COSTA, Aline Silva; NAMIUTI-TEMPONI, Cristiane. WebSinC: Buscas online em corpora sintaticamente anotados. "E-Book do Congresso de Humanidades Digitais em Portugal: Construir pontes e quebrar barreiras na era digital – 2015". Lisboa: Universidade Nova de Lisboa, 2017 (no prelo).

# WEBSINC: BUSCAS ONLINE EM CORPORA SINTATICAMENTE ANOTADOS

Aline Silva Costa

Cristiane Namiuti

## 1. Introdução

Buscas automáticas por categorias sintáticas ou morfossintáticas em textos de *corpora* anotados têm sido utilizadas como metodologia de obtenção de dados para várias pesquisas na área da Linguística. A utilização de softwares que realizem tais buscas é fundamental, uma vez que permitem a análise de grandes *corpora* com grande volume de dados textuais. Apresentamos neste trabalho[1] o software WebSinC[2], um aplicativo web que fornece o recurso de buscas automáticas com interface gráfica. A ferramenta também provê o recurso de gerenciamento de *corpora* anotados, disponibilizando os textos e as imagens dos manuscritos na Internet, para fins de pesquisa.

O software WebSinC foi criado para gerenciar e disponibilizar os textos do *corpus* digital DOViC (Documentos Oitocentistas de Vitória da Conquista e região)[3] que, formado por documentos manuscritos do século XIX guardados nos arquivos do Fórum de Vitória da Conquista - Bahia, forneceu requisitos para implementação e testes com o software. O WebSinC foi

[1] Este trabalho apresenta resultados de pesquisa relativos aos projetos: Fapesp 2012/06078-9, Fapesb PET0034/2010, CNPq 471753/2014-9, CNPq 485098/2013-0.

[2] Cristiane Namiuti, Jorge Viana Santos e Aline Silva Costa, *WebSinC* (Vitória da Conquista: Universidade Estadual do Sudoeste da Bahia), 2015, http://memoriaconquistense.uesb.br/websinc/. Aplicativo Web idealizado pelos professores coordenadores do Laboratório de Pesquisa em Linguistica de *Corpus* (LAPELINC/UESB), Profª. Drª. Cristiane Namiuti e Prof. Dr. Jorge Viana Santos e programado por Aline Silva Costa como produto de sua dissertação de mestrado: Aline Silva Costa, *WebSinC: Uma Ferramenta Web para buscas sintáticas e morfossintáticas em corpora anotados: Estudo de Caso do Corpus DOViC-Bahia*. (Vitória da Conquista: Universidade Estadual do Sudoeste da Bahia, 2015)..

[3] *Corpus* compilado no âmbito do projeto "Memória conquistense: implementação de um *corpus* digital" (Namiuti e Santos 2013), que dá continuidade ao trabalho iniciado no projeto "Memória Conquistense: recuperação de documentos oitocentistas na implementação de um *corpus* digital" (Santos e Namiuti 2009).

programado para buscas por categorias morfossintáticas e sintáticas em *corpora anotados* nos moldes do *Corpus* Histórico do Português Tycho Brahe (CTB)[4] com todo o esquema de anotação e buscas no padrão XML (*Extensible Markup Language*)[5]. O CTB é um *corpus* digital composto atualmente de textos em português de autores nascidos entre 1380 e 1845, desenvolvido na Universidade Estadual de Campinas (Unicamp).

Tanto o CTB quanto o *Corpus* DOViC são *corpora* de textos antigos. Faz parte da metodologia de trabalho destes *corpora*, a anotação de edições realizadas nos textos para tratamento computacional, com a preservação de características originais para estudos linguísticos e filológicos. A motivação para o desenvolvimento do software WebSinC consiste no fato de que os recursos desenvolvidos e aplicados para a compilação, anotação e busca de dados no CTB não seguem um padrão único de linguagem. A linguagem XML é utilizada para a anotação de edição e para a anotação morfossintática. Já a anotação sintática segue o formato *Penn TreeBank*[6], um outro formato que implica duplicação do texto, perda de informação e uso de outra linguagem para as buscas morfossintáticas. Com o WebSinC, as buscas morfossintáticas podem explorar as potencialidades das anotações de edições no formato XML, possibilitando implementação do retorno de resultados na versão original ou editada, o que até então não era possível com outras ferramentas de busca existentes.

