UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA (UESB)
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA (PPGLIN)

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 – BAHIA
2015

ALINE SILVA COSTA

WEBSINC: UMA FERRAMENTA WEB PARA BUSCAS SINTÁTICAS E MORFOSSINTÁTICAS EM *CORPORA* ANOTADOS - ESTUDO DE CASO DO *CORPUS* DOVIC - BAHIA

Dissertação apresentada ao Programa de Pós-Graduação em Linguística (PPGLin), da Universidade Estadual do Sudoeste da Bahia (UESB), como requisito parcial e obrigatório para obtenção do título de Mestre em Linguística.

Área de Concentração: Linguística

Linha de Pesquisa: Descrição e Análise de Línguas Natur

Orientador: Profª. Drª. Cristiane Namiuti

Coorientador: Prof. Dr. Jorge Viana Santos

VITÓRIA DA CONQUISTA – BAHIA

2015

Costa, Aline Silva.

C87w WebSinc: uma ferramenta Web para buscas sintáticas e morfossintáticas em corpora anotados – estudo de caso do corpus DOViC – Bahia / Aline Silva Costa; orientadora: Cristiane Namiuti; coorientador: Jorge Viana Santos – Vitória da Conquista, 2015.
187f.

Dissertação (mestrado) – Universidade Estadual do Sudoeste da Bahia, Programa de Pós-graduação em Linguística, Vitória da Conquista, 2015.
Inclui referências.

1. Análise sintática e morfossintática. - Web. 2. Ferramenta de busca - Web. 3. Linguagem – XML I. Namiuti, Cristiane. II. Santos, Jorge Viana. III. Universidade Estadual do Sudoeste Bahia, Programa de Pós-Graduação Lingüística. IV. Título.

CDD: 469.5

Catalogação na fonte: Elinei Carvalho Santana – CRB 5/1026
UESB – Campus Vitória da Conquista – BA

Título em inglês: WebSinc: a Web tool for annotated syntactic and morphosyntactic searches in corpora - corpus case study DOViC - Bahia

Palavras-chave em inglês: Annotated Corpora. Search Tools. XML. Syntax.

Área de concentração: Linguística

Titulação: Mestre em Linguística

Banca examinadora: Profa. Dra. Cristiane Namiuti Temponi (Presidente-Orientadora); Prof. Dr. Jorge Viana Santos (Coorientador-UESB); Profa. Dra. Adriana Stella Cardoso Lessa de Oliveira (UESB); Prof. Dr. Leonel Figueiredo de Alencar Araripe (UFC).

Data da defesa: 29 de abril de 2015

Programa de Pós-Graduação: Programa de Pós-Graduação em Linguística.

ALINE SILVA COSTA

WEBSINC: UMA FERRAMENTA WEB PARA BUSCAS SINTÁTICAS E MORFOSSINTÁTICAS EM *CORPORA* ANOTADOS - ESTUDO DE CASO DO *CORPUS* DOVIC - BAHIA

Dissertação apresentada ao Programa de Pós-graduação em Linguística, Universidade Estadual do Sudoeste da Bahia, como parte dos requisitos para a obtenção do título de Mestre em Linguística.

Data da aprovação: 29 de abril de 2015.

BANCA EXAMINADORA

Profa. Dra. Cristiane Namiuti Temponi (UESB)
(Orientadora)

Prof. Dr. Jorge Viana Santos (UESB)
(Coorientador)

Profa. Dra. Adriana Stella Cardoso Lessa de Oliveira (UESB)

Prof. Dr. Leonel Figueiredo de Alencar Araripe (UFC)

AGRADECIMENTOS

Agradeço primeiramente a Deus, meu principal colaborador nessa trajetória.

À Prof.ª Dr.ª Cristiane Namiuti, pela orientação, paciência, e pela compreensão nos momentos difíceis. Ao Prof. Dr. Jorge Viana Santos, pela coorientação e contribuição neste trabalho. Deixo também aqui meu registro de admiração aos dois por serem excelentes professores e pesquisadores, tendo papel fundamental na realização desta pesquisa.

Aos professores do Programa de Pós-Graduação em Linguística da UESB, pela dedicação e por contribuírem com minha formação acadêmica.

Aos professores Dr.ª Adriana Lessa Oliveira e Dr. Leonel Figueiredo de Alencar, pela participação no exame de qualificação desta dissertação e pelas contribuições ao meu trabalho. Obrigada também pelo aceite do convite para a banca de defesa.

Aos colegas professores de Informática do IFBA, pelo encorajamento.

Ao meu cunhado-irmão Igor, pelas diversas ajudas com o código, especialmente com CSS e SVG.

Ao meu amado esposo Bruno, pelo companheirismo, encorajamento e por estar sempre ao meu lado nesta caminhada. Obrigada por ter fé em mim.

Ao meu filhinho amado Arthur, que sem entender direito o que é esse "tal de mestrado", entendeu muitas vezes que a mamãe precisava estudar e não podia brincar.

Finalmente, agradeço àqueles que me ajudaram de alguma forma com a concretização deste trabalho, com quem compartilhei minhas pequenas conquistas e em quem encontrei suporte em oração e conforto nos momentos de angústia: meus pais, minha irmã Fá, Nícia, Paulo, os amigos Cauhina e Fernando Lacerda. Por todo amor e por acreditarem em mim. Obrigada porque tiveram fé.

RESUMO

As necessidades de quantidade de dados, agilidade e automação, itensificaram a produção de corpora de línguas naturais, computacionalmente trabalháveis, anotados morfológica e sintaticamente, para pesquisas na área de Gramática. Com isso, a ciência Linguística passou a contar com a possibilidade de utilização de recursos para buscas automáticas por categorias sintáticas ou morfossintáticas em textos de *corpora* anotados. 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. No entanto, grande parte das pesquisas que utilizam recursos automatizados para a busca de dados em corpora anotados não contam com ferramentas com interface gráfica, tendo, o pesquisador, que aprender uma linguagem de consulta que exige certo conhecimento de programação para aplicá-la em interface texto. O uso de um software que forneça o recurso de buscas automáticas com interface gráfica facilita o processo de busca, dispensando o aprendizado de comandos ou linguagens de consulta pelo linguista, contribuindo, desta maneira, com os estudos gramaticais, sobretudo da área de sintaxe. Consideramos que um esquema de anotação linguística baseado em padrões, como a linguagem XML (*Extensible Markup Language)*, aliado a um aparato tecnológico para essa mesma linguagem, propicia mais flexibilidade às buscas, além de reuso e independência de tecnologias. Nesse contexto, o presente trabalho teve como objetivo o desenvolvimento de um sistema web de buscas morfossintáticas e sintáticas, denominado de WebSinC, para ser utilizado em corpora digitais com anotação XML baseados na metodologia do Corpus Tycho Brahe, seguido de aplicação e testes no *corpus* digital DOViC. O software provê também o gerenciamento e a publicação do *corpus*, disponibilizando-o na Internet para pesquisadores interessados. A metodologia de pesquisa utilizada no trabalho caracteriza-se como pesquisa aplicada. O WebSinC foi modelado utilizando-se da Linguagem de Modelagem Unificada (UML) e sua implementação utilizou a linguagem de programação Java e o *framework Java Server Faces* (JSF). O banco de dados utilizado no software foi o PostgreSQL. Os testes das buscas sintáticas e morfossintáticas implementadas no software foram realizados utilizando-se como dados uma carta do *corpus* DOViC, entitulada *Carta de Alforria da cabra de nome Sofia*, escrita em 1845, e um texto do *corpus* Tycho Brahe, escrito em 1502 por Pero Magalhães de Gandavo. Os testes foram realizados comparando os resultados do sistema WebSinC com os resultados produzidos pela ferramenta de busca *Corpus Search*, já utilizada em muitas outras pesquisas. Foi possível demonstrar a adequação dos resultados das buscas produzidos pelo WebSinC aos resultados esperados e/ou a igualdade com os resultados produzidos pelo *Corpus Search*. A utilização da

linguaguem XML para todo o esquema de anotação e buscas conferiu maior possibilidade de recuperação de informação dos textos, explorando potencialidades de extração de dados em diferentes versões nas buscas, contribuindo assim para a possibilidade de garantia de fidedignidade das versões e controle das edições dos documentos. Também foi demonstrada a aplicabilidade da ferramenta em pesquisas realizadas em corpora anotados, dando exemplos de buscas automáticas que poderiam ser feitas com este recurso do WebSinC, o que leva à conclusão de que o WebSinC é uma ferramenta singular que trará possibilidades que até então não haviam sido exploradas no mundo dos corpora anotados para a pesquisa linguística.

PALAVRAS-CHAVE

Corpora Anotados. Ferramentas de Busca. XML. Sintaxe.

ABSTRACT

Recent research in the Grammar area requires amount of data data that can be retrieved automatically. This need intensified the production of annotaded corpora of natural languages. Thus, Linguistics science has, nowadays, the possibility of using resources for automatic searches for syntactic or morphosyntactic categories in annotated corpora. The use of software to perform such a search is crucial, since they allow the analysis of large corpora, with a large amount of textual data. However, much of the research that use automated tools for data search in annotated corpora do not have tools with graphical interface, and the researcher needs to learn a query language and some knowledge in computer programming. Using a software that provides on automatic searches with graphical interface facilitates the search process, eliminating the learning commands or query languages by linguist, and thus can contribute to the grammatical studies, especially the syntax area. The linguistic annotation scheme based on standards such as XML (Extensible Markup Language), combined with a technological apparatus for the same language, provides more flexibility to search, and reuse and independence technologies. In this context, this study aimed to develop a web system with a search tool for morphosyntactic and syntactic categories, that we called WebSinC. This web system can be used in any digital corpora with XML annotation based in Corpus Tycho Brahe methodology, followed by application and testing in digital corpus DOViC. The software also provides management and publication of the corpus, making it available on the Internet for interested researchers. The research methodology used in the work is characterized as applied research. The WebSinC was modeled using up the Unified Modeling Language (UML) and its implementation used the Java programming language and the Java Server Faces framework (JSF). The database software was used in PostgreSQL. The tests of syntactic and morphosyntactic search implemented in software have been conducted using data as a letter from DOViC corpus, entitled *Carta de Alforria da cabra de nome Sofia*, written in 1845, and a text of Corpus Tycho Brahe, written in 1502 by Pero Magellan Gandavo. Tests were performed comparing the results of WebSinC system with the results produced by the search engine search corpus already used in many other research. It was possible to demonstrate the adequacy of the search results produced by WebSinC the expected results and / or equal to the results produced by the Corpus Search. The use of XML linguaguem for all annotation scheme and searches has increased the possibility of recovering information from texts, exploring data extraction capabilities in different versions in searches, thus contributing to the reliability of the control versions and editions of the documents. It was also demonstrated the applicability of the tool in

research in annotated corpora, giving examples of automated queries that could be made with this feature WebSinC, which leads to the conclusion that the WebSinC is a unique tool that will bring possibilities that until then had not been explored in the world of annotated corpora for linguistic research.

KEYWORDS

Annotated Corpora. Search Tools. XML. Syntax.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURA E SIGLAS

AHDig	Associação das Humanidades Digitais
BNC	British National Corpus
CE-DOHS	Corpus Eletrônico de Documentos Históricos do Sertão
CES	Corpus Encoding Standard
CORDIAL-SIN	Corpus Dialectal para o Estudo da Sintaxe
CRPC	Corpus de Referência do Português Contemporâneo
CTB	Corpus Tycho Brahe
DD	Documento Digital
DF	Documento Físico
DOViC	Corpus de Documentos Oitocentistas de Vitória da Conquista
FLWOR	for, let, where, order by, return
GUI	Graphic User Interface
HTML	Hiper Text Markup Language
IcePaHC	Parsed Historical Corpus
JSF	Java Server Faces
NILC	Núcleo Interinstitucional de Linguística Computacional
OANC	Open American National Corpus
OCR	Optical Character Recognition
PFC	Pesquisador Formador de Corpora
PHPB	Projeto Para a História do Português Brasileiro
PLN	Processamento de Linguagem Natural
POS	Part-of-Speech
PPCEME	Penn-Helsinki Parsed Corpus of Early Modern English
PPCMBE	Penn Parsed Corpus of Modern British English
PPCME2	Penn-Helsinki Parsed Corpus of Middle English
PTB	Penn TreeBank
SGDB	Sistema Gerenciador de Banco de Dados
TEI	Text Encoding Initiative
UML	Linguagem de Modelagem Unificada
W3C	World Wide Web Consortium
XCES	Corpus Encoding Standard for XML
XML	Extensible Markup Language

XPath XML Path

SUMÁRIO

1 APRESENTAÇÃO

Baseando-se nas soluções técnicas para a edição especializada de textos antigos em meio eletrônico, propiciadas pela intensificação do trabalho com textos antigos no âmbito da Linguística Histórica e de *Corpus*, o presente trabalho apresenta um estudo de ferramentas para análise e exploração de *corpora,* realizado no âmbito da pesquisa de mestrado que culminou nesta dissertação e no desenvolvimento de uma ferramenta web, a qual denominamos WebSinC. A pesquisa foi motivada por lacunas identificadas em trabalhos com dado de língua nesse novo suporte do texto, especialmente no que se refere às possibilidades de investigação e busca de dados propiciadas pela criação e implementação de analisadores automáticos empregados nos textos do *Corpus* Tycho Brahe (GALVES, 1998): o *tagger* - software de anotação morfológica - integrado à ferramenta E-Dictor (PAIXÃO DE SOUZA; KEPLER; FARIA, 2010) e o *parser* - ferramenta de anotação sintática de *corpora* desenvolvido na Universidade da Pensilvânia (SANTORINI, 2010; MARCUS; TAYLOR, 2002).

O projeto Tycho Brahe para a construção de grandes corpora anotados é resultado do trabalho pioneiro, no Brasil, de Charlotte Galves (UNICAMP) e colaboradores, sendo o *Corpus Anotado do Português Histórico* o principal produto do projeto temático *Padrões Rítmicos, Fixação de Parâmetros e Mudança Linguística* (1998-2010), financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e coordenado pela professora. No entanto, os recursos desenvolvidos e aplicados para a compilação, anotação e busca automática de dados desse corpus não seguem um padrão de linguagem. A tecnologia XML é utilizada para a edição das versões dos textos do *corpus* (original e modernizada), para a anotação morfossintática automática, e, também, para a ferramenta de correção da anotação morfossintática acoplada no editor XML E-Dictor. Já a anotação sintática segue o padrão *Penn TreeBank.* O *parser* requer um arquivo no formato texto com anotação morfossintática, o qual precisa ser gerado do arquivo XML, fato este que tem como conseqüência a duplicação do texto e perda de informação. Como *output* do *parser* se tem um novo arquivo texto com a anotação sintática, ou seja, uma nova duplicação do texto. Para a correção da anotação sintática e para a busca automática no formato *Penn TreeBank*, foram desenvolvidas, na Universidade da Pensilvânia, as ferramentas *Corpus Search* e *Corpus Draw*.

O software WebSinC foi programado como recurso para buscas automáticas baseadas em categorias morfossintáticas e sintáticas para ser utilizado em *corpora* digitais anotados com a mesma metodologia do Corpus Tycho Brahe (CTB). A ferramenta segue uma homogeneidade na linguagem de anotação e tecnologia de buscas, com todo o aparato tecnológico e anotações

linguísticas baseados no padrão XML. Nossa questão e motivação para investigar uma alternativa para acoplar a anotação sintática ao formato XML consiste em uma desvantagem de se utilizar o formato *Penn TreeBank*, uma vez que no corpus Tycho Brahe ou noutro com a mesma metodologia, XML é a tecnologia já utilizada para anotações morfossintática e de edições no *corpus*. 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 resultados, e independência tecnológica para grupos de pesquisa interessados neste *corpus*.

Neste trabalho, o WebSinC foi aplicado no corpus DOViC (Documentos Oitocentistas de Vitória da Conquista), que forneceu requisitos para implementação e testes com o software. O corpus DOViC é um *corpus* eletrônico de documentos manuscritos dos séculos XVIII e XIX guardados nos arquivos do Fórum de Vitória da Conquista - Bahia, compilado e anotado nos mesmos moldes do corpus Tycho Brahe. A ferramenta também foi aplicada para gerenciamento e disponibilização na Internet dos textos deste *corpus*.

No capítulo introdutório desta dissertação, são discutidas questões referentes à linguística computacional e a pesquisa em *corpora* anotados.

No segundo e terceiro capítulos, são elencados alguns *corpora* eletrônicos disponíveis e analisadas algumas ferramentas para trabalhar com compilação, análise e pesquisas em *corpora*. Padrões para anotação e formatos de anotação também são apresentados, além de uma breve abordagem sobre a linguagem XML e aspectos teóricos de morfologia e sintaxe.

No capítulo quarto, é explicitada a metodologia utilizada no *corpus* Tycho Brahe, tratando das ferramentas tecnológicas e sistema de anotação utilizados, recursos nos quais o *corpus* DOViC se embasa para sua compilação.

O *corpus* DOViC é apresentado no capítulo quinto e o capítulo sexto traz alguns artefatos produzidos no projeto e desenvolvimento da ferramenta WebSinC, bem como as suas funcionalidades pretendidas e telas dos requisitos funcionais implementados.

Resultados e discussões são apresentadas no sétimo capítulo. Finalmente, o oitavo capítulo realiza uma conclusão acerca do estudo realizado e também aponta para os desdobramentos que podem ser realizados em trabalhos futuros.

2 DELINEAMENTOS INICIAIS: A LINGUÍSTICA, A COMPUTAÇÃO E A PESQUISA EM *CORPORA* ANOTADOS NO CONTEXTO DAS HUMANIDADES DIGITAIS

A revolução tecnológica do século XX trouxe a informática como o setor das grandes transformações que culminaram na criação e avanço de novas áreas da ciência sob o escopo da computação. O acelerado desenvolvimento da informática, na segunda metade do século XX, propiciou grandes mudanças nos estudos científicos de uma maneira profunda e geral, atuando decisivamente no progresso das ciências e na produção de conhecimento. No caso específico do estudo das línguas naturais, a computação possibilitou o surgimento de novas abordagens a problemas descritivos, teóricos e práticos que antes ou eram impossíveis de serem abordados ou não podiam ser tratados adequadamente pela linguística. O programa científico para a descrição e análise de línguas naturais que melhor representa esta estreita relação entre a Computação e a Linguística é o da Gramática Gerativa.

