bahia (UESB), Brasil
Endereço eletrônico: cristianenamiuti@uesb.edu.br

Bruno Silvério Costa
Instituto Federal da Bahia (IFBA), Brasil
Endereço eletrônico: bsilveriocosta@gmail.com

INTRODUÇÃO

Estudos em Linguística Histórica dependem de fontes documentais antigas/históricas. As investigações diacrônicas enfrentam dificuldades e limites quanto ao suporte material desses documentos. No caso do objeto manuscrito com suporte físico em papel, impõe-se a limitação de que o pesquisador deve estar no mesmo espaço físico do documento e deve ter permissão para acessá-lo. A raridade de muitos documentos antigos associada à fragilidade do suporte material traz ainda a complexidade de dificuldade de manuseio e/ou restrições para sua preservação (NAMIUTI; SANTOS, 2017).

As tecnologias digitais trouxeram novos suportes e ferramentas para a exploração de fontes documentais, a exemplo do suporte digital para o texto, e de ferramentas computacionais para pesquisas linguísticas. “O livro digital, apesar de poder ser decodificado de forma direta e visual é uma realidade virtual, codificado computacionalmente [...]” (NAMIUTI; SANTOS, 2017, p. 2). Paixão de Sousa (2006) defende que a versão digital deve manter a fidelidade ao texto original na preparação para o tratamento computacional. Santos e Namiuti (2017) corroboram o requerimento da fidelidade quando postulam que o trabalho de investigação para construir *corpora* deve

buscar a fidedignidade entre o documento físico e sua versão digital com a proposta de *transposição* do suporte físico do livro manuscrito para o livro digital.

As pesquisas com textos antigos no Brasil têm sido intensificadas e conferido destaque à Linguística de *Corpus*, fomentando o desenvolvimento de vários *corpora* digitais de textos da língua portuguesa. Para atender aos objetivos de pesquisa com o dado de língua, os textos precisam ser anotados eletronicamente, um processo que consiste em adicionar novas informações aos textos fontes. As anotações de caráter linguístico podem ser de vários tipos e níveis, representando informações morfológicas/morfossintáticas, sintáticas ou semânticas (NAMIUTI; SANTOS, 2017).

Uma vez que os estudos diacrônicos dependem de fontes antigas em papel, questionamos: Como agregar as tecnologias para anotação dos dados de língua para buscas automáticas, mantendo a fidedignidade do texto original representado no digital, garantindo *performance* com grandes volumes de dados com baixo custo de memória?

Contribuindo para responder tal questionamento, neste trabalho[1], considerando que os recursos computacionais para exploração de *corpora* exercem um papel importante na pesquisa linguística, mostraremos os resultados obtidos do estudo de uma alternativa para anotação morfossintática com a linguagem JSON (*JavaScript Object Notation*), motivados pelo aspecto computacional de desempenho, sem prescindir da fidedignidade ao texto original, requisito necessário ao trabalho com fontes históricas, como uma alternativa à linguagem XML (*eXtensible Markup Language*), uma vez que esta pode apresentar problemas de *performance* para grande volume de dados, além de alto custo de memória (W3SCHOOLS, 2017).

1 Este trabalho vincula-se aos projetos temáticos financiados pela FAPESB (APP 007/2016 e APP 014/2016) e CNPq (436209/2018-7), pois seus autores são ou coordenador ou pesquisadores dos projetos. Nesse sentido, agradecemos à(s) agência(s) de fomento pelo apoio sem o qual a pesquisa que aqui se apresenta não seria possível. Por se tratar de pesquisa colaborativa envolvendo alunos e professores orientadores e coorientadores, este trabalho também contou com a colaboração/autoria de Jorge Viana Santos; todavia, por conta da limitação de número de autores por trabalho somada a número de trabalhos por autor, expressa nas regras de submissão de trabalhos para o XIII Colóquio Nacional e VI Colóquio Internacional do Museu Pedagógico-UESB, sua contribuição/autoria só pode ser mencionada nesta nota.