A homogeneidade na linguagem de edição e buscas favorece a criação de recursos padronizados, permitindo reuso de tecnologia, oferecendo mais flexibilidade para as buscas e exibição dos resultados, e independência tecnológica para grupos de pesquisa interessados no *corpus*. A ferramenta WebSinC é aplicável a qualquer *corpus* com esquema de anotação baseado no CTB e não apenas ao *Corpus* DOViC.

[4] Charlotte Galves e Pablo Faria, *Corpus Histórico Anotado do Português Tycho Brahe Parsed Corpus of Historical Portuguese* (Campinas: UNICAMP), 2010, http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html.

[5] XML é uma linguagem de editoração que oferece um formato universal para estruturação de documentos e dados na Web. Proposta pelo W3C (*World Wide Web Consortium*) como uma nova alternativa à linguagem HTML, linguagem dominante na Web, a XML combina extensibilidade, poder e flexibilidade com a simplicidade exigida pela Web. (Antonio M. Silva Filho, *Programando com XML* (Rio de Janeiro: Elsevier, 2004).

[6] O *Penn TreeBank Format* (Formato *Penn TreeBank*) é um esquema de anotação sintática de *corpora* desenvolvido na Universidade da Pensilvânia. O esquema utiliza uma representação arbórea delimitada por parênteses etiquetados. Beatrice Santorini, *Annotation manual for the Penn Historical Corpora and the PCEEC*, 2010, http://www.ling.upenn.edu/hist-corpora/annotation/index.html.

## 2. Metodologias de anotação do *Corpus* Tycho Brahe e do *Corpus* DOViC

Os textos que compõem o CTB passam primeiro pela etapa de transcrição, que é a reprodução do texto original no meio digital. O arquivo é salvo no formato de texto simples (TXT) e em seguida passa pelas fases de edição e anotação morfossintática e sintática, nessa ordem. As etapas de transcrição, edição e anotação morfossintática são realizadas com o auxílio da ferramenta eDictor[7] – editor de marcação extensível XML.

Os textos antigos possuem características gráficas e grafemáticas que dificultam o processamento computacional posterior à etapa de transcrição. Por essa razão, os textos precisam ser editados, mas as características do texto original devem ser preservadas devido à sua importância para estudos linguísticos e filológicos. As edições dos textos, portanto, são anotadas, segundo esquema de anotação proposto por Paixão de Sousa[8], mantendo as informações sobre a interferência realizada e sobre o texto original no mesmo arquivo de anotação morfossintática em formato XML.

Através da ferramenta eDictor os textos do CTB recebem também anotações acerca da estrutura e formatação dos textos, das edições de grafia e segmentação e de informações linguísticas no nível morfossintático. Metadados, tais como nome dos autores e data de seu nascimento e/ou morte, ano de publicação do documento, gênero, dados sobre edição e editores, créditos do trabalho de edição e correção da anotação morfossintática também são inseridos na anotação XML.

Atualmente, não há uma ferramenta com a funcionalidade de recuperação das informações de edições anotadas no arquivo XML de textos do CTB. As buscas automáticas baseadas em categorias morfossintáticas são realizadas em um arquivo de texto com anotação *Part-Of-Speech* (POS) realizada pelo etiquetador POS. Assim, as anotações de edições com intervenções dos editores e versões originais dos textos contidas no arquivo XML gerado pelo eDictor não podem ser exploradas nas buscas. Os resultados com as ferramentas utilizadas pelo

[7] Maria Clara Paixão de Sousa, Fábio Kepler e Pablo Faria,"eDictor: novas perspectivas na codificação e edição de corpora de textos Históricos", 2010, in *Caminhos da Linguística de Corpus*, orgs. Tania M. Shepherd, Tony .B. Sardinha, e Marcia Pinto (Campinas: Mercado de Letras, 2012).

[8] Maria Clara Paixão de Sousa., "Memórias do Texto", *Revista Texto Digital*, n. 2 (2006), http://www.textodigital.ufsc.br/num02/paixao.htm.