A Teoria da Gramática Gerativa, postulada por Noam Chomsky na era da informação, influencia e é influenciada pela ciência da computação. Chomsky lança mão de termos computacionais em seus postulados teóricos: O programa minimalista da teoria gerativa postula a existência de uma computação sintática que parte de uma numeração composta de um conjunto de itens selecionados do léxico da língua para formar um par de forma e sentido nas interfaces de produção e percepção dos enunciados linguísticos (CHOMSKY, 1995). Deixando as analogias entre as ciências a parte, os estudos gramaticais ganharam, nos finais do século XX e no século XXI, uma vertente computacional efetiva para as pesquisas .

As possibilidades de investigação com base em grandes amostras de dados mantendo e garantindo controle e rigor nas pesquisas se multiplicaram exponencialmente com a era digital. Também nessa era da tecnologia digital, a modelagem virtual do processamento linguístico é uma realidade e os algoritmos de aprendizagem e as tecnologias de inteligência artificial são bastante explorados.

A necessidade de estudiosos em se apoiar em usos reais da língua para fazerem generalizações ou esboçarem teorias a respeito do funcionamento linguístico levou ao surgimento da linguística de *corpus*. Atualmente, esta área está intimamente ligada ao uso do computador, visto que os *corpora* são conjuntos de dados eletronicamente manipuláveis através de *softwares*. Por serem objetos de tratamento computacional, os *corpora* eletrônicos necessitam de ferramentas computacionais para sua compilação e tratamento (MELLO; SOUZA, 2012).

Estudos em *corpora* eletrônicos estão sendo desenvolvidos no Brasil. Sardinha (2004) destaca os seguintes *corpora* de língua portuguesa: o *corpus* do Núcleo Interinstitucional de Linguística Computacional (NILC), o *Corpus* de Referência do Português Contemporâneo (CRPC), o *corpus* do Projeto Para a História do Português Brasileiro (PHPB), o Banco de Português, o PORTEXT, o Português falado do Ceará, o *corpus* Histórico do Português Anotado Tycho Brahe, entre outros.

Ferramentas computacionais para exploração de *corpora* exercem um papel importante na investigação linguística de *corpus*, pois permitem análises estatísticas e buscas por padrões em grandes volumes de dados. Assim, softwares para análises de *corpora* têm sido desenvolvidos e utilizados como aliados em pesquisas acerca de fenômenos linguísticos. Este trabalho propõe o desenvolvimento de uma ferramenta computacional para gerenciamento, disponibilização ao público e buscas automáticas sobre o padrão XML no *corpus* digital DOViC, podendo ser utilizada para qualquer outro *corpus* compilado nos mesmos moldes do *corpus* Tycho Brahe.

O *Corpus* DOViC (Documentos Oitocentistas de Vitória da Conquista) é um *corpus* digital de documentos manuscritos dos séculos XVIII e XIX guardados nos arquivos do Fórum de Vitória da Conquista-Bahia. Esse *corpus* é compilado no âmbito do projeto temático **"**Sintaxe diacrônica em *corpus* eletrônico: do português pré-clássico às variantes modernas" (NAMIUTTI, 2010), cujo objetivo é realizar pesquisas na sintaxe diacrônica do português e também contribuir com a construção e implementação de *corpora* anotados em meio digital. O trabalho é desenvolvido em parceria com o projeto "Memória conquistense: implementação de um *corpus* digital" (NAMIUTI, 2013), que dá continuidade ao trabalho iniciado no projeto "Memória Conquistense: recuperação de documentos oitocentistas na implementação de um *corpus* digital" (SANTOS, 2009).

Tanto o corpus Tycho Brahe, quanto o DOViC e outros, são desenvolvidos em projetos que estão ligados à Associação das Humanidades Digitais (AHDig), uma rede de

> pesquisadores de diferentes áreas das humanidades e das ciências da informação e da computação, em universidades brasileiras, portuguesas e outras, envolvidos em diversos grupos de pesquisas, projetos e iniciativas ligados às Humanidades Digitais [...] unidos pela língua portuguesa e pela inclusão da perspectiva digital em seus horizontes de pesquisa (ASSOCIAÇÃO DAS HUMANIDADES DIGITAIS, 2013).

Criada por pesquisadores brasileiros e portugueses em 25 de outubro de 2013, a AHDig tem o objetivo de "fortalecer as iniciativas em Humanidades Digitais já ativas no universo dos

falantes do português, e promover novas iniciativas nesse campo". Inicialmente, a rede foi formada por um grupo de vinte e seis pesquisadores e aberta a "novos participantes ligados à reflexão sobre o digital [...], que falem ou conduzam suas pesquisas em português, ou que tenham interesse em investigar as múltiplas esferas da expressão cultural nessa língua" (ASSOCIAÇÃO DAS HUMANIDADES DIGITAIS, 2013).

Os textos do *corpus* DOViC são transcritos, editados e anotados nos mesmos moldes do *Corpus* Tycho Brahe. 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), no âmbito do Projeto "Padrões Rítmicos, Fixação de Parâmetros e Mudança Linguística" (UNICAMP, 1998).

Paixão de Sousa (2007b) aponta que na etapa de transcrição ou digitalização dos textos de um *corpus* – ou seja, na passagem do meio físico para o meio digital – há grande potencial de perda de informações em detrimento da fidelidade ao texto original. O potencial de perda deve-se às especificidades técnicas do meio eletrônico e uma questão central que se coloca no trabalho com textos antigos é que a fidelidade aos textos originais deve ser mantida. Assim, a autora (2006) preconiza que a produção de textos para a construção de *corpora* de língua no meio eletrônico deve fazer uso de um processo que permita a codificação de uma grande variedade de informações em meio digital, de maneira que estas possam ser depois "traduzidas" ou "lidas" por uma programação que gera a apresentação final do texto de modo confiável.

Partindo dessa premissa, no processamento eletrônico, as estruturas variantes, como interferências realizadas pelo editor, precisam ser anotadas e posteriormente traduzidas ou filtradas. No *Corpus* Tycho Brahe utiliza-se o esquema de anotação desenvolvido no Projeto "Memórias do Texto" (PAIXÃO DE SOUSA, 2006), fazendo uso da linguagem XML (*Extensible Markup Language*). A transcrição e edição dos textos são feitas com o auxílio da ferramenta E-Dictor (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010), a qual provê uma interface gráfica que evita o contato direto do editor com a estrutura XML. O *corpus* DOViC segue os mesmos moldes do CTB em relação às etapas de transcrição, edição e anotação da morfossintaxe dos textos e ao uso da ferramenta E-Dictor.

Edições como junção, segmentação e modernização de grafia são feitas por meio da interface gráfica do E-Dictor, produzindo como resultado um arquivo anotado na linguagem XML. O software realiza anotação das informações morfossintáticas dos textos, também no formato XML e ambas as anotações são feitas num único arquivo. As anotações, se processadas por ferramentas computacionais que façam a tradução e filtragem, consistem num importante

recurso para pesquisas linguísticas, permitindo buscas baseadas em categorias morfossintáticas ou a leitura e estudo dos textos em diversas versões (com/sem modernização de grafia).

O sistema de anotação utilizado também provê o suporte à anotação de metadados dos textos, como títulos, autores, localização do original, data de escrita, data de publicação, etc. Outras informações referentes às características físicas dos documentos, como características da capa (cor, material e forro), altura, largura, profundidade, quantidade de folhas e a imagem dos textos originais, se armazenados digitalmente de maneira adequada, e com o suporte à recuperação por meio de recursos computacionais, proveem também um recurso importante para pesquisas diversas, não apenas linguísticas, mas também filológicas e históricas.

2.1 Problema

A disponibilização de *corpora* na Internet traz diversos benefícios tanto à Instituição que possui ou compilou o *corpus* quanto ao usuário pesquisador (SANTOS, 1999). No entanto, para garantir qualidade e confiabilidade do material, é necessário que o acesso aos textos do *corpus* disponível seja controlado. O controle abrange tanto o gerenciamento do *corpus*, permitindo que apenas pesquisadores autorizados realizem upload de arquivos, quanto a restrição de acesso ao mesmo, permitindo que apenas pesquisadores previamente cadastrados e de acordo com o termo de compromisso ou licença de uso façam pesquisas com o *corpus*.

Para satisfação do requisito de acesso controlado, se faz necessário não apenas a disponibilização de arquivos do *corpus* para *download*, mas também a implementação de um conjunto de características adicionais no software que torna o *corpus* disponível para atender à necessidade de controle de acesso ao mesmo.

A disponibilidade de *corpora* na Internet constitui-se mais significativa se agregada a possibilidades de busca automática que auxiliem os pesquisadores em suas investigações. Para atender aos estudos linguísticos é desejável que o *corpus* forneça a possibilidade de fazer buscas automáticas por categorias gramaticais. Com o objetivo de realizar pesquisas linguísticas automáticas, os textos do *corpus* DOViC e outros baseado no CTB são anotados morfossintaticamente utilizando a linguagem XML. Anotações de intervenções realizadas pelo editores durante o processo de transcrição também são realizadas, com o intuito de preservar características originais dos textos, fundamentais para as pesquisas filológicas e que ora são perdidas na fase de preparação para o tratamento computacional (PAIXÃO DE SOUSA, 2006). Todas essas informações são preservadas em um arquivo anotado gerado pela ferramenta E-Dictor.

Atualmente, não há implementação no software E-Dictor ou ainda o desenvolvimento de outra ferramenta com a funcionalidade de recuperação das informações XML que o programa gera. Assim, buscas automáticas baseadas em categorias morfossintáticas nos textos do *corpus* DOViC anotados em XML demandam implementações em software que as possam viabilizar. A ferramenta WebSinC provê tais buscas como parte do conjunto de recursos que implementa, os quais poderão ser compartilhados e utilizados não apenas por pesquisadores que se interessem pelo DOViC mas também em diversos outros *corpora* que usam o mesmo sistema de anotação do *Corpus* Tycho Brahe.

A anotação sintática utilizada pelo CTB e DOViC é feita no formato *Penn TreeBank*, desenvolvido pela Universidade da Pensilvânia. Existem ferramentas computacionais disponíveis gratuitamente para buscas automáticas baseadas em categorias sintáticas neste formato. No entanto, a utilização destes recursos requer do usuário pesquisador o conhecimento da linguagem de consulta utilizada pela ferramenta e também da anotação utilizada, além da sua instalação no computador.

Dessa forma, questiona-se: (i) como disponibilizar o *corpus* DOViC (ou similares com a metodologia CTB) na Internet permitindo a realização de buscas sintáticas ou morfossintáticas, sem que o linguista necessite aprender comandos, linguagens de consulta ou formatos de anotação?; (ii) de que maneira pode-se contribuir para a fidedignidade das versões dos textos de *corpora* desse tipo, o controle das edições e a confiabilidade nos dados extraídos?; (iii) como a ferramenta WebSinC pode contribuir com os estudos gramaticais da língua portuguesa, mais especificamente com o avanço das pesquisas em sintaxe?

2.2 Resolução do Problema/Formulação de Hipótese

XML é uma linguagem que permite descrever qualquer tipo de dado e é um padrão aberto para interoperabilidade e intercâmbio de informações. Dessa maneira, usar XML para todas as representações nos textos de um *corpus* favorece a criação de recursos padronizados, permitindo reuso de tecnologia e intercâmbio de dados do *corpus*, oferecendo mais flexibilidade para buscas e visualização de resultados, além de promover independência tecnológica para grupos de pesquisa interessados.

Baseando-se nessas premissas, neste trabalho defende-se que: (i) a implementação de um software Web para buscas sintáticas e morfossintáticas em *corpora* anotados permite a disponibilização e a manipulação destes para fins de pesquisas linguísticas, eliminando a necessidade do aprendizado de qualquer linguagem de consulta, sistema de anotação ou

instalação de software pelo pesquisador; (ii) a homogeneidade na linguagem de edição, anotação e busca pode contribuir para um maior controle das edições e reutilização de recursos linguísticos; e, (iii) a implementação de um software com funções de buscas por relações estruturais como dominância, precedência e c-comando, em textos portugueses anotados, pode contribuir com pesquisas gramaticais da língua portuguesa e mais especificamente com o avanço de pesquisas em sintaxe.

2.3 Objetivos

2.3.1 Objetivo Geral

O objetivo geral deste trabalho é construir um software Web para buscas sintáticas e morfossintáticas em corpora com metodologias baseadas no Corpus Tycho Brahe, e aplicá-lo no *corpus* eletrônico DOViC, no intuito de auxiliar as pesquisas linguísticas com o *corpus*. A aplicação também deve disponibilizar os textos do *corpus*, tornando-o acessível pela Internet.

2.3.2 Objetivos Específicos

O trabalho tem os seguintes objetivos específicos:

- Implementar uma aplicação baseada na Web integrada a tecnologias de armazenamento e recuperação de informação em textos XML, considerando o sistema de anotação utilizado no *corpus* em estudo.
- Verificar a solução implementada utilizando documentos do *corpus* DOViC e *Corpus* Tycho Brahe como entrada para verificação.
- Realizar buscas sintáticas e morfossintáticas em cartas do *corpus* DOViC utilizando o software construído.

2.4 Justificativa

Pesquisas com textos antigos têm se intensificado no Brasil e a disponibilidade de recursos computacionais para elaboração e exploração de grandes *corpora* de língua confere destaque à Linguística de *Corpus*, aproximando as áreas de Ciência da Computação e Filologia. Essas pesquisas fomentam o desenvolvimento de diversos *corpora* eletrônicos, dentre os quais

muitos pretendem promover os debates e as pesquisas sobre a história do Português Brasileiro (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010).

Diversos *corpora* eletrônicos construídos no Brasil têm sido disponibilizados na Internet para o público leitor em geral, contribuindo para pesquisas acadêmico-científicas em diversas áreas. Na Web, a recuperação do conteúdo de um *corpus* pode ser feita rapidamente, a partir de qualquer lugar, transpondo barreiras geográficas e, consequentemente, diminuindo os custos da realização das pesquisas, em virtude da não locomoção até as fontes originais (GALVES; BRITTO, 2008; SANTOS, 2009; SARDINHA, 2004; CÂNDIDO JÚNIOR; ALUÍSIO, 2008).

Especificamente acerca da disponibilização de *corpora* na Internet, Santos (1999) enfatiza as seguintes vantagens para a instituição que possui ou compilou o *corpus*:

1. A possibilidade de restringir o acesso ao *corpus* de forma que o usuário não receba o *corpus* inteiro no seu computador, se este for um requisito importante;
2. A possibilidade de saber o que os usuários fazem com os *corpora*, monitorando o acesso;
3. A possibilidade de corrigir/melhorar o *corpus* sem que vários usuários usem versões diferentes;
4. A possibilidade de contabilizar o acesso, e o trabalho feito sobre o *corpus*, e portanto, a eventual resposta da comunidade científica e dos parceiros comerciais;

Para o usuário, a autora elenca também os seguintes benefícios:

1. O acesso, de qualquer local com Internet, a um recurso que se encontra fisicamente longe;
2. A minimização dos conhecimentos técnicos necessários para ter acesso ao *corpus*: não é preciso instalar programas, mudar de sistema operacional, aprender uma sintaxe de linguagem de consulta para conseguir fazer pesquisas no *corpus*;
3. A minimização dos recursos tecnológicos necessários (espaço de memória, requisitos de sistema operacional) – basta ter um navegador com acesso à Internet;
4. A existência de um apoio remoto e de uma comunidade de outros usuários pesquisadores com quem pode trocar experiências, resolver problemas e colaborar.

Além das vantagens já citadas, Santos (1999) enfatiza a possibilidade de avaliação de uma dada ferramenta ou teoria em relação a um padrão comum como uma das maiores vantagens obtidas com a disponibilização de *corpora* na Internet. Apesar da relevância em disponibilizar *corpora* na Web, uma vez que a Internet tem se tornado acessível a um número

cada vez maior de pessoas, apenas a disponibilização dos textos não é suficiente. A linguística de *corpus* não consiste apenas na compilação de *corpora* mas também na exploração deles para o avanço da linguística em geral e do processamento e investigação de uma dada língua em particular. Como objetos de tratamento computacional, os *corpora* eletrônicos necessitam de ferramentas computacionais para sua compilação e tratamento (MELLO; SOUZA, 2012).

Existem diversas ferramentas e recursos computacionais disponíveis para exploração de *corpora* eletrônicos. Entretanto, a utilização de tais recursos requer do usuário pesquisador o conhecimento da linguagem de consulta utilizada pela ferramenta ou da anotação utilizada, além da sua instalação em um computador. Tais ferramentas permitem a consecução de tarefas gerais relacionadas a análises de *corpora*, mas não atendem a requisitos inerentes a determinados estudos, uma vez que não tenham sido elaboradas para análises de projetos específicos (SARDINHA, 2006; GERBER; VASILÉVSKI, 2007; PAIXÃO DE SOUSA, 2006).

Dentro desse cenário, este trabalho se justifica através da necessidade do desenvolvimento de uma ferramenta de software, com uma interface de fácil utilização, que permita a recuperação de informação de *corpora* para auxiliar as pesquisas sobre as estruturas gramaticais do português, sobretudo da morfologia e da sintaxe da língua, contribuir para os estudos de mudança linguística, alem de possibilitar avanços em diversas áreas que perpassam a construção de *corpora* diacrônicos, tais como a história, a filologia, e outras subáreas da linguística, a exemplo da semântica e do discurso. Soma-se a esta justificativa geral, a aplicação de recursos computacionais de software para disponibilização e exploração de dados linguísticos no *Corpus* digital DOViC (*Corpus* de Documentos Oitocentistas de Vitória da Conquista), promovendo a ampliação de pesquisas sobre a história do Português do Brasil e contribuindo para a preservação do patrimônio histórico e linguístico da cidade de Vitória da Conquista.

2.5 Metodologia

A metodologia científica adotada no trabalho caracteriza-se como Pesquisa Aplicada, uma vez que desenvolveu uma solução de software para a disponibilização e recuperação de informação de *corpora* eletrônico anotados nos moldes do Corpus Tycho Brahe, aplicando-a imediatamente ao corpus DOViC. De acordo com Barros e Lehfeld (2000, p. 78) apud Vilaça (2010), a pesquisa aplicada tem como motivação a necessidade de produzir conhecimento para

aplicação de seus resultados, com o objetivo de contribuir para fins práticos visando à solução imediata do problema encontrado na realidade.