METODOLOGIA

Os textos do *corpus* DOViC (*Corpus* de Documentos Oitocentistas de Vitória da Conquista) (SANTOS; NAMIUTI, 2016) são editados e anotados com a linguagem XML, com o sistema de anotação proposto por Paixão de Souza (2006) para anotações morfossintática e de edições. Para este estudo, considerou-se o acesso aos documentos que compõem o *corpus* e o trabalho já desenvolvido por Costa (2015) que usa, na ferramenta *Websinc*, o padrão XML para anotação e busca. O documento intitulado "Carta de Alforria da cabra de nome Sofia", escrita em 1845, recebeu anotação morfossintática e de edições utilizando a linguagem JSON. A anotação foi feita manualmente e o arquivo completo foi armazenado num SGBD (Sistema Gerenciador de Bando de Dados) NoSQL, o MongoDB. No SGBD foram realizados os mesmos tipos de buscas morfossintáticas feitas por Costa (2015) sobre o mesmo documento, utilizando também os mesmos parâmetros. As consultas realizadas testaram funções de Existência, Precedência, Precedência Imediata, Palavra na n-ésima posição das sentenças e Palavras no início ou no fim das sentenças. As funções de buscas por Vizinhança não foram implementadas.

Para avaliar os resultados produzidos pelas buscas morfossintáticas, foi utilizado o método de comparação: compararam-se a saída da ferramenta desenvolvida por Costa (2015), o *WebSinc*, com a saída produzida pelo MongoDB. Os resultados foram comparados verificando o número total de ocorrências para a busca em cada ferramenta e a igualdade das sentenças retornadas. Para a comparação da *performance* entre as anotações, foram feitos testes de desempenho para as buscas de Costa (2015) e as buscas na anotação proposta, contrastando os tempos médios obtidos em cada tipo de consulta. Detalhes técnicos computacionais a respeito dos testes podem ser consultados em Damaceno (2018).

RESULTADOS E DISCUSSÃO

Apresenta-se aqui o resultado de uma proposta de anotação morfossintática e de edições para qualquer *corpus* anotado nos mesmos moldes do DOViC usando o formato JSON. Existe similaridade entre os formatos XML e JSON: ambos são formatos de texto

simples, capazes de representar, no formato tabular, informação complexa e difícil, como a estrutura hierárquica de uma anotação sintática. Enquanto a XML se baseia na sintaxe de etiquetas, atributos e valores, o JSON utiliza um formato mais simples, o formato com pares "chave: valor", em que, para cada valor representado, atribui-se um nome (chave) que descreve o seu significado (W3SCHOOLS, 2017).

Os critérios adotados na elaboração da proposta de anotação JSON e o completo mapeamento de etiquetas XML para pares "chave:valor" são detalhados em Damaceno (2018). A figura 1 mostra o resultado do trecho de um arquivo do *corpus* DOViC*, a "*Carta de Liberdade da Cabra de nome Sofia", com a anotação no formato JSON.

Figura 1 - Trecho do *Corpus* DOViC com anotação morfossintática em JSON.


```
{
  p:[ { _id : "p_1" },
      { s :[    {_id : "s_1"},
                { id: 1, o: "Carta", m: "NPR"},
                { id: 2, o: "de", m: "P",  },
                { id: 3, o: "Liberdade", m: "N" },
                { id: 4, o: "da",m: "P+D-F" },
                { id: 5, o: "Cabra", m: "NPR" },
                { id: 6, o: "de", m: "P"},
                { id: 7, o: "nome",m: "N", "bk":{"t":"1","id":"bk_1"}},
                { id: 8, o: "Sofia", m: "NPR"},
                { id: 9, o: "passada", m: "VB-AN-F"},
                { id: 10, o: "por", m: "P",  },
                { id: 11, o: "Antonio", m: "NPR" },
                { id: 12, o: "Jose", "e":{"t":"mod","c":"José"}, m: "NPR" },
                { id: 13, o: "de", m: "P" },
                { id: 14, o: "Souza", m: "NPR", "bk":{"t":"1","id":"bk_2"}},
                { id: 15, o: "Paes",m: "NPR" },
                ....
                {path: "NPR, P, N, P+D-F, NPR, P, N, NPR, VB-AN-F, P, NPR, NPR, P, NPR, NPR, ADV, NPR, P+D-F,"}
            ],
        }
    ]
},
```


Fonte: Damaceno (2018).