CTB trazem apenas sentenças em versões editadas. A ferramenta WebSinC foi programada para realizar as buscas morfossintáticas no arquivo XML, podendo explorar as potencialidades das anotações com retorno de sentenças de resultado em versão original ou editada, minimizando a perda de informação.

A versão atual do programa eDictor (versão 1.0 beta 10) não realiza anotação da estrutura sintática. Tal informação é gerada separadamente utilizando um *parser* que recebe como entrada um arquivo anotado no texto (.txt) no formato POS, com as etiquetas morfossintáticas, e gera como saída um outro arquivo texto no formato *Penn TreeBank.* Os textos recebem anotação morfossintática (categoria POS) e sintática dentro dos moldes propostos pelo *Penn-Helsinki Parsed Corpus of Middle English* (PPCME), cuja proposta sugere que a etiquetagem POS deve preceder o processo de anotação sintática[9].

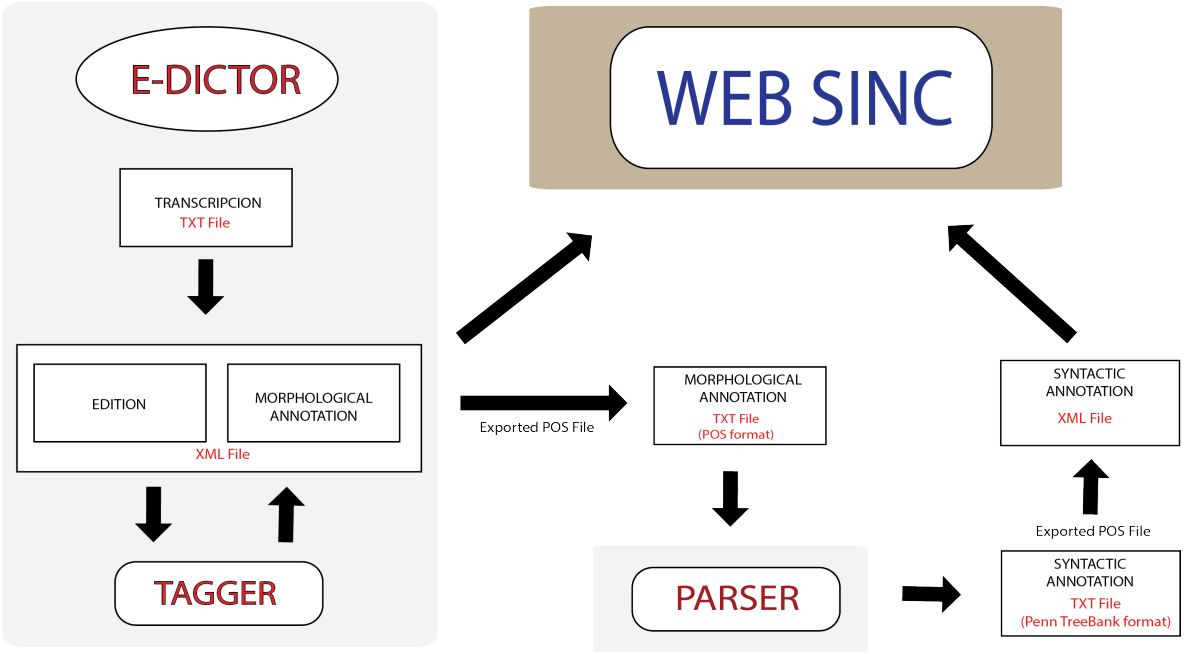
O processo de compilação de corpora realizado para a construção do CTB apresenta, portanto, heterogeneidade de linguagem. Entretanto, na busca por obter homogeneidade de linguagem para reuso de tecnologia, o WebSinC converte a anotação *Penn TreeBank* para o formato XML, reutilizando assim a mesma tecnologia para todos os tipos de buscas. Além do ganho com reuso de tecnologia, é possível obter mais flexibilidade para as buscas e exibição dos resultados, bem como obter independência tecnológica, uma vez que ferramentas de busca para o formato *Penn TreeBank* tornam-se dispensáveis[10].