A ferramenta desenvolvida, denominada WebSinC, foi construída com uma metodologia de desenvolvimento de software iterativa, na qual diversos ciclos envolvendo atividades de levantamento de requisitos, análise e projeto, implementação e testes são continuamente repetidos até a conclusão da mesma. O levantamento de requisitos e análise abrangem atividades em que as funções, restrições e objetivos do sistema são estabelecidos por meio da consulta aos usuários do sistema. O projeto abrange os detalhes da especificação técnica baseando-se no que foi definido na análise. A implementação consiste em codificar o sistema através de uma linguagem de programação e os testes permitem que o sistema seja validado a fim de garantir que os requisitos de software foram atendidos (SOMMERVILLE, 2003).

Para o levantamento dos requisitos da aplicação, foram utilizadas técnicas de prototipação, além de considerar ferramentas existentes para propósito semelhante. A análise e o projeto da aplicação 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 tornou-se 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, uma linguagem padronizada e recomendada pelo W3C (*World Wide Web Consortium*) para consultas XML.

O banco de textos de um *corpus* eletrônico tende a crescer e gerar volumes de dados cada vez maiores. Para obtenção de melhor desempenho computacional na extração de informações do *corpus* se fez necessário seu armazenamento num Sistema Gerenciador de Banco de Dados (SGBD). Sistemas típicos de processamento de arquivos possuem várias desvantagens se comparados aos SGBDs, dentre as quais destacam-se problemas de segurança e integridade (KORTH; SILBERSCHATZ; SUDARSHAN, 2006). Cabe ressaltar ainda que Paixão de Sousa e Trippel (2006) relataram problemas com o tempo de resposta relacionados à leitura dos arquivos do *corpus* Histórico do Português Tycho Brahe quando houve crescimento do mesmo, e destacaram que para um grande volume de dados um SGBD XML deveria ser considerado. Assim, a aplicação Web implementada utilizou um SGBD livre com suporte a armazenamento XML, o PostgreSQL.

Para a verificação da qualidade do sistema, foram utilizadas estratégias de teste de software caixa preta. A metodologia de testes de caixa preta examina o sistema de software como uma função, com entradas e saídas, sem se preocupar com os detalhes de implementação (RIOS; MOREIRA, 2006).

Para a verificação dos requisitos de gerenciamento, os documentos do *Corpus* DOViC já catalogados, editados e revisados foram armazenados no Banco de Dados através da aplicação implementada. Para verificação das funções de buscas automáticas, foram realizadas buscas sintáticas no texto História da Província de Santa Cruz, de Pero Magalhães de Gandavo, escrito em 1502, pertencente ao *corpus* Tycho Brahe, e buscas morfossintáticas em uma carta do *corpus* DOViC, entitulada Carta de Liberdade da cabra de Nome Sofia, escrita em 1845 . Os resultados produzidos foram comparados com os resultados produzidos pelas buscas equivalentes na ferramenta *Corpus Search*. Os testes das buscas são detalhados e discutidos no capítulo sete.

Após a verificação e validação do sistema, o WebSinC foi implantado em um servidor Web ligado à Internet, destinado à disponibilização do *corpus DOViC*. Assim, os textos do *corpus* DOViC catalogados e armazenados no banco de dados serão disponibilizados na Internet, permitindo que sejam consultadas informações como título, data de publicação, autor, gênero, local de depósito, cor e material da capa, altura, largura, profundidade, entre outras, e permitindo também a realização de buscas por categorias sintáticas e morfossintáticas.

3 COMPILAÇÃO DE *CORPORA* PARA A DESCRIÇÃO E ANÁLISE DE ESTRUTURAS GRAMATICAIS DE LÍNGUA NATURAL.

A Linguística Computacional é a área da ciência linguística que cuida de investigar o tratamento computacional da linguagem e das línguas naturais para diversos fins práticos. De acordo com Vieira e Lima (2001, p.1), "é a área de conhecimento que explora as relações entre linguística e informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural." A área circunda diversas áreas de pesquisa da "Linguística Teórica e Aplicada, como a Sintaxe, a Semântica, a Fonética e a Fonologia, a Pragmática, a Análise do Discurso, etc." (OTHERO; MENUZZI, 2005, p.22).

Uma coleção de textos, disponível em formato eletrônico ou não, com objetivo de embasar com dados à investigação linguística, é considerada o *corpus* da pesquisa. De acordo com Sardinha (2004), um *corpus* pode ser definido como uma extensa coleção de dados linguísticos (sejam eles textos escritos ou a transcrição de fala), reunidos criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Segundo Kennedy (1998), a maioria dos *corpora* está armazenada atualmente em formato eletrônico para exploração pelas ferramentas computacionais, mas este tipo de armazenamento começou a ser usado, mais intensamente, apenas a partir dos anos 60. De acordo com Bennet (2010), um *corpus* é uma grande coleção de exemplos de ocorrências naturais de uma língua armazenada eletronicamente.

Sardinha (2000, p. 325) define a Linguística de *Corpus* como a área da Linguística que se ocupa da coleta e exploração de *corpora* e, “como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador”, dialogando com a Linguística Computacional em muitos aspectos.

Do ponto de vista linguístico, Mello e Souza (2012) enfatizam a abordagem empirista, tendo como central a noção de linguagem enquanto sistema probabilístico. Baseando-se nessa noção, os autores afirmam que os traços linguísticos não ocorrem de forma aleatória, sendo possível a evidência e quantificação de padrões. Na linguística de *corpus*, afirma-se que a linguagem é padronizada (*patterned*), ou seja, "existe uma correlação entre os traços linguísticos e os contextos situacionais de uso da linguagem", o que se evidencia por colocações, coligações ou estruturas que se repetem significativamente.

Ainda de acordo com Mello e Souza (2012), as principais áreas da Linguística de *Corpus* são:

1. Compilação de *corpora*;
2. Desenvolvimento de ferramentas para análise de *corpora*;
3. Descrição da linguagem;
4. Exploração do uso de descrições baseadas em *corpora* para várias aplicações como ensino-aprendizagem de línguas e gêneros linguísticos, processamento da linguagem natural por máquinas, reconhecimento de voz, construção de gramáticas e dicionários, etc.

Nesse contexto, nas seções seguintes, abordaremos assuntos importantes acerca de compilação de *corpora*; apresentaremos alguns *corpora* eletrônicos disponíveis para estudo; e, trataremos, ainda que sucintamente, das teorias de sintaxe e morfologia, cujos aspectos gerais são pré-requisito para compreensão das anotações de informações linguísticas em *corpora* para os estudos gramaticais, o que consiste numa das etapas da compilação. No capítulo seguinte, apresentaremos as possibilidades de ferramentas para análise e exploração de *corpora*, trazendo mais informações acerca de anotações e buscas em *corpora* anotados.

3.1 *Corpora* eletrônicos disponíveis

O rápido desenvolvimento da informática permitiu um maior número de pesquisas com *corpora* eletrônicos. Nos anos 90 muitos projetos de compilação de *corpora* surgiram no mundo todo e, atualmente, diversos *corpora* eletrônicos estão disponíveis para análise em várias línguas. A maior parte destes são *corpora* de língua inglesa, mas também há *corpora* existentes em línguas como português, francês, espanhol, alemão, tcheco, chinês, entre outras (SARDINHA, 2000).

Alguns dos maiores *corpora* linguísticos disponíveis em língua inglesa são:

- **Open American National *Corpus* (OANC)** - O OANC é um *corpus* de 15 milhões de palavras do inglês falado na América, incluindo textos de todos os gêneros e transcrições de voz, produzidos a partir de 1990. Os textos são anotados automaticamente com informações como estrutura do texto, estrutura sintática (limites das sentenças), classes de palavras, entre outras. Todos os dados do *corpus* incluindo as anotações são disponíveis abertamente, sem restrição de uso (AMERICAN NATIONAL *CORPUS*, 2012).
- **British National *Corpus* (BNC)** - O BNC é um *corpus* de 100 milhões de palavras de amostras do Inglês Britânico, produzidas a partir do século XX, oriundas de língua escrita e falada. A última edição do *corpus* é a XML BNC, distribuída em 2007. 90% do

corpus correponde a dados de escrita e inclui textos de jornais regionais e nacionais, revistas especializadas e periódicos de interesses e idades diversos, livros acadêmicos, textos universitários, entre vários outros tipos de texto. A parte falada, que corresponde a 10% do *corpus*, é compilada a partir de transcrições ortográficas de conversas informais improvisadas ou dados de fala recolhidos em contextos diferentes, como reuniões de negócios ou do governo e programas de rádio. O *corpus* é anotado segundo diretrizes da iniciativa TEI (BRITISH NATIONAL *CORPUS*, 2009).

- **Brown *Corpus*** - O *Standard Corpus of Present-Day American English*, ou simplesmente Brown *corpus*, foi criado em 1964 na Borwon University. Composto por textos de amostra do inglês americano impressos nos Estados Unidos no ano de 1961, é o *corpus* mais conhecido e historicamente importante por ter sido o primeiro *corpus* eletrônico a ser criado. Possui cerca de 1 milhão de palavras e apesar de pequeno e pouco atualizado, ainda é muito utilizado, uma vez que serviu de modelo para outros *corpora* (FRANCIS; KUCERA, 1979; SILVEIRA, 2008; SARDINHA, 2000).
- ***Corpus Penn TreeBank*** - O *Penn TreeBank* é um *corpus* com cerca de 4,5 milhões de palavras, oriundas de textos coletados do Wall Street Journal. Os textos do *corpus* recebem anotação da estrutura sintática e anotação POS marcando as classes gramaticais das palavras. O *corpus* não está disponível gratuitamente (MEGERDOOMIAN, 2003; MARCUS; SANTORINI; MARCINKIEWICZ,1993).

Na língua portuguesa, há vários *corpora* eletrônicos de destaque. O projeto AC/DC (Acesso a *corpora/*Disponibilização de *corpora)* da Linguateca é uma iniciativa criada com o objetivo de juntar recursos disponíveis, como *corpora,* ferramentas e serviços computacionais, num único ponto na Internet, facilitando a comparação e a reutilização do material. O projeto tornou-se uma importante fonte de *corpora* em língua portuguesa, pois muitos dos *corpora* existentes nesta língua estão disponibilizados no site do AC/DC. Alguns destes *corpora* disponíveis são listados a seguir.

- ***Corpus* NILC/São Carlos**: O *corpus* NILC/São Carlos do Núcleo Interinstitucional de Linguística Computacional (NILC), contém textos do português brasileiro contemporâneo. Os textos que compõem o *corpus* são textos jornalísticos, literários, jurídicos, universitários (relativos a atividades acadêmicas, como dissertações, relatórios, aulas, etc.), técnicos e científicos e estão disponíveis na Internet (PINHEIRO; ALUISIO, 2003).

- ***Corpora* do Projeto Lácio-Web**: O projeto Lacio-Web tem o objetivo de disponibilizar, principalmente para linguistas e cientistas da computação, *corpora* de textos escritos em português brasileiro contemporâneo e ferramentas linguístico-computacionais. O projeto disponibiliza atualmente seis *corpora*, que agregam textos jornalísticos, acadêmicos, jurídicos, entre outros. Dos seis *corpora*, o Lacio-Ref é o *corpus* de referência do projeto, composto de textos provenientes de diversos jornais, revistas, teses, dissertações, livros, cartas, biografias e informativos. Existe uma sobreposição parcial deste *corpus* com o material do *corpus* NILC (ALUISIO et al., 2003; LÁCIO-WEB, 2004).
- ***Corpus* Brasileiro**: O projeto *Corpus* Brasileiro, desenvolvido no Centro de Pesquisas, Recursos e Informação de Linguagem (CEPRIL), Programa de Pós-Graduação em Linguística Aplicada da Pontifícia Universidade Católica de São Paulo (PUC-SP), com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), é uma coletânea de aproximadamente um bilhão de palavras de português brasileiro contemporâneo, de vários tipos de linguagem. O projeto tem o objetivo de construir e disponibilizar *online* este *corpus* com grande variedade de registros e gêneros. O usuário pode fazer buscas no *corpus* tendo acesso a informações sobre frequência de ocorrência dos termos e de linhas de concordância onde os termos ocorrem, mas ele não tem acesso ao texto integral (PUCSP, 2014).
- **CETEMPúblico**: O CETEMPúblico (*Corpus de Extractos de Textos Electrónicos* MCT/Publico) é um *corpus* com aproximadamente 180 milhões de palavras em português europeu da área jornalística. O *corpus* foi criado pelo projeto Processamento Computacional do Português (projeto que deu origem a Linguateca) após a assinatura de um protocolo entre o Ministério da Ciência e da Tecnologia (MCT) português e o jornal Público, um jornal diário português de edição de grande circulação que disponibiliza sua edição eletrônica na rede. O *corpus* está disponível gratuitamente para fins de pesquisa na página da Linguateca (ROCHA; SANTOS, 2000).
- **CETENFolha**: O CETENFolha (*Corpus de Extratos de Textos Eletrónicos* do NILC/Folha de São Paulo) é um *corpus* de cerca de 24 milhões de palavras em português brasileiro criado no âmbito do projeto Processamento Computacional do Português (mesmo projeto que criou o CETEMPúblico). O *corpus* foi compilado com base nos textos do jornal Folha de São Paulo e pode ser consultado para fins de pesquisa

como parte do *corpus* NILC/São Carlos. O *corpus* está disponível na página da Linguateca através do projeto AC/DC (LINGUATECA, 2014).

- **CRPC**: O CRPC (*Corpus* de Referência do Português Contemporâneo) é um *corpus* com amostras do Português Europeu e outras variedades (português do Brasil, Angola, Cabo Verde, Guiné-Bissau, Moçambique, São Tomé e Príncipe, Macau, Timor leste e Goa). O *corpus* foi compilado pelo Centro de Linguística da Universidade de Lisboa e contem atualmente 311,4 milhões de palavras. O CRPC é constituído por amostras de diversos tipos de texto escrito (literário, jornalístico, técnico, etc.) e de falas transcritas (informais e formais). O sub*corpus* escrito do CRPC encontra-se disponível para download no site do *corpus* e para pesquisas online através da interface CQPWeb (CENTRO DE LINGUÍSTICA DA UNIVERSIDADE DE LISBOA, 2014).
- **PHPB-RJ**: O *corpus* do Projeto Para a História do Português Brasileiro - RJ (PHPB-RJ) é um *corpus* diacrônico disponível na Internet, constituído de documentos escritos no Brasil, recolhidos nos principais acervos do Rio de Janeiro e de Lisboa. Os documentos são transcrições de impressos do século XIX e edições diplomático-interpretativas de manuscritos dos séculos XVIII e XIX. O empreendimento do *corpus* se deu agregado à proposta de trabalho coletivo do Projeto Para a História do Português Brasileiro (PHPB), nascido em 1997, que integra equipes regionais por todo o país. (UFRJ, 2000).
- ***Corpus* Histórico do Português Tycho Brahe** - O *Corpus Tycho Brahe* é um *corpus* digital composto de textos em português escritos por autores nascidos entre 1380 e 1845, compilado pela Universidade Estadual de Campinas. O desenvolvimento deste *corpus* se deu a partir de 1998, no âmbito do Projeto "Padrões Rítmicos, Fixação de Parâmetros e Mudança Linguística", com objetivo de investigar mudanças ocorridas do Português Clássico ao Português Europeu Moderno. Atualmente o *corpus* conta com 63 textos e mais de 2 milhões de palavras. Está disponível para download e pesquisas no site do *corpus* e também no site da Linguateca, através do projeto AC/DC (UNICAMP, 1998).

3.2 Compilação de *corpora* eletrônicos

"Compilar – ou criar – um *corpus* é projetar e codificar uma coleção de documentos coletados dentro de determinados padrões ou exigências, para a realização de estudos linguísticos ou computacionais de aprendizagem de máquina" (SILVEIRA, 2008, p.29). Vários aspectos envolvem a compilação de um *corpus* e não é consenso entre investigadores

quais os atributos que um *corpus* deve ter. Mesmo quando concordam na utilização de determinados requisitos, pode haver discordâncias na maneira prática de consegui-los. Características como formato, representatividade, organização, anotações utilizadas e padrões de codificação devem ser estudados, discutidos e planejados previamente de acordo com o propósito de cada *corpus* (KENNEDY, 1998; SANTOS, 1999).

O empreendimento de compilar um *corpus* eletrônico envolve vários processos, que compreendem a coleta, a preparação, a segmentação e a anotação dos textos. O processo de coleta intenta obter documentos textuais no formato eletrônico que atendam aos requisitos estabelecidos para a composição do *corpus*. Se os documentos de interesse forem livros ou documentos antigos impressos ou manuscritos, o processo de coleta implicará na digitação ou digitalização dos impressos, ou ainda na transcrição de áudios ou manuscritos. Quando documentos que farão parte do *corpus* já estão disponíveis em formato eletrônico, como arquivos gravados em CDs, DVDs, discos rígidos, bancos de dados ou documentos armazenados na Internet, o processo de coleta de textos torna-se mais abreviado (EVANS, 2008).

O processo de preparação envolve conversão de formatos dos textos. Mesmo já em formato eletrônico, muitos documentos podem não estar no formato adequado para serem lidos pelas ferramentas disponíveis. Em geral, as ferramentas suportam documentos no formato de texto puro, sem nenhuma formatação (formato TXT) ou algum formato específico de análise de *corpora*. Se o *corpus* for constituído de arquivos provenientes da Web, haverá necessidade de conversão do formato HTML (*Hiper Text Markup Language*) para o texto puro. Para textos transcritos, a necessidade de conversão depende do editor utilizado para fazer a digitação. Textos digitalizados através de ferramentas OCR (*Optical Character Recognition*) requerem revisão, pois geralmente o texto de saída contém inconsistências em relação ao original. Elementos não textuais, como figuras por exemplo, também podem estar incluídos e precisam ser removidos. Os formatos gerados por esse tipo de ferramenta geralmente são PDF ou DOC e também podem demandar a conversão para texto puro (EVANS, 2008).