Nos testes para verificar igualdade entre as sentenças retornadas, foram feitas 24 consultas, com 6 funções de busca morfossintáticas. Como resultado dos testes, todas as 24 buscas retornaram as mesmas sentenças e idênticos totais de ocorrências para ambas as anotações. Isso mostra que a alteração da linguagem de anotação (de XML para JSON) não afeta os resultados da busca. Não obstante, com relação ao desempenho computacional das buscas morfossintáticas, a proposta de anotação JSON deste trabalho

se sobressaiu obtendo um tempo de resposta aproximadamente 99% menor do que as buscas nos arquivos XML.

CONCLUSÕES

Com base nos resultados obtidos nesta pesquisa, uma representação morfossintática e de edições de *corpora* utilizando o formato JSON mostrou-se viável para textos anotados nos moldes do DOViC. Os testes mostram que a anotação mantém as possibilidades que a linguagem XML proporciona, uma vez que as buscas realizadas nos arquivos/documentos processados tanto com JSON quanto com XML retornaram as mesmas ocorrências. Os testes de desempenho mostraram melhor resultado na utilização do formato JSON, conferindo-lhe vantagem no aspecto de desempenho, fato que mostra a viabilidade da tecnologia NoSQL nos estudos da Linguística de *Corpus*, no âmbito das Humanidades Digitais.

PALAVRAS-CHAVE: Linguística Histórica; *Corpus*; Anotação; XML; JSON.

REFERÊNCIAS

COSTA, A. S. **WebSinC: Uma Ferramenta Web para buscas sintáticas e morfossintáticas em *corpora* anotados - Estudo de Caso do *Corpus* DOViC – Bahia**. Orientador: Cristiane Namiuti Temponi. 2015. Dissertação (Programa de Pós-Graduação em Linguística) - Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista, 2015.

DAMACENO, R.P. **Representação de um *corpus* linguístico em um banco de dados NoSQL: Estudo de caso do *corpus* DOViC**. Orientador: Aline Silva Costa. 2018. Monografia (Curso de Bacharelado em Sistemas de Informação) - Instituto Federal de Educação, Ciência e Tecnologia da Bahia, Vitória da Conquista, 2018.

PAIXÃO DE SOUSA, M.C. Memórias do Texto. **Revista Texto Digital**, Universidade Federal de Santa Catarina, Florianópolis, n.2, 2006. Disponível em: http://www.textodigital.ufsc.br/num02/paixao.htm. Acesso em: 24 fev. 2019.

SANTOS, J. V.; NAMIUTI, C. **DOViC - Documentos Oitocentistas de Vitória da Conquista. Memória Conquistense**. Vitória da Conquista: UESB/LAPELINC, 2016. Disponível em: http://memoriaconquistense.uesb.br/websinc.Acesso em 24 mar. 2019.

SANTOS, J. V.; NAMIUTI, C. **O objeto livro: a complexidade da forma e o digital**. In: Anais do X Congresso Internacional da Associação Brasileira de Linguística: pesquisa linguística e compromisso político, 7 a 10 de março de 2017, Niterói, RJ /

organizado por Luciana Sanchez Mendes, Nadja Pattresi de Souza e Silva e Silmara Cristina Dela da Silva. – Niterói: UFF, 2017.

W3SCHOOLS. **JSON VS XML**. 2017. Disponível em: https://www.w3schools.com/js/js_json_xml.asp. Acesso em: 17 nov. 2017.