A figura 1 mostra o resumo dos processos realizados com os textos do *Corpus* DOViC. Toda a parte à esquerda é idêntica aos processos do CTB. Os retângulos brancos representam os processos realizados e o texto em vermelho no interior dos retângulos representa o formato da saída gerada. As ferramentas computacionais utilizadas nos processos estão indicadas (i. eDictor; ii. Tagger incorporado ao eDictor; iii. Parser; iv. WebSinC). As setas indicam a interação e direção entre os processos.

[9]Charlote Galves e Helena Brito. *A Construção do Corpus Anotado do Português Histórico Tycho Brahe – o sistema de anotação morfológica*, 2008, http://www.tycho.iel.unicamp.br/~tycho/pesquisa/artigos/GALVES_Cetal-Fase1a.pdf.

[10]Cristiane Namiuti e Aline S. Costa, "Reflexão sobre anotação sintática e ferramentas de busca – Uso da linguagem XML para anotação sintática no *corpus* digital DOViC" *Letras & Letras,* 30.2 (2014): 82-103.

**Figura 1 - Resumo dos processos no *Corpus* DOViC**




## 3. O software WebSinC

O software WebSinc provê uma interface gráfica de fácil utilização pelo usuário, que pode ser um pesquisador interessado no *corpus*, ou um administrador do sistema que o gerencie, fazendo *upload* e cadastro de documentos digitais imagem (DDI) e documentos digitais texto (DDT)[11], tornando o *corpus* disponível na Internet. Ao pesquisador da Internet estarão disponíveis os documentos do *corpus* para visualização do documento original em imagem digital (fotos do original físico) e também das versões do texto original (transcrita, editada e morfossintaticamente anotada). A ferramenta provê o recurso de buscas baseadas em categorias sintáticas ou morfossintáticas para auxiliar pesquisas na área da Linguística, sem que o usuário aprenda qualquer linguagem de consulta, pois as buscas poderão ser feitas graficamente através de componentes GUI (*Graphic User Interface*) como links, botões e caixas de seleção.

[11]Cristiane Namiuti e Jorge Viana Santos. "New challenges for ancient sources: DOViC experience in the new Historical Linguistics" In: *Congresso de Humanidades Digitais em Portugal: construindo pontes e quebrando barreiras na era digital*, (Lisboa: Universidade Nova de Lisboa, 2015).

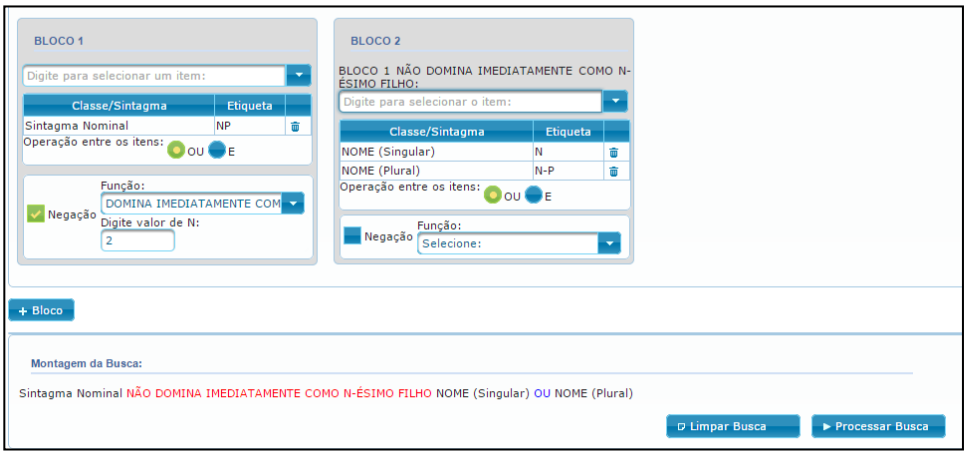
A figura 2 mostra a tela de apresentação do software, onde são exibidas informações gerais e os campos para acesso com *login* e senha. O usuário deve ser registrado para ter acesso ao *corpus*.