O processo de segmentação ou *tokenização* consiste em dividir o texto em unidades distintas ou *tokens*, que são as menores unidades de um texto. A segmentação também determina limites de palavras, sentenças e parágrafos. Existem muitas ferramentas computacionais, conhecidas como *tokenizers* ou analisadores léxicos, que fazem este processo de forma automática (MEGERDOOMIAN, 2003; JACKSON; MOULINIER, 2002).

O processo de anotar os textos consiste em realizar inserção automática, semi-automática ou manual de anotações, também conhecidas como marcações, etiquetas ou *tags*,

em um *corpus* de estudo. As anotações são informações adicionais, inseridas com objetivo de facilitar a análise linguística, auxiliando o pesquisador a extrair do *corpus* o máximo possível de informação. A segmentação do texto é um pré-requisito para a inserção das anotações (KENNEDY, 1998; EVANS, 2008; GERBER; VASILÉVSKI, 2007). Os processos de segmentação e anotação serão abordados com maior ênfase no capítulo seguinte.

As etiquetas podem identificar divisões do texto como parágrafos e números de linhas ou informações referentes a autoria e origem dos textos, como autor, data de publicação do documento, título, etc., as quais são normalmente obtidas na fase de coleta e registradas em seções na forma de cabeçalhos ou *headers*. Os textos podem receber anotações com informações linguísticas de vários tipos, como a classe gramatical de cada palavra ou a estrutura das sentenças. O processo de identificação dos principais constituintes do texto, tais como sintagmas nominais, sintagmas verbais, etc. é conhecido como *parsing* e deve ser realizado para inserção da anotação de estrutura sintática (EVANS, 2008; JACKSON; MOULINIER, 2002).

Evans (2008) também destaca que qualquer documento preparado para a análise de *corpus* é apenas uma representação do documento original. Devido a isso, muita informação contextual pode ser perdida na etapa de preparação do texto para um formato exigido pela ferramenta de análise. A maneira usual de lidar com este problema é também através da inserção de anotações. Assim, outros tipos de anotação podem ser acrescentadas, com informações sobre a estrutura e formatação do original, como a presença de cabeçalhos, rodapés, quebras de páginas, notas, trechos em negrito, itálico, etc.

3.3 Morfologia

A morfologia é "frequentemente definida como o componente da Gramática que trata da estrutura interna das palavras" (Sandalo, 2001, p.181). Goldsmith (2010) afirma que esta área envolve o estudo do que uma palavra é nas línguas naturais e defende que, na prática, a morfologia abrange o estudo de quatro aspectos autônomos: a identificação do léxico de uma língua, a morfofonologia, a morfossintaxe, e a decomposição morfológica, ou estudo da estrutura interna das palavras. Nesta seção, trataremos da decomposição morfológica e focalizaremos a morfossintaxe, por serem estes os aspectos da morfologia relevantes para a anotação e busca automática, tarefas contempladas nesta pesquisa.

A morfossintaxe lida com a relação entre os morfemas encontrados dentro de uma palavra e as outras palavras que a cercam na mesma sentença; é uma responsabilidade

partilhada entre as disciplinas de sintaxe e morfologia. A sintaxe é o domínio da análise linguística responsável pela formação de sentenças; do ponto de vista sintático, palavras são unidades que compõem uma sentença, e são agrupadas de acordo com a função gramatical na mesma.

A decomposição morfológica se preocupa com a estrutura interna das palavras, e quais as regras para formá-las a partir de pequenas unidades - os morfemas, ou seja, é a área da morfologia que estuda a relação dos morfemas no interior da palavra. No entanto, os morfemas que aparecem no interior de uma palavra também podem especificar informações sobre outras palavras na sentença - por exemplo, o sufixo verbal *-s* em *Sincerity frightens John* especifica que o sujeito do verbo está no singular. Assim, um estudo da sintaxe, inevitavelmente, leva a abordagens que contemplam a estrutura interna de pelo menos algumas palavras em uma língua, o que em muitas línguas é a regra e não a exceção (TROST, 2003; GOLDSMITH, 2010).

Enquanto a palavra é a unidade máxima da morfologia, o morfema é a sua unidade mínima, sendo os morfemas os menores elementos que carregam significado dentro de uma palavra (SANDALO, 2001). Trost (2003) os conceitua como entidades abstratas que expressam características, relacionadas a conceitos semânticos, que são as raízes, como *door*, *play, smart, etc,* ou conceitos abstratos, de natureza gramatical (morfossintática ou morfossemântica), como o *s* plural em *door**s***, o tempo passado *ed* em *play**ed***, o modificador superlativo ***est*** em *smart**est***, etc.

O tipo e a quantidade de informações expressas pela morfologia diferem bastante entre as línguas. Informações expressas pela sintaxe em uma língua são expressas morfologicamente em outra. Por exemplo, o inglês usa um verbo auxiliar para marcação do tempo verbal futuro enquanto que o espanhol usa um sufixo. O quadro 1 mostra um exemplo.

Quadro 1 - Exemplo de informação marcada pela morfologia e sintaxe - inglês x espanhol.

INGLÊS	**ESPANHOL**
I speak	Hablo
I will speak	hablaré

Fonte: Trost (2003).

Enquanto o tempo futuro é marcado pela palavra *will* (verbo auxiliar) em inglês, o mesmo traço morfossintático é marcado por um morfema no interior da palavra *hablaré* (*ré*) em espanhol.

Algum tipo de informação pode estar presente em uma língua e ausente em outra. Por exemplo, o japonês não marca plural de nomes, enquanto muitas outras línguas o fazem. Um exemplo pode ser visto no quadro 2.

Quadro 2 - Exemplo de marcação de número (singular/plural) em nomes no inglês e japonês.

INGLÊS	JAPONÊS
book	Hon
books	Hon

Fonte: Trost (2003).

Em inglês *s,* quando preso à forma nominal *book,* representa o número plural do substantivo que faz oposição a ausência de morfema para a expressão do singular. Em japonês a palavra *hon* é invariável com relação ao número, ou seja, não apresenta um desdobramento para efetuar a oposição singular x plural. Tal oposição não é feita por morfema preso à *hon*.

Em li nguas isolantes como o chine s, cada palavra carrega apenas um significado, mas em li nguas polissinte ticas como o kadiwe u, falada no Mato Grosso do Sul, certas palavras, carregam significados traduzidos por frases em li nguas como o portugue s. Segue exemplo no quadro 3.

Quadro 3 - Exemplo de palavra e frase em kadiwéu e português.

KADIWÉU	PORTUGUÊS
jotagangetagadomitiwaji	Eu falo com ele por vocês

Fonte: Sandalo (2001).

Jotagangetagadomitiwaji, do kadiwe u, é formada, segundo Sândalo (2001), pela composição de unidades de sentido menores – *j*: sujeito de primeira pessoa; *otagan*: falar; *gen*: transitivizador; *t*: objeto indireto; *ga*: segunda pessoa; *dom:*objeto direto benefactivo; *it* plural do objeto indireto; *waji*: plural do objeto direto. A mesma proposição semântica em português é composta por seis palavras que constituem uma frase na língua, cada palavra da composição sintática – *eu falo com ele por vocês* – possui uma estrutura interna específica e o sentido derivado dos morfemas que compõem as palavras portuguesas são realizados em kadiwéu por formas presas que compõem uma única palavra na língua.

Segundo Anderson (1982), a flexão é uma operação morfológica com relevância sintática, uma vez que as categorias flexionais refletem uma profunda relação entre a estrutura das palavras e a estrutura das sentenças. Algumas propriedades das palavras individuais são dependentes da sua posição na estrutura sentencial. Pode-se dizer ainda que "A flexão é

requerida pela sintaxe da sentença, isto é, um contexto sintático apropriado leva à expressão das categorias flexionais, o que não acontece com a derivação, isenta do requisito “obrigatoriedade sintática”" (GONÇALVES, 2011, p.12).

A flexão não altera a categoria *Part-of-Speech* (POS - Parte do Discurso) de uma palavra mas a sua função gramatical. Todas as formas diferentes que a flexão produz para uma palavra formam seu paradigma. Pode-se dizer também que a flexão é uma operação completa, uma vez que todas as formas do paradigma existem para uma determinada palavra. Por exemplo, para nomes que se referem a seres animados, podemos encontrar as quatro formas flexionais: masculino singular, feminino singular, masculino plural e feminino plural (TROST, 2003).

A derivação é uma operação morfológica que cria novas palavras. Nesta, uma nova palavra é produzida pela adição de um morfema à forma base, usualmente com categoria POS diferente da palavra que deu origem à derivada. A derivação é incompleta, uma vez que um morfema derivacional pode não ter aplicabilidade a todas as palavras de uma classe. "Em outros termos, a sintaxe impõe o uso de afixos flexionais, mas é cega à constituição interna da palavra derivada, sendo, portanto, insensível à existência de afixos derivacionais." (GONÇALVES, 2011, p. 12).

O quadro 4 mostra um exemplo de derivação com o sufixo *bar* no alemão, que pode ser aplicado à maioria dos verbos para produzir adjetivos.

Quadro 4 - Exemplo de derivação de verbos em adjetivos no alemão.

VERBO ALEMÃO (tradução em português)	**ADJETIVO ALEMÃO (tradução em português)**
essen (comer)	essbar (comestível)
absehen (conceber)	absehbar (concebível)
sehen (ver)	*sehbar (visível)

Fonte: Adaptado de Trost (2003).

O verbo *essen* (comer) em alemão, pode ser acrescido do afixo de adjetivalização ***bar***, e transformado no adjetivo *ess**bar*** (comestível). O mesmo pode ocorrer com *absehen* (conceber) tranformando-o no adjetivo *absehbar*. No entanto, o afixo ***bar*** não pode ser aplicado ao verbo *sehen* (ver).

A morfologia computacional lida com o processamento de palavras, seja na forma escrita ou falada, e tem uma extensa variedade de aplicações práticas. Nesta área de estudo, a tarefa mais básica consiste em transformar uma *string* de caracteres de entrada em uma *string* mapeada com os morfemas que formam a palavra e a sua interpretação morfossintática como saída. O quadro 5 mostra um exemplo de uma *string* no inglês dada como entrada, seu

mapeamento em morfemas e sua interpretação morfossintática dados como saída (TROST, 2003).

Quadro 5 - Mapeamento de uma *string* do inglês em morfemas e interpretação morfossintática.

Entrada	incompatibilities
Morfemas da *string* de entrada	in+con+patible+ity+s
Interpretação morfossintática	incompatibility+NounPlural

Fonte: Adaptado de Trost (2003).

A palavra *incompabilities* é mapeada no conjunto de morfemas que a compõem: *in*, *con*, *patible*, *ity* e *s* (designa plural). A interpretação morfossintática produz como saída a raiz *incompatibility* com a informação de categoria gramatical, que é um nome no plural. Um pouco mais desse assunto será abordado no capítulo seguinte. Passemos agora para algumas considerações sobre a sintaxe.

3.4 Sintaxe

A sintaxe é a disciplina linguística que estuda como as palavras são combinadas para formar sintagmas e como estes se combinam para formar sentenças. Um sintagma é uma unidade sintática construída hierarquicamente, apesar das sentenças pronunciadas ou escritas na língua sejam realizadas pela fonologia ou pela escrita como uma sequência linear de sons ou letras (MIOTO; SILVA; LOPES, 2013).

Na estrutura sintática das sentenças, há dois aspectos distintos, porém inter-relacionados. O primeiro compreende as relações gramaticais, como a função de sujeito e a função objeto em uma sentença, e também engloba relacionamentos como modificador-modificado, possuidor-possuído, etc. (VAN VALIN JR, 2001; RAPOSO,1992; CHOMSKY, 1993).

O segundo aspecto da sintaxe está relacionado à organização das unidades constituintes das sentenças. Uma sentença não consiste simplesmente de uma sequência de palavras. As palavras estão organizadas dentro de unidades, que por sua vez, estão inseridas em unidades maiores, chamadas de constituintes. A organização hierárquica das unidades na sentença é chamada de estrutura de constituinte (VAN VALIN JR, 2001; RAPOSO,1992; CHOMSKY, 1993).

3.4.1 O que são constituintes sintáticos e como são organizados

Todo constituinte é construído a partir de um núcleo que projeta uma estrutura sintagmática máxima consoante às suas relações semânticas e categoriais. Os núcleos podem ser de natureza lexical ou gramatical/funcional. Os núcleos lexicais compreendem os verbos, nomes, preposições e adjetivos. Outras categorias como flexão de tempo e modo verbais, determinantes, conjunções, são categorias funcionais. A teoria gramatical convencionou rótulos associados a cada categoria morfossintática: V para verbos, N para nomes, P para preposições, etc (MIOTO; SILVA; LOPES, 2013; RAPOSO,1992; CHOMSKY, 1993).

O constituinte composto por um substantivo e um artigo é comumente chamado de *Noun Phrase* [NP] no inglês, traduzido como Sintagma Nominal no português. Uma preposição seguida de um NP também formam um constituinte numa sentença, que por sua vez é chamado de *Preposition Phrase* [PP], ou Sintagma Preposicional. Consideremos a sentença *The teacher read a book in the library*, dado como exemplo por Van Valin Jr (2001). O substantivo *teacher* e o artigo *the* formam um NP, assim como *the* e *library* formam outro NP. A preposição *in* seguida do último NP mencionado formam um PP. O constituinte composto de um verbo mais um NP que o segue é chamado de *Verbal Phrase* [VP], ou Sintagma Verbal. Na sentença de exemplo, o verbo *read* mais o NP *a book* e o PP *in the library* formam um VP. Essa estrutura de constituintes pode ser representada na notação de colchetes aninhados, como mostra a figura 1 (VAN VALIN JR, 2001).

Figura 1- Exemplo de estrutura constituinte usando notação de colchetes aninhados.

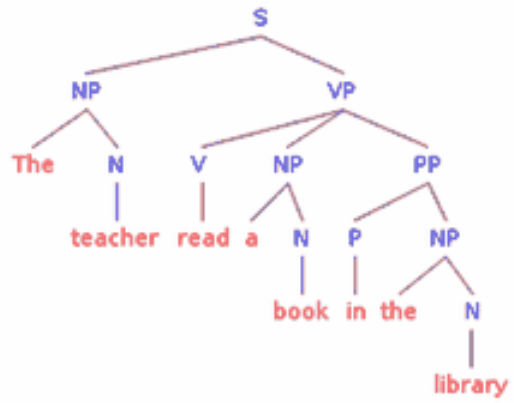
$[_S\ [_{NP}$ The $[_N$ teacher$]]\ [_{VP}\ [_V$ read$]\ [_{NP}$ a $[_N$ book$]]\ [_{PP}\ [_P$ in$]\ [_{NP}$ the $[_N$ library$]]\ _{PP}]\ _{VP}]\ _S]$

Fonte: Van Valin Jr (2001).

O aninhamento dos colchetes denota a estrutura hierárquica. S representa toda a sentença e portanto, é o colchete mais externo. O primeiro NP está interno a S e portanto, está dentro de S na estrutura hierárquica. O colchete marcado com o subscrito N está interno a este NP, e portanto, está dentro do sintagma NP. O encerramento do núcleo N e do NP se dão com o fechamento dos dois colchetes seguintes. Segue-se com esta mesma lógica de raciocínio para o restante da sentença na figura.

Uma outra representação para a mesma sentença pode ser ainda dada graficamente em formato de árvore, conforme mostra a figura 2. A árvore foi construída com auxílio da ferramenta phpSyntaxTree (EISENBACH; EISENBACH,2003).

Figura 2 - Exemplo de estrutura constituinte usando notação gráfica arbórea.



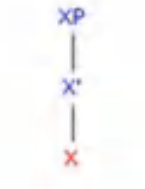
A sentença S é a raiz da árvore. Um NP e um VP estão um nível abaixo na estrutura hierárquica, sendo portanto, internos a S ou dominados diretamente por ele. O NP por sua vez, domina o item *The* e o núcleo N, que possui o item *teacher* como filho. Segue-se com esta mesma lógica de raciocínio para o restante da sentença na figura.

3.4.2 A Teoria X-Barra

A teoria X-barra trata de mostrar como um sintagma é estruturado, explicitando as relações que se estabelecem dentro dele e como eles estão organizados de forma hierárquica para formar a sentença. Esta teoria representa o núcleo do sintagma por uma variável X, que recebe a atribuição de um valor dependente da categoria do núcleo. Assim, se a categoria é nome, o valor atribuído à variável X será N (nome). Para núcleos de categoria verbal, a variável X recebe o valor V (verbo), para categoria preposicional, valor P (preposição), entre outros (MIOTO; SILVA; LOPES, 2013; RAPOSO,1992; CHOMSKY, 1993).

O núcleo X determinará as relações internas que se estabelecem em dois níveis: o nível X' (X-Barra) e o nível XP (onde P abrevia o termo *Phrase*). Estas relações estão representadas na figura 3.

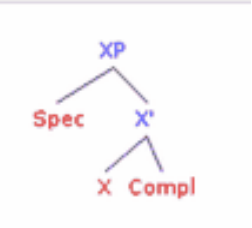
Figura 3 - Representação de relações sintagmáticas na Teoria X-Barra.



X é uma categoria mínima às vezes também representada como X_0. X' é chamado de nível intermediário e XP de nível sintagmático ou projeção máxima de X. "Na projeção

intermediária X’ o núcleo pode estar relacionado com complementos (Compl) e na projeção máxima pode estar relacionado com um especificador (Spec).” A representação desse esquema pode ser dada em forma de árvore, como mostra a figura 4 (MIOTO; SILVA; LOPES, 2013; RAPOSO,1992; CHOMSKY, 1993).

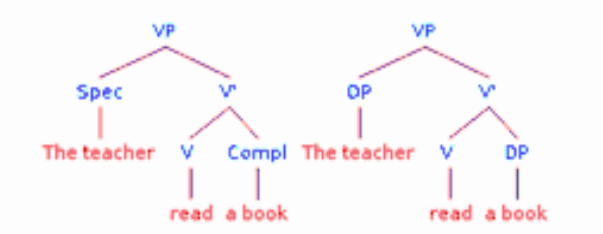
Figura 4 - Representação de relações sintagmáticas na Teoria X-Barra com projeções máxima, mínima e intermediária.



A projeção máxima de uma categoria lexical (XP) é a categoria de nível mais elevado e na teoria proposta por Chomsky (1970), ela é designada por X''. Esta projeção é obtida através da composição da projeção intermediária X' com o especificador da categoria lexical X. A projeção X', por sua vez, é obtida através da compisção da categoria lexical X com seus complementos.