**Figura 2 - Tela de login do WebSinC**


### 3.1. Buscas sintáticas no WebSinC

A figura 3 mostra a tela do WebSinC para o recurso de buscas sintáticas. As buscas podem ser feitas pela manipulação dos componentes gráficos na interface.

**Figura 3 - Tela de buscas sintáticas do WebSinC**




Não é necessário que o usuário conheça o conjunto de etiquetas da anotação do *corpus* pesquisado. A busca é construída através da montagem de um conjunto ou *container* de itens, denominado na ferramenta como um "bloco" com categorias de itens lexicais e/ou sintagmas. Os itens do bloco podem ser relacionados pela operação lógica "OU" ou "E". A opção "OU" vem pré-selecionada por *default*. O usuário pode selecionar um item lexical/sintagma usando o recurso de pesquisa do componente (recurso do tipo *autocomplete*), que funciona de maneira que, a cada caracter digitado, um filtro é aplicado retornando apenas itens que se iniciem com os caracteres digitados. Após a seleção do item, o mesmo é inserido automaticamente no bloco.

Feita a montagem do bloco, o usuário deve selecionar uma função de busca. Há funções que requerem apenas um argumento, como é o caso da função "Existência". As outras funções requerem dois ou mais argumentos, e, ao selecioná-las, os componentes da interface são exibidos dinamicamente para seleção. As funções de busca sintática implementadas no WebSinC são: 1) Existência; 2) Precedência; 3) Precedência imediata; 4) Dominância; 5) Dominância imediata; 6) Irmandade; 7) C-Comando; 8) Dominância de N palavras; 9) Dominância de mais de N palavras; 10) Dominância de menos de N palavras; 11) Dominância imediata como primeiro filho; 12) Dominância imediata como último filho; 13) Dominância imediata como N-ésimo filho; 14) Dominância imediata como único filho; 15) Dominância imediata de N filhos; 16) Dominância imediata de menos de N filhos; 17) Dominância imediata de mais de N filhos.

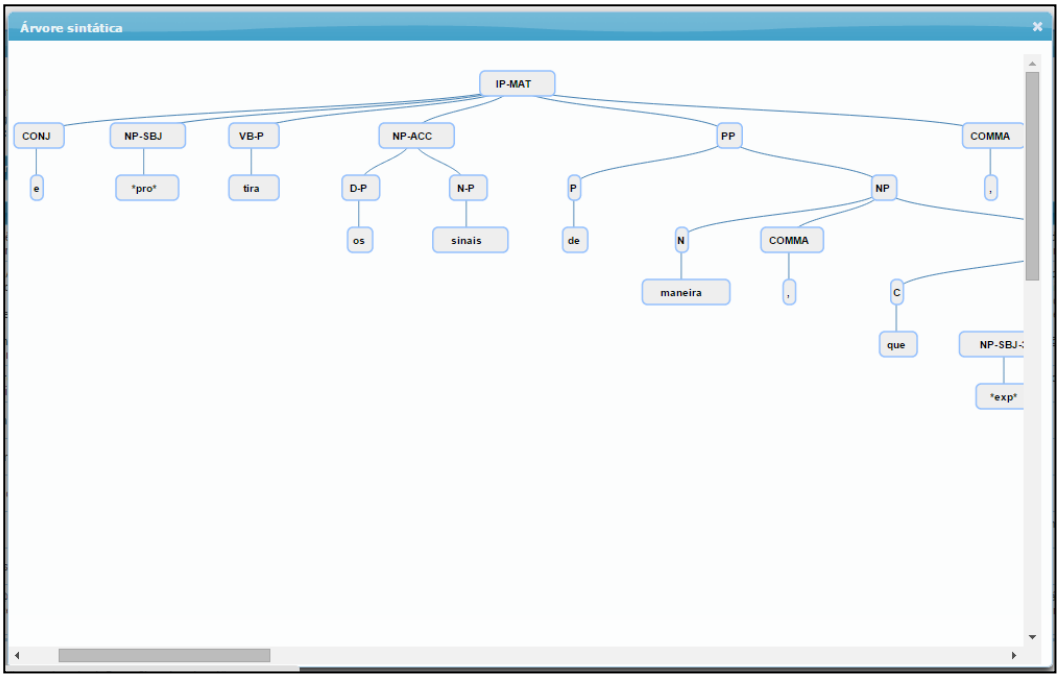
Para realização das buscas sintáticas no WebSinc, é necessária a transformação do arquivo com anotação sintática gerado pelo *parser* em um arquivo XML com representação da estrutura sintagmática. Tal conversão é realizada pelo WebSinC. Sendo assim, o arquivo XML para buscas sintáticas se trata de um outro arquivo diferente e não relacionado ao arquivo com anotações morfossintáticas e de edições do texto gerado pelo eDictor. Isto porque o *parser* utilizado atualmente não pode ser aplicado a um arquivo XML, mas unicamente ao formato POS.

Ao processar a consulta, o resultado da busca é exibido numa tabela, com cada sentença de resultado numa linha. Na tabela também são exibidos um atalho para visualização de sentença no formato gráfico de árvore, além de informações sobre a busca: Texto utilizado para a busca, a consulta realizada e o total de sentenças encontradas. A figura 4 mostra como exemplo a tela de resultado para uma busca sintática realizada com o documento "História da Província de Santa Cruz", do CTB. A busca procurou por sentenças onde um *NP sujeito* e um *sintagma adverbial* tem como irmão na árvore um *verbo estar no tempo passado*.

**Figura 4 - Tela de apresentação de resultados de buscas no WebSinC**


A figura 5 mostra a representação gráfica arbórea de uma das sentenças do resultado da busca.

**Figura 5 - Visualização arbórea de sentença de resultado de busca no WebSinC**




### 3.2. Buscas morfossintática no WebSinC

As buscas automáticas por categorias morfossintáticas no WebSinC também são realizadas através de uma interface gráfica, que segue o mesmo modelo das buscas sintáticas, com a montagem de blocos. A diferença nesta interface de busca é que não existe a exibição de sintagmas. Apenas categorias de itens lexicais são usadas em buscas por categorias morfossintáticas. Assim, os blocos serão compostos apenas por classes de itens lexicais. As funções de busca disponíveis para categorias morfossintáticas também são diferentes, uma vez que esse tipo de busca pressupõe apenas a linearidade e não uma estrutura hierárquica, como nas buscas sintáticas.

As funções de busca morfossintática implementadas no WebSinC são: 1) Existência; 2) Precedência; 3) Precedência imediata; 4) Vizinhança à direita; 5) Vizinhança à esquerda; 6) Início da sentença; 7) Fim da sentença; 8) Posição N da sentença. O arquivo pesquisado na busca morfossintática do WebSinC é o arquivo XML gerado pelo eDictor, com anotações da estrutura dos textos, informações morfossintáticas e anotações de edições.

### 3.3. Tecnologias empregadas no desenvolvimento do WebSinC

A análise e o projeto da aplicação WebSinC foram feitos utilizando a Linguagem de Modelagem Unificada (UML - *Unified Modeling Language*), que obedece aos padrões internacionais de análise e modelagem de software. Para implementação da interface gráfica do sistema foi utilizada a tecnologia *Java Server Faces* (JSF), que se tornou um padrão para construção de interfaces com usuário na Web baseadas em Java. Para programação da lógica do software foi utilizada a linguagem de programação Java. Para implementação de buscas automáticas nos textos do *corpus* foi utilizada a linguagem de consulta XQuery[12], uma linguagem padronizada e recomendada pelo W3C (*World Wide Web Consortium*) para consultas XML. A aplicação utilizou um Sistema Gerenciador de Banco de dados (SGBD) livre com suporte a armazenamento XML, o PostgreSQL[13].

## 4. Conclusão

O desenvolvimento e utilização do software WebSinC com o *Corpus* DOViC trouxeram como resultados o recurso de gerenciamento e a disponibilização do *corpus* na Internet, permitindo também a realização de buscas automáticas para fins de pesquisa. Ressaltamos que o uso de XML para anotação sintática tem a vantagem de reutilizar a mesma tecnologia já utilizada para anotações morfossintática e de edições no *Corpus* DOViC. Como XML é um padrão, usá-lo para todas as representações nos textos do *corpus* favorece a criação de recursos padronizados, permitindo reuso de tecnologia, oferecendo mais flexibilidade para as buscas e exibição dos

[12] XQuery é mais simples de trabalhar e mais fácil de manter do que muitas outras alternativas. É uma linguagem que possui diversas implementações, flexível o suficiente para consultar um amplo espectro de fontes de informação XML, incluindo as bases de dados e documentos (W3C, *XQuery*, 2012, http://www.w3.org/XML/Query/).