Como exemplo de representação dos níveis estabelecidos dentro do sintagma, tomemos parte da sentença vista na figura 1, *The teacher read a book*. Duas representações simplificadas, explicitando as projeções de VP sem detalhar as projeções DP são mostradas na figura 5.

Figura 5 - Representação de relações sintagmáticas da sentença *The teacher read a book.*

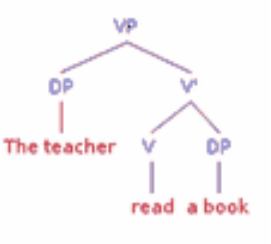


Existem relações entre os nós presentes na estrutura arbórea: a relação de dominância, dominância imediata, maternidade, irmandade, precedência e c-comando, cujas definições são dadas abaixo:

- **Dominância** - Um nó α domina um nó β na árvore se β é descendente de α, ou seja, é possível fazer um percurso descendente de α até β, por meio de uma conexão entre os galhos. A relação de dominância denota a relação de inclusão existente entre dois sintagmas, de maneira que o sintagma dominado (que está abaixo) está incluído no que domina (o que está acima) (MIOTO; SILVA; LOPES, 2013). Na figura 6, VP domina

DP, domina *The teacher*, domina V' e também domina V, DP, *read* e *a book*. V' domina V, DP, *read* e *a book*.

Figura 6 - Representação de relações estruturais na árvore.



- **Dominância imediata** - Um nó α domina imediatamente um nó β na árvore se β é descendente de α e não há nenhum outro nó γ entre α e β (MIOTO; SILVA; LOPES, 2013). Na figura 6, VP domina imediatamente DP, mas não domina imediatamente *The teacher*. V também domina imediatamente V', o qual domina imediatamente V e DP. V domina imediatamente *read* e DP domina imediatamente *a book*.
- **Maternidade** - Um nó α é mãe de β se α domina β imediatemente (MIOTO; SILVA; LOPES, 2013). Na figura 6, VP é mãe de DP e de V'.
- **Irmandade** - Um nó α é irmão de β se α e β tem a mesma mãe (MIOTO; SILVA; LOPES, 2013). Na figura 6, DP e de V' são irmãos, assim como V e DP.
- **Precedência** - Um nó α precede β se α estiver à esquerda de β e não dominá-lo nem ser dominado por ele (MIOTO; SILVA; LOPES, 2013). Na figura 6, DP precede V', V, DP, *read* e *a book*.
- **C-Comando** - Um nó α c-comanda β se β é irmão de α ou se β é dominado pelo irmão de α (MIOTO; SILVA; LOPES, 2013). Na figura 6, DP c-comanda V' e também c-comanda os descendentes deste, que são: V, DP, *read* e *a book*.

4 FERRAMENTAS DE ANÁLISE E EXPLORAÇÃO DE *CORPORA*, ANOTAÇÕES E BUSCAS EM *CORPORA*

A Linguística Computacional é um campo da linguística que se preocupa com o processamento da linguagem por meio de computadores (KAY, 2010). A área cuida de investigar o tratamento computacional da linguagem e das línguas naturais para diversos fins práticos (OTHERO; MENUZZI, 2005).

Othero (2006) diz que apenas para fins didáticos, a Linguística Computacional pode ser dividida em duas subáreas: a Linguística de *Corpus* e o Processamento de Linguagem Natural

(PLN). Segundo o autor, a primeira se ocupa do trabalho com *corpora* eletrônicos com amostras de linguagem natural, objetivando a investigação de fenômenos linguísticos. Já a área de PLN preocupa-se diretamente com o estudo da linguagem voltado para a construção de softwares, aplicativos e sistemas computacionais específicos, capazes de interpretar e/ou gerar informações em língua natural. O termo *Processamento de Linguagem Natural* é normalmente usado para descrever a função dos componentes de software ou hardware em um computador que podem analisar ou sintetizar a linguagem falada ou escrita (JACKSON; MOULINIER, 2002).

McEnery (2003) compreende *corpus* como um recurso multifuncional, útil não apenas para a linguística e para o PLN, mas também para diversas outras áreas de pesquisa. O tratamento computacional de dados linguísticos faz uso de ferramentas que empregam técnicas de PLN e portanto, para muitas aplicações desta área, os dados de um *corpus* são dados brutos de entrada e/ou testes para avaliação destas aplicações. Por uma estreita relação das aplicações de PLN com a compilação de *corpora*, trataremos destas ferramentas neste capítulo.

A maioria dos sistemas de PLN não são entidades monolíticas, mas consistem em distintos componentes de software, dispostos para um processamento *pipeline*. Por exemplo, a identificação de papel semântico (por exemplo, o agente, paciente, tema) depende de análise sintática, que por sua vez, depende de marcação com categoria POS, que por sua vez depende da *tokenization* (separação do texto em *tokens*) (RESNICK; LIN, 2010). Jackson e Moulinier (2002) usam o termo "*layered fashion"* para definir esse tipo de processamento, explicando que para análise linguística de textos o processamento normalmente é feito da seguinte forma: Os documentos são divididos em parágrafos, parágrafos em sentenças e sentenças em palavras individuais. Palavras em uma sentença são então marcadas ou etiquetadas com informações POS (*Part of Speech*- Parte do Discurso) e outras características, antes que a sentença seja analisada sintaticamente. Assim, analisadores sintáticos ou *parsers*, estão numa "camada" mais alta, ou seja, fazem a análise depois dos trabalhos realizados por delimitadores de sentenças (*sentence delimiters*), *tokenizers, stemmers e taggers* POS. No entanto, nem todas as aplicações de PLN requerem um conjunto completo de tais ferramentas. Trataremos destes recursos nas seções a seguir.

4.1 Delimitadores de sentenças (*Sentence delimiters*)

Um delimitador de sentença consiste num componente de software cuja tarefa é detectar os limites das sentenças. A realização deste trabalho com precisão deixa de ser simples visto que

existe ambiguidade frequentemente presente nos sinais de pontuação que marcam o fim de uma sentença. Por exemplo, o símbolo de ponto "." pode denotar um ponto decimal, ou uma abreviação, além de marcar o final de uma frase. Da mesma maneira, sentenças começam com letras maiúsculas, mas nem todas as palavras que começam com uma letra maiúscula iniciam uma sentença (JACKSON; MOULINIER, 2002).

Para lidar com a ambiguidade dos sinais de pontuação, delimitadores de sentenças muitas vezes utilizam expressões regulares ou regras de exceção. Outras ferramentas de segmentação utilizam técnicas empíricas, baseadas no aprendizado de máquina, e são treinadas em um *corpus* segmentado manualmente (JACKSON; MOULINIER, 2002).

4.2 *Tokenizers*

Tokenizers, também conhecidos como analisadores léxicos, segmentam uma sentença em cadeias de caracteres com significado, chamadas de *tokens*. A simples abordagem de considerar um *token* como qualquer sequência de caracteres separados por um espaço em branco pode ser adequada para algumas aplicações, mas pode levar a imprecisões. É preciso considerar sinais de pontuação, vírgulas e hífens. Além disso, existem línguas que permitem o espaço em branco dentro de uma palavra, como por exemplo em *pomme de terre* no francês (termo francês para "batata") (JACKSON; MOULINIER, 2002).

As línguas não segmentadas, como muitas línguas orientais, não colocam espaços entre palavras, escrevendo cada *token* adjacente a outro. Outros idiomas, como alemão, finlandês ou coreano, mantém a maioria dos espaços em branco, mas permitem a criação dinâmica de palavras compostas, como por exemplo *Lebensversicherungsgesellschaft* (termo alemão para "empresa de seguro de vida"). Estes compostos podem ser considerados como uma única palavra, mas em uma tarefa de recuperação de documentos, é importante que haja uma segmentação identificando cada "parte" dela (JACKSON; MOULINIER, 2002; MIKHEEV, 2003).

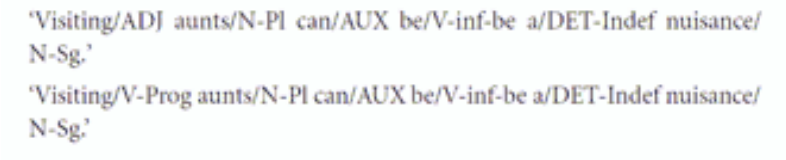
A separação do texto em *tokens* e a delimitação de sentenças podem ser considerados segmentação do texto do tipo *low-level*, uma vez que são desenvolvidos nos estágios iniciais do processamento do texto. Outras tarefas pode ser consideradas como segmentação *high-level*: Segmentação intra-sentencial ou inter-sentencial. A primeira envolve a segmentação de grupos linguísticos, como as entidades nomeadas, e segmentação de grupos verbais e grupos nominais, que também é chamado de *chunking* sintático. A segunda envolve agrupamento de sentenças e parágrafos dentro de tópicos do discurso, que são chamados *text tiles* (MIKHEEV, 2003).

4.3 *Taggers POS (Part-of-Speech Taggers)*

Part- of-Speech tem sido reconhecido na linguística desde muito tempo. A primeira gramática da tradição ocidental, a *Téchné Grammatiké* (Arte da Gramática), escrita por Dionísio Trácio (*Dionysios Thrax*), por volta de 100 A.C., fez distinção entre oito classes de palavras: nomes, verbos, particípios, artigos (incluindo os pronomes relativos), pronomes, preposições, advérbios e conjunções (VOUTILAINEN, 2003).

Taggers POS são componentes de software que realizam o trabalho de etiquetar cada palavra em uma sentença com uma *tag* apropriada, que indica se determinada palavra é um substantivo, verbo, adjetivo, etc. Tal trabalho é realizado "sobre" o trabalho do *tokenizer* e do delimitador de sentenças (MANNING; SCHUTZE, 2000). A figura 7 mostra um exemplo de duas possibilidades de marcação, associadas a uma frase ambígua do inglês. Cada palavra na figura é marcada com uma etiqueta ou *tag*, logo após a mesma, separada pelo símbolo "/".

Figura 7 - Exemplo de uma sentença do inglês com anotação POS realizada por um *tagger*.



Fonte: Van Valin Jr (2001).

Na primeira sentença, *visiting* é um adjetivo que modifica o sujeito *aunts*. Já na segunda, *visiting* é um verbo no gerúndio e *aunts* um objeto. Como o exemplo mostra, mais de uma *tag* POS pode ser atribuída a uma mesma palavra, e a responsabilidade do *tagger* é atribuir a etiqueta correta. No exemplo, *aunts* não é uma informação suficiente na sentença para decidir entre as duas tags.

Part-of-Speech é apenas uma parte da informação que um *tagger* POS produz. Informação flexional e léxico-semântica também são produzidas. Voutilainen (2003) diz que o POS *tagger* também pode ser chamado de *tagger* morfológico, *tagger* para classes de palavras e até mesmo *tagger* lexical.

No PLN há duas abordagens principais para etiquetagem POS: abordagem baseada em regras ou estocática. Um *tagger* baseado em regras faz aplicação de conhecimento linguístico para descartar sentenças sintaticamente incorretas. Um exemplo de regra seria: Se um termo desconhecido é precedido por um determinante e seguido por um substantivo, etiquete-o como um adjetivo. Informações morfológicas também auxiliam no processo de remoção de

ambiguidade. Por exemplo, se uma palavra ambígua/desconhecida termina em *"ing"* e é precedida por um verbo, deve ser rotulada como um verbo (para o inglês) (JACKSON; MOULINIER, 2002).

Na abordagem estocástica, os *taggers* dependem de dados de treinamento, e abrangem métodos que confiam na informação de frequência ou probabilidades para remover as ambiguidades na atribuição de tags. Os *taggers* mais simples baseiam-se na probabilidade que uma palavra ocorre com uma determinada etiqueta. Esta probabilidade é tipicamente calculada a partir de um *corpus* de treinamento, consistindo num conjunto de palavras que foram etiquetadas manualmente. Uma desvantagem dessa abordagem é que sequências sintaticamente incorretas podem ser geradas, embora cada atribuição de etiquetas individuais possa ser válida. Assim, em "*visiting aunts"* do exemplo dado, *visiting* pode ser marcado como um verbo, em vez de um adjetivo, simplesmente porque esta palavra ocorre com mais frequência como um verbo do que como um adjetivo no *corpus* de treinamento. *Taggers* mais complexos usam modelos estocásticos mais avançados (JACKSON; MOULINIER, 2002; MANNING; SCHUTZE, 2000).

Muitos *taggers* têm uma arquitetura similar, que é composta por componentes que realizam as seguintes tarefas (VOUNTILAINEN, 2003):

- *Tokenização* (abordada na seção anterior) - envolve a divisão do texto em *tokens* para análise adicional;
- Procura por ambiguidade - envolve uso de um léxico e algoritmos classificadores para *tokens* não reconhecidos pelo léxico;
- Resolução da ambiguidade (ou desambiguação) - É a principal fase de um *tagger*, que consiste em decidir a *tag* correta para uma ambiguidade encontrada.

4.4 *Parser*

De maneira geral, um *parser* é um software responsável por verificar se uma sequência de símbolos dada como entrada está sintaticamente correta. A construção de um *parser* baseia-se em uma gramática, que descreve as construções válidas da linguagem que ele deve reconhecer. Além de aceitar as entradas corretas, um *parser* deve rejeitar as incorretas, indicando a ocorrência de erros sintáticos (DELAMARO, 2004).

No campo das linguagens formais, uma gramática é, basicamente, um conjunto finito de regras que, aplicadas sucessivamente, geram palavras. No contexto de Processamento de

Linguagem Natural, o termo *parsing* refere-se ao processo de análise automática de uma dada sentença, vista como uma sequência de palavras, a fim de determinar as estruturas sintáticas que a compõem. O processo de *parsing* requer um modelo matemático da sintaxe da linguagem de interesse, que corresponde a uma gramática formal. A gramática formal consiste em um conjunto de regras que especificam como elementos da língua, por exemplo, palavras, podem ser combinados para formar frases, e como frases são estruturados. Regras podem estar relacionadas a informação puramente sintática, tais como funções gramaticais, concordância entre sujeito e verbo, ordem de palavras, etc., mas alguns modelos também podem incorporar questões como semântica lexical (NEDERHOF; SATTA, 2010).

Existe uma vasta gama de formalismos gramaticais, que dependem de diferentes teorias sintáticas, e as estruturas resultantes da análise podem diferir substancialmente entre um formalismo e outro. Muitos formalismos especificam a análise sintática de uma sentença em termos de uma estrutura de sintagmas, que é uma árvore rotulada e ordenada que expressa as relações hierárquicas entre certos grupos de palavras que formam os sintagmas. Uma representação alternativa é a estrutura de dependências, o que indica relações gramaticais binárias entre palavras em uma frase (NEDERHOF; SATTA, 2010).

A análise sintática, ou *parsing*, está relacionada com o reconhecimento, que é o processo de determinar se uma sentença de entrada pertence a uma língua escolhida ou, de forma equivalente, se alguma estrutura sintática pode ser atribuída a esta dada sentença (NEDERHOF; SATTA, 2010).

Como o *parsing* é feito em relação a uma gramática, que é basicamente um conjunto de regras que dizem quais combinações de que categorias POS podem gerar uma sentença bem formada, tomemos como exemplo a sentença dada em (2), retirada de Jackson e Moulinier (2002) :

Colorless green ideas sleep furiously. (2)

A sentença pode ser julgada pelo *parser* como sintaticamente bem formada, já que ADJETIVO+ADJETIVO+NOME é um padrão válido para um sintagma nominal no inglês, VERBO+ ADVÉRBIO é válido para um sintagma verbal, e um sintagma nominal (NP) + um sintagma verbal (VP) formam uma sentença válida. Em contrapartida, a sentença (3) seria julgada como agramatical uma vez que nenhum dos padrões do quadro 6 são aceitos pelas regras sintáticas do inglês (JACKSON; MOULINIER, 2002).

Furiously sleep ideas green colorless. (3)

Quadro 6 - Padrões não aceitos pelas regras sintáticas do inglês.

- ADVÉRBIO + VERBO + NOME + ADJETIVO + ADJETIVO
- ADVÉRBIO + VERBO + NOME + NOME + ADJETIVO
- ADVÉRBIO + NOME + NOME + ADJETIVO + ADJETIVO
- ADVÉRBIO + NOME + NOME + NOME + ADJETIVO

Fonte: Jackson e Moulinier (2002).

A análise semântica envolve a identificação de diferentes tipos de palavras ou sentenças, por exemplo, o reconhecimento de uma ou mais palavras como um nome próprio, e também identificar o papel que eles desempenham na sentença, por exemplo, o papel de sujeito ou objeto. Diferentes tipos semânticos têm características diferentes, por exemplo, uma palavra ou sintagma nominal pode referir-a algo animado ou inanimado, a uma empresa, uma organização, um lugar, um data, ou uma soma em dinheiro (JACKSON; MOULINIER, 2002) .

Em sistemas de linguagem natural, o *parsing* é geralmente uma fase de processamento entre vários outros. Como resultado do *parsing*, várias análises para um mesma entrada podem ser obtidas, em virtude das entradas ambíguas. Tais resultados são chamados de "*parses*". A eficácia das etapas que se seguem à análise geralmente dependem de ter obtido um pequeno conjunto de *parses* preferidos. Este processo é chamado de desambiguação sintática (NEDERHOF; SATTA, 2010). Uma abordagem comum para remoção da ambiguidade é incrementar cada regra gramatical com algum tipo de valor numérico, ou peso. Durante a análise estes valores são utilizados para determinar qual o *parse* preferido. Um caso especial desta abordagem é a análise probabilística, que conta com a atribuição de probabilidades à regras da gramática. A probabilidade de uma análise é definida como o produto das probabilidades das regras a partir das quais o resultado é construído. A desambiguação é conseguida selecionando a análise com a mais alta probabilidade. O sucesso de análises probabilísticas e de análises ponderadas em geral, é devido à sua flexibilidade e escalabilidade, em contraste com as abordagens para desambiguação sintática que contam com conhecimento aprofundado da língua (NEDERHOF; SATTA, 2010).

A acurácia do *parser* deve ser medida relacionando a saída produzida por ele com a saída esperada de um conjunto de testes de entrada, denominada de "*gold standard"*. Entretanto, a exata correspondência da saída do *parser* com a saída esperada pode não ser muito apropriada,

porque algumas aparentes diferenças nas análises podem não fazer qualquer diferença no significado, e a gramática utilizada pelo *parser* pode ter sido projetada para analisar certas construções de maneira diferente da *gold standard* (CARROLL, 2003).

4.5 Anotação e Buscas em *Corpora* anotados

Corpora são normalmente utilizados para a extração de informação estatística e linguística e para testar hipóteses sobre a linguagem natural. Uma vez que os dados são coletados em um *corpus* de texto ou de voz, eles podem ser usados para investigar fenômenos lingüísticos ou para extrair informações sobre uso da língua. Nesse contexto, os *corpora* podem ser usados para uma série de aplicações distintas e as manipulações feitas para os mesmos dependem principalmente dos objetivos finais do pesquisador (MEGERDOOMIAN, 2003).

A adoção de uma abordagem baseada em *corpus* revela padrões que são difíceis de prever ou observar em textos encontrados no dia a dia. O computador desempenha um papel imprescindível para os estudos nesta área, pois os *corpora* geralmente são extensos e demandam recursos computacionais que trabalhem com grandes volumes de texto. Além da capacidade de armazenamento dos dados e dos recursos de preparação e anotação dos textos, as ferramentas computacionais para exploração de *corpora* fornecem novas perspectivas para a análise linguística, uma vez que podem gerar extração de informações do *corpus* e organização, possibilitando a observação e interpretação de dados (MELLO; SOUZA, 2012).

Lingüistas de *corpus* usam os padrões resultantes e as informações extraídas das buscas para testar hipóteses sobre a língua, sobre gramáticas, ou sobre a linguagem de maneira geral. Em contraste, lingüistas computacionais costumam reunir informações a fim de criar gramáticas e sistemas computacionais mais efetivos. Em lingüística computacional, os resultados dos testes realizados em *corpora* e as informações obtidas a partir de análise de *corpus* são frequentemente utilizados como uma base para a melhoria das aplicações de PLN (MEGERDOOMIAN, 2003).

Embora muita informação possa ser obtida a partir dos textos "crus" dos *corpora*, um pré-requisito para a maioria dos trabalhos de análise de *corpus* é a inserção de anotações no texto, uma vez que as etiquetas facilitam a recuperação das informações no texto eletrônico. A extensão da marcação (por exemplo, palavras *versus* sentenças) e o conjunto de etiquetas utilizados são determinados pelas necessidades da aplicação. Buscas em um *corpus* de texto escrito propiciam ao pesquisador a recuperação de palavras, frases ou sentenças por correspondência de padrão. As buscas associadas a ferramentas computacionais permitem identificar e analisar padrões de uso da língua dentro do *corpus* dado. A recuperação desse tipo de informação é frequentemente usada em análise de colocações, tradução automática, aquisição lexical, análise sintática parcial e sumarização de texto. Análises dos resultados de

buscas podem mostrar concordâncias, repetição de itens coocorrentes, sequências semânticas, etc. (SINCLAIR, 1991 apud HUNSTON, 2012; MEGERDOOMIAN, 2003).

As buscas para recuperação de informação em *corpus* eletrônico estão diretamente associadas ao formato de codificação e anotação utilizados e do tipo de informação anotada no *corpus*. As anotações linguísticas podem ser de vários tipos e níveis, representando informações morfológicas, sintáticas ou semânticas. Há padrões que especificam apenas um nível de informação e outros podem especificar dois níveis ou todos. Quando um mesmo formato de anotação é utilizado para todos os níveis, as informações podem ser mantidas num único arquivo. Caso contrário, a informação referente a cada nível deve ser mantida em arquivos separados (IDE; BONHOMME; ROMARY, 2000). Como muitos padrões de anotação baseiam -se na linguagem XML, faremos uma breve abordagem sobre essa linguagem em uma das subseções seguintes.

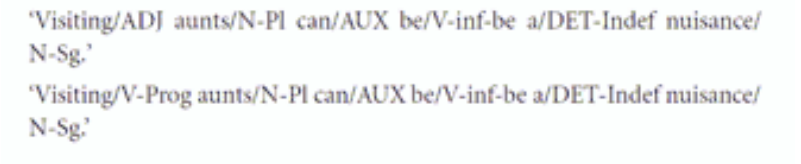
4.5.1 Anotação POS (*Part-Of-Speech*)

A anotação que marca as palavras com suas classes gramaticais é conhecida como *Part-Of-Speech tagging (POS* tag*ging*) e pode ser feita manualmente ou com o auxílio de um software etiquetador (um *tagger* POS, cf. seção 3.3), como o software Palavras (BICK, 2000).

A anotação da classe POS também costuma ser remetida como anotação morfológica, a exemplo das publicações sobre o *Corpus* Tycho Brahe (GALVES; BRITTO, 2008; BRITTO; FINGER; GALVES, 1998). A especificação da anotação POS em morfológica ou morfossintática está diretamente relacionada ao conjunto de tags utilizado (*tag set*) e à atribuição destas pelo *parser*. Marcus, Santorini e Marcinkiewicz (1993) pontuam essa diferença existente entre o conjunto de etiquetas do *corpus* Brown e do *corpus* Penn TreeBank. No Brown, as palavras são etiquetadas independentemente da função sintática. Por exemplo, no sintagma *the one*, *one* é sempre etiquetado com CD (*Cardinal Number* - número cardinal), equanto que no sintagma correspondente no plural *the ones*, *ones* é sempre etiquetado como um NNS (*Plural common noun* - Plural de nome). Já o Penn TreeBank considera o contexto sintático na *tag* POS sempre que possível. Assim, *one* é etiquetado como NNS (Singular common noun - Singular de nome) ao invés de CD quando for núcleo de um sintagma. A anotação POS no contexto desta pesquisa leva em consideração não apenas a categoria POS das palavras mas também a função sintática. Portanto, utilizaremos a terminologia "anotação morfossintática" neste texto para fazer referência à anotação POS e "busca morfossintática" para busca em arquivos de um *corpus* com esta anotação.

Etiquetadores ou *taggers* geralmente são capazes de marcar grandes quantidades de textos com alto nível de precisão. A escolha do software que faz a etiquetagem implica na escolha de um conjunto de *tags* específico. A figura 8 mostra um exemplo de duas possibilidades de marcação POS, associadas a uma frase ambígua do inglês. Cada palavra é marcada com uma etiqueta ou *tag*, logo após a palavra, separada pelo símbolo "/" (JACKSON; MOULINIER, 2002; HALTEREN, 1999 apud SILVEIRA, 2008).

Figura 8 - Exemplo de uma sentença do inglês com anotação POS realizada por um POS *tagger*.



Fonte: Van Valin Jr (2001).

4.5.2 Buscas em *corpora* anotados morfossintaticamente

Buscas por informações em arquivos POS requerem uma pesquisa linear, uma vez que a anotação segue um padrão de linearidade em seu formato. Cada item lexical é seguido imediatamente da etiqueta com a informação morfossintática correspondente. Buscas automáticas para esse tipo de arquivo podem ser feitas usando o recurso de "Localizar" de um editor de texto. No entanto, esse tipo de busca é limitado, não possibilitando uma pesquisa por padrões. Maior potencialidade nos resultados podem ser conseguidos fazendo buscas com uso de expressões regulares. A linguagem Perl é comumente utilizada para buscas com expressões regulares, pois se trata de uma linguagem que fornece este recurso. Entretanto, isso demanda por parte do linguista pesquisador o aprendizado da referida linguagem de programação.

Um exemplo de expressão regular escrita em Perl para busca baseada em categorias morfossintáticas é mostrada no quadro 7. A expressão regular representa um padrão para buscas em textos com anotação no formato POS utilizando o sistema de etiquetas utilizado no *corpus* Tycho Brahe definido por Finger (1998). O objetivo desta expressão é encontrar ocorrências contendo classes de palavras com várias formas do verbo TER, que são palavras anotadas com etiquetas TR (infinitivo), TR-F (infinitivo flexionado), TR-SP (presente do subjuntivo), TR-SD (passado do subjuntivo), TR-SR (futuro do subjuntivo), TR-RA (formas com morfema -ra), TR-P (presente), TR-R (futuro) ou TR-D (passado), seguidas não imediatamente de itens lexicais que não sejam preposições (etiqueta P), preposições combinadas a determinantes e morfemas de número e gênero (P+D-P; P+D-F; P+D-F-P), preposições combinadas a clíticos (P+CL), verbos (VB), verbo HAVER (HV), verbo SER (SR) ou verbo ESTAR (ET).

Quadro 7 - Expressão regular em Perl para busca baseada em categorias morfossintáticas.

```
/([^\/]*\/(TR|TR\-F|TR\-SP|TR\-SD|TR\-SR|TR\-RA|TR\-P|TR\-R|TR\-D)
[^\/]*\/[^(P|P\+D|P\+D\-F|P\+D\-P|P\+D\-F\-P|P\+CL|VB|HV|ET|SR)])/
```

Fonte: Namiuti, 2005.

Apenas a escrita da expressão regular não garante a realização das buscas dentro dos textos. A expressão deve estar contida dentro de um programa completo, que faça a abertura dos arquivos, a leitura do conteúdo buscando pelo padrão descrito e exiba os resultados para o pesquisador. Assim, é necessário conhecimento do uso da linguagem de programação, do uso de expressões regulares e também de lógica de programação.

4.5.3 Anotação sintática

Um *corpus* com anotação no nível sintático é considerado um *treebank*. Um *treebank* é constituído por um conjunto de textos ou coleção de sentenças nos quais a estrutura sintática de constituintes é marcada, convencionalmente por um processo de inclusão de colchetes ou parênteses etiquetados. A forma mais comum de representar a estrutura de sentenças é por meio de uma estrutura árbórea. No entanto, o termo *treebank* não é limitado a *corpora* contendo representações de estrutura sintática, mas aplica-se a todos os tipos de *corpora* gramaticalmente analisados em forma de árvore (JOHANSSON; STENSTROM, 1991).

Assim como há vários formatos para representação e armazenamento de *corpora* linguísticos, há também um variado número de formatos para representação e anotação da estrutura sintática dos textos que os compõem, como TIPSTER, *Penn TreeBank*, Susanne e NeGra (MENGEL; LEZIUS, 2000).

O *Penn TreeBank Format* (Formato *Penn TreeBank*) é um esquema de anotação sintática de *corpora* desenvolvido pela Universidade da Pensilvânia. O esquema utiliza uma representação arbórea delimitada por parênteses etiquetados. Todos os parênteses abertos têm uma etiqueta associada, sendo uma etiqueta *phrase* (NP, ADJP, etc), associada a projeções máximas da teoria X-Barra, ou uma etiqueta *word* (N, ADJ , etc), associadas a núcleos da mesma teoria, representando os nós de uma árvore (SANTORINI, 2010; MARCUS; TAYLOR, 2002).

A cada palavra está associada uma etiqueta *word*, mas nem sempre uma etiqueta *phrase* será associada a cada nó correspondente em uma árvore da teoria sintática. As projeções intermediárias da teoria X-Barra (N', ADJ', etc) não são incluídas nessa representação. Outras

categorias também são omitidas nesse esquema de anotação, como por exemplo, VP e DP. A categoria DP é omitida porque o custo de incluí-la supera sua utilidade. Já os VPs são omitidos porque suas fronteiras normalmente são indeterminadas (caso do Inglês Médio). Mesmo no Inglês moderno, há muitos casos em que não é claro se algum sintagma deve ser um filho de VP ou se deve estar mais acima na árvore (SANTORINI, 2010).

A representação parcial da estrutura sintática se dá por razões práticas, e por esse motivo não se mantém a mesma estrutura correspondente à árvore teórica. Outra diferença para as árvores da teoria sintática é que nesse esquema de representação as árvores não são obrigatoriamente binárias, ou seja, cada nó pode ter mais de duas ramificações (SANTORINI, 2010).

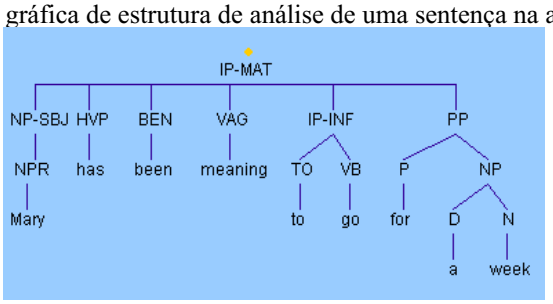
Uma estrutura típica de análise sintática com anotação nesse formato é dada como exemplo no quadro 8. A figura 9 mostra a representação gráfica correspondente a esta mesma estrutura de análise.

Quadro 8 - Estrutura de análise de uma sentença na anotação *Penn TreeBank*.

```
((IP-MAT (NP-SBJ (NPR Mary))
          (HVP has)
          (BEN been)
          (VAG meaning)
          (IP-INF (TO to)
                  (VB go))
          (PP (P for)
              (NP (D a) (N week)))))
```

Fonte: Santorini (2010).

Figura 9 - Representação gráfica de estrutura de análise de uma sentença na anotação *Penn TreeBank.*



Conforme mostram a figura 9 e o quadro 8, além da identificação dos constituintes sintáticos e sua informação categorial, o formato *Penn TreeBank* apresenta anotação para o tipo de sentença (IP-MAT, para oração matriz; IP-INF para oração subordinada reduzida de infinitivo) e para certas funções como a função sujeito (NP-SBJ).

TIPSTER é uma arquitetura que define algumas anotações padrão e atributos associados. Outras anotações e atributos podem ser definidos para uma aplicação TIPSTER e se forem bem definidas podem ser apresentadas para inclusão na arquitetura.
Comparando-o com o padrão CES (*Corpus Encoding Standard*), que segue o método de incorporar as *tags* no texto, o TIPSTER ao contrário, carrega a anotação separada do texto, associando a informação original ao elemento *span* (GRISHMAN, 1998).

O quadro 9 mostra um exemplo de um trecho de documento anotado retirado de Grishman (1998). Na parte superior da tabela está o trecho do documento a ser anotado; imediatamente abaixo da linha com o texto há uma tabela aninhada com a indicação numérica que mostra a posição de cada caracter. Debaixo desta aparecem as anotações, uma anotação por linha. Para cada anotação é mostrado o seu Id, Tipo, *Span*, e Atributos. Para simplificar o exemplo, um único *span* para cada anotação é mostrado. Os atributos são apresentados sob a forma *nome = valor*. O exemplo mostra uma única frase e o resultado de três procedimentos de anotação: separação em *tokens* com atribuição de anotação POS, o reconhecimento do item lexical e o reconhecimento do limite da sentença. Cada símbolo tem um único atributo, a sua parte do discurso (POS), usando a *tag* definida a partir do formato *Penn Tree Bank.* Para a anotação da sentença, os constituintes que a compõem são representados como "[anotação Id]". Por exemplo, "[3]" representa uma referência à anotação 3. (GRISHMAN, 1998).

Quadro 9 - Trecho de documento com anotação TIPSTER.

Texto

```
Cyndi savored the soup.
C y n d i   s a v o r e d   t h e   s o u p .
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
```

Anotações

Id	Tipo	Início do span	Fim do span	Atributos
1	*token*	0	5	pos=NP
2	*token*	6	13	pos=VBD
3	*token*	14	17	pos=DT
4	*token*	18	22	pos=NN
5	*token*	22	23	
6	Name	0	5	name_type=person
7	*sentence*	0	23	constituents= [1],[2],[3],[4],[5]
8	*parse*	0	5	symbol="NP",constituents= [1]
9	*parse*	14	22	symbol="NP",constituents=[3],[4]
10	*parse*	6	22	symbol="VP",constituents=[2],[9]
11	*parse*	0	22	symbol="S",constituents=[8],[10]

Fonte: Adaptado de Grishman (1998).

O quadro 9 mostra, portanto, que o programa identificou cinco *tokens*, aos quais foi atribuído Id numérico de 1 a 5, delimitando seu caracter inicial e final e seus atributos POS (1:0-

5:NP: Cyndi; 2:6-13:VBD: savored; 3:14-17:DT:the; 4:18-22:NN: soup; 5:22-23:.). Os cinco *tokens* foram organizados hierarquicamente em quatro constituintes encaixados: [8], [9], [10] e [11], sendo **[11]** o constituinte superior - a sentença completa - composto pelos constituintes **[8]** e **[10]**, cada um com seus elementos constituintes: [**8**:NP:*Cyndi*] e [**10**:VP: [2:VBD:*savored*],**[9**:[3: DT:*the*],[4:NN:*soup*]**]**.

4.5.4 Buscas em *corpora* anotados sintaticamente

Buscas por informações em arquivos anotados sintaticamente requerem uma pesquisa não linear, uma vez que a anotação segue um padrão hierárquico em seu formato. A estrutura de constituintes é representada hierarquicamente, e portanto, as buscas sintáticas demandam pesquisas dentro de uma estrutura arbórea, que é não linear. Buscas automáticas nesse tipo de estrutura requerem o uso de ferramentas computacionais que façam interface entre o pesquisador e a estratégia de busca, uma vez que estas requerem algoritmos mais complexos. Como *corpora* anotados sintaticamente constituem um *treeBank*, buscas por categorias sintáticas requerem implementação de pesquisas por relações hierárquicas dentro da árvore, denominadas frequentemente de funções de busca. As relações pesquisadas em ferramentas de busca sintática abrangem relações de dominânica, dominânica imediata, c-comando, precedência, precedência imediata e irmandade (MIOTO; SILVA; LOPES, 2013; CORPUS SEARCH, 2009).

4.5.5 Padrões para Anotações em *corpora*

O aumento das pesquisas em linguística de *corpus* e o crescimento na disponibilidade de *corpora* eletrônico fizeram com que diversos formatos de codificação e anotação de textos surgissem. Cada projeto de compilação de *corpus* pode criar e/ou definir um formato, com o objetivo de atender requisitos das ferramentas de anotação e exploração de *corpus* específicas. A diversidade de formatos aumentou a importância e a necessidade de estabelecimento de padrões que facilitassem o compartilhamento, a combinação e o intercâmbio desses recursos. Entre os principais projetos e iniciativas com o propósito de definir um padrão de codificação e anotação de textos, podemos destacar: MuchMore, Tiger- XML, Text Encoding Initiative (TEI) , *Corpus* Encoding Standard (CES), *Corpus* Encoding Standard for XML (XCES) e padrão ISO TC37/SC4.