[13] O PostgreSQL é um sistema robusto, confiável e largamente utilizado como SGBD de sistemas empresariais e baseados na Web por todo o mundo (Álvaro Pereira Neto. *PostgreSQL. Técnicas avançadas: Versões open source: Soluções para desenvolvedores e administradores de Banco de Dados* (São Paulo: Editora Érica, 2003)).

resultados, e independência tecnológica para grupos de pesquisa interessados em estudos a partir desse *corpus*.

**Referências**

Costa, Aline S. "WebSinC: Uma Ferramenta Web para buscas sintáticas e morfossintáticas em corpora anotados - Estudo de Caso do *Corpus* DOViC – Bahia." Dissertação de Mestrado, Universidade Estadual do Sudoeste da Bahia, 2015.

Galves, Charlotte, coord. "Padrões Rítmicos, Fixação de Parâmetros & Mudança Linguística", 1998. Acessado em 31 de julho de 2014, http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/index.html.

Galves, Charlotte, e Helena Brito. *A Construção do Corpus Anotado do Português Histórico Tycho Brahe – o sistema de anotação morfológica*, 2008, Acesso em 5 de agosto de 2014, http://www.tycho.iel.unicamp.br/~tycho/pesquisa/artigos/GALVES_Cetal-Fase1a.pdf.

Galves, Charlotte, e Pablo Faria. "*Corpus* Histórico Anotado do Português Tycho Brahe", 1998. Acessado em 30 de julho de 2014, http://www.tycho.iel.unicamp.br/~tycho/corpus.

Namiuti, Cristiane, e Aline S. Costa. "Reflexão sobre anotação sintática e ferramentas de busca – Uso da linguagem XML para anotação sintática no *corpus* digital DOViC." *Letras & Letras* 30.2 (2014): 82-103. Acessado em Setembro 15, 2015.

Namiuti, Cristiane, e Jorge Viana Santos, coords. *Memória Conquistense: implementação de um corpus digital*, 2013, CNPq 485098/2013-0, Vitória da Conquista (Projeto de Pesquisa).

Namiuti, Cristiane, e Jorge Viana Santos. "New challenges for ancient sources: DOViC experience in the new Historical Linguistics." Trabalho apresentado no congresso de Humanidades Digitais em Portugal: construindo pontes e quebrando barreiras na era digital, Lisboa, Portugal. Outubro 8-9, 2015.

Namiuti, Cristiane, Jorge Viana Santos e Aline S. Costa. "New challenges for ancient sources: an important dialogue between Computer Science and new Historical Linguistics." Trabalho apresentado no Workshop: The New Historical Linguistics and the World of Annotated Corpora, Campinas, Brasil, Março 9-13, 2015.

Paixão de Sousa, Maria Clara. "Memórias do Texto", in *Revista Texto Digital*, n.2., 2006. Acessado em 5 de agosto de 2014, http://www.textodigital.ufsc.br/num02/paixao.htm.

Paixão de Sousa, Maria.C., Fábio Kepler e Pablo Faria. "eDictor: novas perspectivas na codificação e edição de corpora de textos Históricos", 2010, in *Caminhos da Linguística de Corpus*, Shepherd, Tania .M., e Sardinha, Tony .B., e Marcia Pinto, organizadores, (2012), Campinas: Mercado de Letras.

Pereira Neto, Álvaro. *PostgreSQL. Técnicas avançadas: Versões open source: Soluções para desenvolvedores e administradores de Banco de Dados,* 2003, São Paulo: Editora Érica.

Santorini, Beatrice. "Annotation manual for the Penn Historical Corpora and the PCEEC", 2010. Acessado em 8 de outubro de 2013, http://www.ling.upenn.edu/hist-corpora/annotation/index.html.