O CES (*Corpus Encoding Standard*) é um padrão de codificação para *corpora* destinado a atender a necessidade do desenvolvimento de práticas de codificação padronizados para *corpora* linguísticos. Tal padrão identifica um nível de codificação mínima que *corpora* devem alcançar para serem considerados padronizados em termos de representação descritiva (marcação de informação estrutural e linguística). O XCES é a versão do padrão CES baseado em XML (IDE, 1998; IDE; BONHOMME; ROMARY, 2000).

O padrão XCES mantem informações sobre estrutura de sentenças e informações morfológicas numa única estrutura. As informações morfossintáticas são anotadas utilizando-se dos elementos <tok>. A integração com os dados primários é feita através do atributo "xlink". Elementos <s> marcam sentenças e etiquetas <par> marcam parágrafos. A figura 10 mostra um fragmento de um texto contendo anotações de informação morfossintática neste padrão.

O Padrão ISO TC37/SC4 é um *framework* para anotação de informação linguística desenvolvido pela Organização Internacional de Padronização (*International Organization for Standardization).* A ISO formou um subcomitê (SC4) no âmbito da Comissão Técnica 37 (TC37, *Terminology and Other Languages Resources*) com o objetivo de estabelecer padrões internacionais e recomendações para a modelagem de dados, anotação, intercâmbio de dados e avaliação de recursos linguísticos. Dentre os diversos grupos de trabalho do TC37/SC4, um grupo foi criado para prover um *framework* para anotação linguística. A intenção não é definir um esquema ou formato único e definitivo de anotação, mas fornecer uma arquitetura que possa servir de referência para diferentes esquemas de anotação, permitindo a fusão ou comparação entre eles. A estrutura do *framework* tem como finalidade prover o máximo de flexibilidade para codificadores e anotadores, e ao mesmo tempo permitir e estimular o intercâmbio e reutilização de recursos linguísticos anotados (IDE; ROMARY; CLERGERIE, 2003).

Figura 10 - Fragmento de texto com anotação no padrão XCES.

```
<?xml version="1.0">
<chunk type="BODY" lang="en"
 xml:base=
"http://www.cs.vassar.edu/~ME/Oen.xcesDoc#">
 <par xlink:href="xptr(substring(//p[1]">
  <s xlink:href="xptr(substring(//p/s[1]">
   <tok type="WORD"
     xlink:href=
     "xptr(substring(//p/s[1]/text(),1,2">
      <orth>It</orth>
      <disamb>
        <base>it</base>
        <msd>Ppins</msd>
        <ctag>PPER3</ctag></lex>
      <lex>
        <base>it</base>
        <msd>Ppins</msd>
        <ctag>PPER3</ctag></lex></tok>
    <tok type="WORD"
     xlink:href=
     "xptr(substring(//p/s[1]/text(),4,2">
      <orth>was</orth>
      <disamb>
        <base>be</base>
        <msd>Vais3s</msd>
        <ctag>PAST3</ctag></lex>
      <lex>
        <base>be</base>
        <msd>Vais1s</msd>
        <ctag>MUI1</ctag></lex>
      <lex>
     <base>be</base>
     <msd>Vais3s</msd>
     <ctag>MUI3</ctag></lex>
   <lex>
     <base>be</base>
     <msd>Vais1s</msd>
     <ctag>PAST1</ctag></lex>
   <lex>
     <base>be</base>
     <msd>Vais3s</msd>
     <ctag>PAST3</ctag></lex></tok>…
```

Fonte: Ide; Bonhomme; Romary (2000).

O projeto MuchMore (*Multilingual Concept Hierarchies for Medical Information Organization and Retrieval*) propõe um formato de anotação linguística capaz de integrar múltiplos níveis de informação: anotação morfológica, sintática e semântica. O formato é baseado em XML e os níveis de informação podem ser organizados separadamente, sendo integrados através de referência a identificadores (BUITELAAR et al., 2003).

4.5.6 A linguagem XML

XML (*Extensible Markup Language*) 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 (SILVA FILHO, 2004; DEITEL et al., 2005).

Documentos XML são documentos de texto que representam dados de maneira estruturada utilizando um conjunto de tag*s* ou elementos. Tal conjunto não é fixo nem limitado, podendo ser estendido. Assim, os autores dos documentos podem criar suas próprias *tags* para atender a necessidades específicas, o que torna a linguagem poderosa para representar qualquer tipo de dado conferindo-lhe a classificação como uma metalinguagem (SILVA FILHO, 2004; DEITEL et al., 2005).

Ainda que baseie-se em texto, "a XML não se limita a descrever somente dados textuais, mas também pode descrever imagens, gráficos vetoriais, animações ou qualquer outro tipo de dado para o qual seja estendida" (DEITEL et al., 2005).

Dados representados por XML são estrutrurados de forma arbórea, e cada *tag* ou marca representa um nó ou elemento na árvore. "A sintaxe de XML requer um único elemento como nó raiz, uma marca de abertura e de finalização para cada elemento, marcas corretamente aninhadas e valores de atributos entre aspas." (DEITEL et al., 2005).

O quadro 10 mostra um exemplo de um documento XML representando os dados de um livro, com as informações de autor, título e ISBN. O nó raiz é <livro> e este possui como filhos três nós <autor> e um nó <titulo>. A informação de ISBN foi representada como atributo do nó <livro> e seu valor no exemplo é "978-85-7244-800-0".

Quadro 10 - Exemplo de um documento XML.

```
<livro ISBN="978-85-7244-800-0">
   <autor> Carlos Mioto </autor>
   <autor> Ruth Lopes </autor>
   <autor> Maria Cristina Figueiredo Silva </autor>
   <titulo> Novo Manual de Sintaxe</titulo>
</livro>
```

Os documentos XML são legíveis para as pessoas e também manipuláveis por computadores. A ausência de instruções de formatação facilita a realização do processamento sintático de sua estrutura, o que a torna uma referência que pode ser usada para o intercâmbio de dados. Para obter funcionalidade e interoperabilidade na Web, desenvolvedores de software em todo o mundo estão integrando XML a seus aplicativos. Contudo, a XML não está limitada a aplicações Web (DEITEL et al., 2005).

Atualmente, a linguagem XML é um dos formatos mais utilizados para compartilhamento de informação estruturada entre aplicativos, independente de plataforma. Como é um padrão aberto, existe uma grande quantidade de opções relacionadas às ferramentas para implementá-la, permitindo que o usuário escolha o que melhor se ajuste às suas necessidades (W3C, 2010; DEITEL et al., 2005).

4.5.6.1 Linguagens de consulta para XML: XQuery e Xpath

A linguagem XML descreve dados de forma flexível e eficiente através da marcação dos dados com tag*s* descritivas. No entanto, ela não fornece uma maneira de localizar dados específicos dentro de um documento (DEITEL et al., 2005).

A linguagem XPath (XML Path), recomendada pelo W3C, fornece uma sintaxe para localizar dados específicos em um documento XML de forma efetiva e eficiente. XPath modela um documento XML como uma árvore de nós. É uma linguagem de expressões, baseada em *strings*, para localizar conteúdo dentro da árvore que representa o documento XML (W3C, 1999; DEITEL et al., 2005).

Exemplos de expressões XPath são dados nos quadros 11 a 13. Todos os exemplos podem ser aplicados ao documento XML dado como exemplo do quadro 10. A expressão no quadro 11 localiza todos os nós <titulo>, que sejam filhos de <livro>. A expressão no quadro 12 localiza o nó <livro> que possua um atributo ISBN cujo valor seja "`978-85-7244-800-0`". E por fim, a expressão no quadro 13 localiza o terceiro nó filho <autor> do nó <livro>.

Quadro 11 - Exemplo de expressão XPath para localizar nós <titulo> filhos de <livro>.

```
/livro/titulo
```

Quadro 12 - Exemplo de expressão XPath para localizar nós <livro> com atributo ISBN com valor "978-85-7244-800-0".

```
/livro[@ISBN="978-85-7244-800-0"]
```

Quadro 13 - Exemplo de expressão XPath para localizar o terceiro nó <autor> filho de nós <titulo>.

```
/livro/autor[3]
```

XQuery é uma linguagem de consulta projetada pelo W3C para suprir as necessidades de localização, formatação e transformação de elementos em documentos XML. Sua sintaxe permite selecionar os elementos de interesse, reorganizar e, possivelmente, transformá-los retornando os resultados em uma estrutura escolhida. XQuery utiliza expressões XPath para localizações de caminho, o que torna a última um subconjunto da primeira (WALMSLEY, 2007).

Expressões de caminho são convenientes por causa de sua sintaxe compacta, de fácil aprendizado. No entanto, elas têm uma limitação: elas só podem retornar elementos e atributos tal como eles aparecem nos documentos de entrada. Quaisquer elementos selecionados em uma

expressão de caminho deve aparecer nos resultados obtidos com os mesmos nomes, os mesmos atributos e conteúdo, e na mesma ordem como no documento de entrada. XQuery ultrapassa essas limitações porque dispõe de outros recursos como funções e expressões FLWOR. FLWOR (pronuncia-se "flower") é a sigla para "for, let, where, order by, return", as palavras-chave usadas nas expressões (WALMSLEY, 2007).

4.6 Ferramentas para exploração de *corpora* eletrônicos

A acurácia ou grau de corretude de ferramentas de PLN é de grande importância na análise linguística. Para fins de avaliação das ferramentas, Hirschman e Mani (2003) classificam as tecnologias aplicadas às línguas naturais em três classes:

- *Analysis systems* (Sistemas de análise) - São ferramentas que recebem uma entrada em língua natural e produzem uma representação ou classificação dessa entrada, como sistemas de extração e recuperação da informação, identificação de tópico e *parsing*.
- *Language output* - São sistemas que produzem língua natural como saída, como por exemplo, ferramentas de tradução, geração e sumarização de textos.
- *Interactive systems* - São os sistemas em que o usuário e o sistema trocam informações por meio de várias interações para alcance de um objetivo.

As ferramentas até aqui abordadas (*parsers, taggers, stemmers e sentence delimiters*) são ferramentas de PLN e na classificação de Hirschman e Mani (2003) são sistemas de análise (*analysis*). A ferramenta desenvolvida nesta pesquisa, o WebSinc, classifica-se como *Interactive systems*, já que é um sistema interativo para buscas em *corpora*, assim como outras ferramentas existentes para exploração de *corpora* eletrônicos. A ferramenta não está envolta na área de PLN, mas na área de Linguística de *Corpus* (cf. áreas da Linguística Computacional elencadas por Mello e Souza, 2012). Assim, apresentaremos nesta seção algumas das ferramentas existentes para exploração de *corpora* eletrônicos, com enfoque naquelas que possuem função de busca, que assim como o WebSinc, são sistemas interativos para análise linguística, inseridos no âmbito da Linguística de *Corpus*.

4.6.1 Ferramentas interativas de busca para exploração de *corpora* eletrônicos

Uma característica intrínseca das ferramentas para exploração de *corpora* eletrônicos é a implementação de mecanismos computacionais de busca que permitam ao pesquisador a busca

por padrões que auxiliem na sua investigação acerca de fenômenos da língua, e que seriam impossíveis ou muito custosos caso fossem procurados manualmente.

Existem ferramentas que oferecem interface gráfica, outras não. Algumas foram projetadas para uso na web, através da interface do navegador, outras requerem instalação no computador do usuário. Existem as caracterizadas como software livre, e as que requerem licença de uso e pagamento. Há também ferramentas gratuitas mas que não podem ser distribuídas livremente. Enquanto há ferramentas que permitem o carregamento e exploração de *corpora* do usuário, outras têm uso restrito apenas com *corpora* determinados. Há ferramentas que além da busca suportam o processo de compilação de *corpora*, permitindo edição e anotação de textos. Outras realizam apenas a busca. A pesquisa por categoria morfológicas pode ser encontrada em todas elas, já a pesquisa sintática está restrita a um número bem menor de ferramentas.

Nesse contexto, abordaremos primeiro três funcionalidades muito comuns em ferramentas de exploração de *corpora*, que são a contagem de ocorrências, a busca de concordâncias e a busca de colocações. A seguir apresentaremos algumas ferramentas de exploração de *corpora*, iniciando com ferramentas sem interface web (WordSmith, Unitex, *Corpus Search*, Tgrep2 e TIGERSearch) seguidas das que possuem interface baseada na web (Corpógrafo, Portal de *Corpus*, EdiSyn e Lacio-Web). Na última subseção faremos um quadro resumo para comparação entre as ferramentas abordadas.

4.6.1.1 Funcionalidades comuns em ferramentas de exploração de corpora

Os programas para exploração de *corpora* podem ter funcionalidades implementadas que os caracterizam como contadores de frequência, buscadores de concordância ou buscadores de colocações.

Os **contadores de frequência** , também chamados de frequenciadores, contadores de ocorrências ou de frequências, fazem a contagem e calculam a frequência de ocorrências de itens lexicais ou palavras em um *corpus*. Algumas ferramentas que dispõe de contadores de palavras são: WordSmith (cf. seção 3.6.1.2), BNCWeb, TACT, CLAN e Corsis (MELLO; SOUZA, 2012; SILVEIRA, 2008).

Os **buscadores de concordância**, também chamados de concordanciadores, são ferramentas muito utilizadas na linguística de *corpus*, pois permitem que padrões de uso da linguagem sejam descobertos ou compreendidos. Proveêm um mecanismo de busca em que o usuário procura por palavras específicas dentro de um *corpus*, e obtem os resultados com a

exibição das ocorrências dessas palavras em um contexto. Isso facilita o entendimento de padrões de uso, pois possibilita que o linguista identifique, por exemplo, quais termos ou classes de palavras costumam seguir ou preceder o termo pesquisado. Há concordanciadores que permitem que a busca seja feita por meio de expressões regulares, o que abre um maior leque de possibilidades para pesquisas. Se forem utilizadas em *corpora* anotados, podem permitir por exemplo, que buscas sejam realizadas sobre classes de palavras específicas. Algumas ferramentas que dispõe de concordanciadores são: Corpógrafo (cf. seção 3.6.1.7), WordSmith (cf. seção 3.6.1.2), Unitex (cf. seção 3.6.1.3), AntConc, WebConc, Corsis, MonoConc, GlossaNet e *Corpus*Eye (MELLO; SOUZA, 2012; SILVEIRA, 2008).

Os **buscadores de colocações** são softwares que realizam a busca de colocações nos textos do *corpus*. Combinações recorrentes revelam maneiras comuns de organização e posicionamentos de palavras em determinados contextos. A identificação de colocações tem aplicações importantes como na escrita de dicionários e no ensino de línguas (SILVEIRA, 2008).

4.6.1.2 WordSmith Tools

WordSmith Tools é um conjunto integrado de ferramentas destinado à análise linguística. Possui um grande número de usuários em todo o mundo, inclusive no Brasil. Os softwares permitem fazer análises baseadas na frequência e na coocorrência de palavras em *corpora*, e também permitem o pré-processamento de arquivos antes da análise, removendo partes indesejadas do texto, inserindo e removendo etiquetas, organizando os arquivos, etc. Dentre os vários programas que compõem o WordSmith, os principais são o WordList, o KeyWords e o Concord. O software possui interface gráfica para o Sistema Operacional Windows e não é um programa gratuito (SARDINHA, 2006).

4.6.1.3 Unitex

Unitex é uma coleção de programas elaborada para a análise de textos de *corpora* usando recursos linguísticos como gramáticas e dicionários eletrônicos. Os dicionários eletrônicos especificam as palavras simples e compostas de uma língua, juntamente com seus lemas e um conjunto de códigos gramaticais (semânticos e flexionais). A disponibilidade desses dicionários é uma grande vantagem em comparação com os utilitários habituais de busca. As informações que eles contêm podem ser usadas para pesquisa por padrão, descrevendo, assim,

grandes classes de palavras usando padrões muito simples. O Unitex surgiu como uma alternativa gratuita a outro sistema de processamento de *corpus*, o Intex.O software possui interface gráfica e está disponível para vários sistemas operacionais. Entre as funcionalidades disponíveis no sistema Unitex encontram-se: um gerador de concordâncias, um contador de frequências, um gerenciador de dicionários DELA (*Dictionnaires Electroniques du LADL*) e um gerenciador de gramáticas. O Unitex permite também que o usuário faça buscas utilizando expressões regulares. Uma desvantagem do programa é que ele exige que os textos do *corpus* estejam todos agrupados num único arquivo para serem analisados, algo que não é usual (PAUMIER, 2003).

4.6.1.4 A Ferramenta Corpus Search

O *Corpus Search* é um programa que realiza pesquisas sintáticas em *corpora* anotados no formato *Penn TreeBank*. Tanto o esquema de anotação quanto o software foram desenvolvidos na Universidade da Pensilvânia. É uma ferramenta do tipo *desktop* e, portanto, precisa ser instalada no computador do usuário para ser utilizada. A ferramenta utiliza a interface de linha de comando para execução (*CORPUS SEARCH*, 2009).

A busca sintática no programa é realizada com uso de um arquivo de entrada que especifica a consulta desejada. Esta especificação deve estar de acordo com a sintaxe exigida pela linguagem de consulta do *Corpus Search*, que compreende chamadas a funções de busca e uso de operações lógicas. As funções de busca pesquisam relações existentes na estrutura sintática como dominância, c-comando, irmandade, entre outras. Os arquivos de entrada do *corpus* a ser pesquisado devem estar no formato Penn TreeBank, com a estrutura arbórea utilizando parênteses. A ferramenta também permite o uso de expressões regulares.

4.6.1.5 TGrep2

O Tgrep2 é uma reescrita do programa TGrep, que é um programa para análise sintática em *corpora* anotados, com características adicionais, distribuída como software livre. A ferramenta realiza buscas na árvore sintática extraindo estruturas correspondentes a um padrão especificado (ROHDE, 2005).

Assim como na ferramenta *Corpus Search*, os arquivos de entrada do *corpus* a ser pesquisado também devem estar no formato TreeBank. A ferramenta utiliza a interface de linha de comando para execução.