Santos, Jorge Viana, e Cristiane Namiuti, coords. *Memória Conquistense: recuperação de documentos oitocentistas na implementação de um corpus digital*, 2009, UESB, Vitória da Conquista, Projeto de Pesquisa.

Silva Filho, Antonio M. *Programando com XML*, 2004, Rio de Janeiro: Elsevier.

W3C, "XML Technology", 2010. Acessado em 8 de outubro de 2014, http://www.w3.org/standards/xml/

W3C, "XQuery", 2012. Acessado em 10 de outubro de 2014, http://www.w3.org/XML/Query/

**Resumo**: Este artigo apresenta o software WebSinc, um aplicativo web de suporte a compilação de *corpora*, com Interface gráfica desenvolvida para o trabalho de registro, armazenamento, disponibilização, visão e busca de dados em *corpora* cientificamente controlados. A ferramenta contempla funcionalidades distribuídas em dois atores que podem interagir com o sistema: o usuário administrador, para gerenciar o *corpus*, e o usuário pesquisador. É permitido ao usuário administrador o registro de informações, o cadastro e upload de documentos digitais imagem (DDIs) e de documentos digitais texto (DDTs), tornando o *corpus* disponível na Internet. Ao pesquisador da Internet, o aplicativo WebSinC, além de disponibilizar documentos e dados do corpus, disponibiliza o recurso de buscas por categorias sintáticas e morfossintáticas em *corpora* anotados nos moldes do *Corpus* Tycho Brahe, com todo o esquema de anotação e buscas no

padrão XML. A ferramenta tem a função de disponibilizar na Internet o *corpus* digital DOViC, permitindo buscas automáticas. A interface gráfica construída amplia o leque de usuários uma vez que seu uso não demanda o aprendizado de qualquer linguagem de consulta, sistema de anotação ou instalação de algum software pelo pesquisador, contribuindo para o avanço científico nas áreas da Linguística e Filologia no contexto das Humanidades Digitais.

**Palavras-chave**: Corpora Anotados. Buscas automáticas. XML.

***Abstract****: This paper presents the WebSinc software, a web application for corpora compilation with graphical interface developed to record, storage, make avaiable, view and search data in scientifically controlled corpora. The tool contain features distributed in two actors that can interact with the system: the administrator user to manage the corpus, and the researcher user: The administrator can record data, register and upload digital document image (DDIs) and digital text documents (DDTS), making the corpus available on the Internet. For the researcher user, in addition to make available documents and data, WebSinC provides the search feature by syntactic and morphosyntactic categories in annotated corpora embased on the Corpus Tycho Brahe, with the annotation scheme and searches in the XML standard. Currently, the tool makes available the DOViC digital corpus in the Internet and allows automatic searches. The graphical interface that we constructed expands the number of users since the use of the tool does not require learning any query language or annotation system and it does not require any software installation by the researcher. So, webSinc is going to contribute to the advancement of scientific areas of Linguistics and Philology in the context of Digital Humanities.*

***Keywords****: Annotated Corpora. Automatic search tools. XML.*

Notas biográficas

**Aline Silva Costa**: Mestre em Linguística (UESB). Professora do Instituto Federal da Bahia (IFBA). Integrante do Grupo de Pesquisa em Linguística de Corpus (GEPELinC/CNPq). Pesquisadora do Laboratório de Pesquisa em Linguística de Corpus (LAPELINC/UESB) e dos projetos FAPESB PET0034/2010, CNPq 485098/2013-0 e FAPESP 2012/06078-9.

**Cristiane Namiuti**: Doutora em Linguística (UNICAMP). Professora do Depto. de Estudos Linguísticos (DELL/UESB) e do Programa de Pós-Graduação em Linguística (PPGLin/UESB). Líder, com Jorge Viana Santos, do Grupo de Pesquisa em Linguística de Corpus (GEPELinC/CNPq) e do Laboratório de Pesquisa em Linguística de Corpus (LAPELINC/UESB). Pesquisadora dos projetos: FAPESB APP0014/2016, APP0007/2016, CNPq 471753/2014-9 e FAPESP 2012/06078-9.