Os padrões especificados no programa TGrep2 são formados por expressões regulares para busca de padrões em árvores. Consistem principalmente de nomes de nós e relacionamentos, que definem links para outros nós. Expressões lógicas também podem ser utilizadas. A expressão regular para especificação de um padrão pode ser inserida como argumento na linha de comando ou pode ser lida a partir de um arquivo (ROHDE, 2005).

4.6.1.6 TIGERSearch

TIGERSearch é uma ferramenta para pesquisa por padrões em *corpora*. O software é desenvolvido na linguagem Java, com uma interface gráfica que permite o processamento de consultas, a visualização de resultados e a exportação de padrões favoritos, dentre várias outras funcionalidades. As consultas podem ser escritas manualmente ou montadas graficamente. O uso de expressões regulares também é um recurso disponível para a especificação de padrões (KÖNIG; LEZIUS; VOORMANN, 2003).

TIGERSearch é um software livre e está disponível para vários sistemas operacionais. Para utilizá-lo o usuário deve realizar o *download* e instalação no computador. A ferramenta requer que o *corpus* em estudo esteja codificado no formato TIGER-XML e portanto, oferece filtros conversores para vários formatos de anotação de *corpora* (KÖNIG; LEZIUS; VOORMANN, 2003).

O editor gráfico de consultas pode ser usado para criar consultas sem o conhecimento da linguagem de consulta TIGERSearch. Podem ser criados nós e arestas que representam as relações de dominância e de precedência. Na estrutura arbórea, podem ser especificadas relações de precedência, precedência imediata, dominância e dominância imediata. A negação de cada um destas também pode ser pesquisada (KÖNIG; LEZIUS; VOORMANN, 2003).

4.6.1.7 A Ferramenta Corpógrafo

O Corpógrafo é um ambiente integrado de ferramentas web projetado para a linguística de *corpus* com a pretensão de apoiar os pesquisadores da língua portuguesa num conjunto de tarefas que compreendem desde a compilação de *corpora* à extração e organização do conhecimento gerado a partir deles. Foi concebido essencialmente para estudo de terminologia, tradução e recuperação da informação, mas também fornece ferramentas para o estudo mais geral da linguagem. O Corpógrafo permite que o usuário possa compilar seus próprios *corpora* e explorá-los, mesmo sem muitos conhecimentos técnicos. O software está disponível gratuitamente no site da Linguateca e sua interface aceita a submissão e extração de textos em diversos formatos. A interface também provê uma ferramenta de edição, onde o usuário pode fazer a limpeza do texto e segmentá-lo em sentenças (MAIA; SARMENTO, 2006; PINTO, 2006).

Análises de itens lexicais e de estruturas sintáticas também são possíveis com as ferramentas de busca do Corpógrafo. A pesquisa pode ser feita não apenas por palavras mas também com uso de expressões regulares da linguagem Perl (MAIA; SARMENTO, 2006; PINTO, 2006).

4.6.1.8 As Ferramentas do Projeto Lacio-Web

O projeto Lacio-Web (ALUISIO et. al., 2003), descrito na seção 2.1, disponibiliza ferramentas linguístico-computacionais aos usuários cadastrados no site do projeto. Dentre as ferramentas podemos citar concordanciadores, frequenciadores e etiquetadores morfossintáticos. Há concordanciadores para *corpora* sem anotação ou anotados morfossintaticamente. Os *corpora* do Projeto Lacio-Web estão disponíveis para uso com as ferramentas. A interface também permite que o usuário faça *upload* de um *corpus* particular em Português do Brasil para que o mesmo seja processado pelos contadores de frequência, concordanciadores para *corpus* sem anotação e etiquetadores (LACIO-WEB, 2004).

4.6.1.9 Portal de Corpus

O Portal de *Corpus* (MUNIZ et al. , 2007) é um portal para *corpora* compatíveis com o padrão XCES que dá acesso a vários *corpora* de textos de jornais brasileiros, compilados no âmbito do projeto PLN-BR (*Resources and tools for information retrieval from Portuguese*

textual bases). Além do acesso aos *corpora*, o portal compartilha ferramentas linguístico-computacionais.

Baseado na Web, o Portal de *Corpus* utiliza tecnologias computacionais livres. Toda a sua documentação e seu código-fonte estão disponíveis gratuitamente, permitindo que seja instalado em um servidor próprio do usuário. Suporta o armazenamento de múltiplos *corpora*, e oferece funcionalidades como: a geração de *subcorpus* a partir de um *corpus* e buscas de textos baseadas nas informações contidas nos cabeçalhos, como informação bibliográfica, seções de jornais, tipos de texto e palavras-chave.

As informações dos textos submetidos pelo usuário são mapeadas para um banco de dados. Depois que um texto é inserido no *corpus*, automaticamente são criadas duas anotações *stand-off* (anotações com marcação lógica e da estrutura sintática, marcando os limites das sentenças) para ele. Além disso, uma versão fundida do texto com essas duas anotações é criada. O sistema do portal mantém sempre uma cópia de cada anotação, do texto e do cabeçalho do texto, tanto no sistema de arquivos quanto no banco de dados.

4.6.1.10 EdiSyn Search Engine

O EdiSyn é uma ferramenta on-line que permite aos usuários consultar e comparar vários *corpora* de dialetos disponíveis na interface. A ferramenta de busca foi desenvolvida no âmbito do projeto EdiSyn. Para garantir interoperabilidade entre as diferentes bases de dados de diferentes *corpora*, um conjunto de *tags* foi criada para o EdiSyn, servindo como um intermediário entre os conjuntos de *tags* de cada banco de dados. As *tags* são compostas por duas partes: Categorias e Características, os quais podem ser combinados ou pesquisados separadamente. As categorias representam a classe de palavras (verbo, advérbio, adjetivo, etc). As características indicam a especificação de uma classe de palavra. Por exemplo uma categoria pode ser "verbo", e "1" (primeira pessoa), "sg" (singular) e "passado" (passado) as especificações desse verbo. Não há nenhuma restrição sobre o número de características que podem ser combinados. Entre os *corpora* disponíveis para consulta encontra-se o CORDIAL-SIN, abordado na seção 4.5.

4.6.1.11 Comparativo entre as ferramentas abordadas

O quadro 14 resume as principais características das ferramentas de exploração de *corpora* abordadas nesta seção, possibilitando um comparativo entre elas. A letra 'X' indica que

a ferramenta possui a característica indicada no cabeçalho da coluna. As características analisadas foram:

1. **Frequenciador** - Foi considerado que a ferramenta é um frequenciador se a funcionalidade de contagem de frequências está implementada e disponível para uso.
2. **Buscador de colocações** - Foi considerado que a ferramenta possui buscador de colocações se esta funcionalidade está implementada e disponível para uso.
3. **Concordanciador** - Foi considerado que a ferramenta é um concordanciador se a funcionalidade de busca de concordâncias está implementada e disponível para uso.
4. **Busca sintática** - Foi considerado que a ferramenta realiza análise da estrutura sintática se permite a realização de buscas automáticas em textos anotados com a representação da estrutura de constituintes das sentenças.
5. **Suporte à compilação de *corpora*** - Foi considerado que a ferramenta fornece suporte à compilação de *corpora* se permite a edição, segmentação ou anotação de textos, seja anotação com etiquetas morfológicas ou morfossintáticas, ou alguma anotação que marque estrutura de sentenças, estrutura do texto ou estrutura de constituintes.
6. **Permite carregamento de textos de *corpus* do usuário** - Foi considerado que a ferramenta permite o carregamento de textos do *corpus* do usuário se as buscas possíveis de serem realizadas não estão restritas apenas a determinados *corpora*, ou seja, o usuário pode utilizar a ferramenta para explorar seu próprio *corpus*.
7. **Disponibiliza o *corpus* para o público** - Foi considerado que a ferramenta possui a funcionalidade de disponibilização de *corpus* se o usuáruio puder tornar seu *corpus* acessível a outras pessoas através da ferramenta. Geralmente esta funcionalidade é implementada nas ferramentas Web, tornando o *corpus* disponível pela Internet.
8. **Permite uso de expressões regulares** - Foi considerado que a ferramenta possui esta funcionalidade se a sintaxe da linguagem de consulta permite uso de expressões regulares na especificação do padrão de busca.
9. **Interface gráfica** - Foi considerado que a ferramenta possui uma interface gráfica se ela possui uma interface com janelas e botões, não demandando que o usuário utilize o *prompt* do computador para execução da ferramenta ou comandos específicos na linha de comando.
10. **Interface Web** - Foi considerado que a ferramenta possui uma interface web se ela pode ser executada a partir de um navegador, sem necessidade de instalação na máquina do usuário.

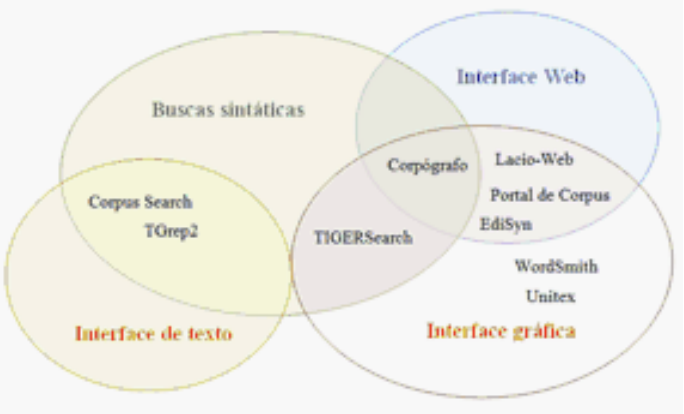
11. **Software Gratuito** - Foi considerado que a ferramenta é gratuita se ela pode ser utilizada ou baixada para instalação e uso, sem requerimento de pagamento.

Quadro 14 - Quadro comparativo entre características de algumas ferramentas de exploração de *corpora.*

	Características										
Ferramentas	**Frequenciador**	**Busca de colocações**	**Concordanciador**	**Busca sintática**	**Suporte à compilação**	***Corpus* do usuário**	**Disponibiliza *corpus***	**Expressões regulares**	**Interface gráfica**	**Interface web**	**Software Gratuito**
WordSmith	X		X		X	X			X		
Unitex	X		X			X		X	X		X
Corpus Search	X		X	X		X		X			X
Tgrep2	X		X	X		X		X			X
TigerSearch	X	X	X	X		X			X		X
Corpógrafo	X		X	X	X	X		X	X	X	X
Ferramentas Lacio-Web	X		X			X	X	X	X	X	X
Portal de *Corpus*	X		X		X	X		X	X	X	X
EdiSyn	X		X						X	X	X

A figura 11 mostra uma divisão entre as ferramentas abordadas considerando as características de interface (web, gráfica e texto) e a funcionalidade de buscas sintáticas.

Figura 11- Recursos disponíveis em algumas ferramentas de exploração de *corpora.*



5 O *CORPUS* HISTÓRICO DO PORTUGUÊS ANOTADO TYCHO BRAHE

O *Corpus* Anotado do Português Histórico *Tycho Brahe* consiste em um *corpus* eletrônico composto de textos escritos em português por autores nascidos entre 1380 e 1845. O desenvolvimento do *corpus* se deu a partir de 1998, no âmbito do Projeto "Padrões Rítmicos, Fixação de Parâmetros e Mudança Linguística" (UNICAMP, 1998), cujo objetivo é investigar a relação entre as mudanças rítmica e sintática no processo que levou do Português Clássico ao Português Europeu Moderno. O objetivo deste *corpus* é disponibilizar publicamente dados históricos do português europeu, permitindo que estudiosos possam ter acesso a informações categoriais e estruturais pertinentes a análises morfossintáticas da língua (GALVES; BRITTO, 2008).

Atualmente, o *corpus* Tycho Brahe conta com 63 textos, que estão disponíveis para pesquisa no site do projeto. 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 (UNICAMP, 1998; GALVES; BRITTO, 2008).

Os textos que compõem o *corpus* Tycho Brahe 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 anotações morfossintática e sintática. As etapas de transcrição, edição e anotação morfossintática são realizadas com o auxílio da ferramenta E-Dictor (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010) – editor de marcação extensível XML.

Através da ferramenta E-Dictor os textos do CTB recebem anotações acerca da estrutura e formatação dos textos, das interferências de edição 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.

5.1 Anotação da estrutura dos textos no CTB

A estrutura XML permite que todas as anotações sejam realizadas e guardadas em camadas em um único arquivo gerado pela ferramenta de edição E-Dictor. Toda a anotação não é *stand-off*, ou seja, é mantida junto com o dado original.

O elemento <format> é usado para informações acerca da formatação, que podem ser do tipo capitular, itálico ou negrito. A figura 12 mostra as possíveis ocorrências de formatação com a respectiva anotação em XML adotada pelo Tycho Brahe.

Figura 12 - Anotação de formatação no *Corpus* Tycho Brahe.

Ocorrência	*Anotação*
capitulares	`<format t="cap"></format>`
itálico	`<format t="i"></format>`
negrito	`<format t="b"></format>`

Fonte: Paixão de Sousa, 2007.

Os textos são divididos em parágrafos, que por sua vez são divididos em sentenças. As marcações para esta divisão são <p> e <s>, respectivamente. Ocorrências como quebra de linha, quebra de página, entre outras, são marcadas com uso da *tag* <sec> e diferenciadas entre si pelo atributo "t". A figura 13 mostra as possíveis ocorrências de estrutura do texto com as respectivas anotações adotadas pelo Tycho Brahe. Informações como número da página, número da linha, entre outras, são anotadas fazendo uso da *tag* <text>. As possíveis ocorrências deste tipo de informação com as respectivas anotações adotadas pelo CTB são mostradas na figura 14 (PAIXÃO DE SOUSA, 2007a).

5.2 Anotação de edições no CTB

Paixão de Sousa (2006) aponta que na etapa de transcrição do texto (passagem do meio físico para o meio digital) existe um grande potencial de perda de informações. A transcrição é naturalmente um processo de interferência no texto, que pode ser em maior ou menor grau. As transcrições com o mínimo possível de interferência editorial são chamadas de edições diplomáticas. Um grau ligeiramente maior de interferência editorial, que inclui a modernização tipográfica ou grafemática e o desenvolvimento das abreviaturas dos textos originais, gera edições chamadas semi-diplomáticas (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010).

Figura 13 - Anotação de divisões do texto no *Corpus* Tycho Brahe

Ocorrência	*Anotação*
parágrafo	
sentença	<s></s>
quebra de linha	<sec t="line"/>
quebra de página	<sec t="pag"/>
quebra de coluna	<sec t="col"/>
capítulo	<sec t="ch"></sec>
prefácio	<sec t="preface"></sec>
prólogo	<sec t="prologue"></sec>
carta	<sec t="letter"></sec>
índice	<sec t="index"></sec>
peça (teatral)	<sec t="play"></sec>
ato	<sec t="act"></sec>
descrição dos personagens	<sec t="char_desc"></sec>
marcações de cena	<sec t="scene_desc"></sec>
nome dos personagens	<sec t="char"></sec>
título	<sec t="title"></sec>
tabela	<sec t="table"></sec>
texto na margem	<sec t="margin"></sec>

Fonte: Paixão de Sousa, 2007a.

Figura 14 - Anotação de elementos do texto no *Corpus* Tycho Brahe

Ocorrência	*Anotação*
número da página	<text_el t="pag_nr"></text_el>
número da linha	<text_el t="line_nr"></text_el>
número do parágrafo	<text_el t="par_nr"></text_el>
cabeçalho da página	<text_el t="pag_head"></text_el>
reclame de pé de página	<text_el t="pag_foot"></text_el>
imagem	<text_el t="image"></text_el>

Fonte: Paixão de Sousa, 2007a.

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 sofrendo interferências maiores que as aceitáveis em edições semi-diplomáticas. No entanto, para estudos filológicos, as características do texto original são importantes e devem ser preservadas. O conflito entre os objetivos filológicos e as peculiaridades do tratamento computacional motivou a criação do projeto “Memórias do

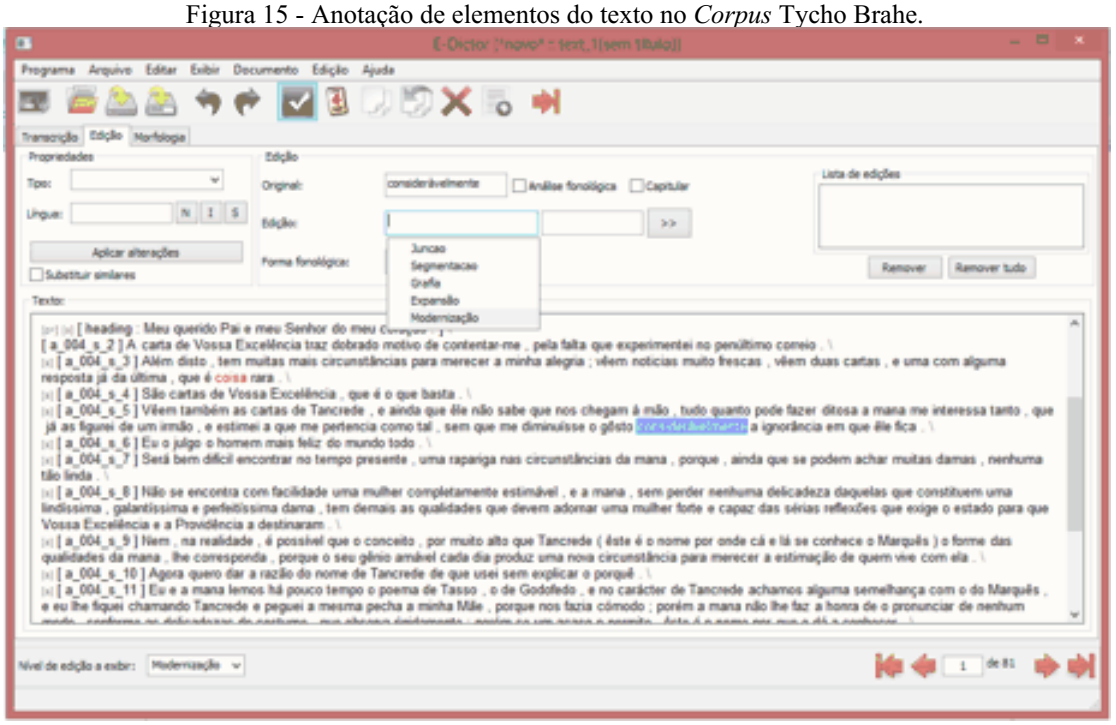
Texto" (PAIXÃO DE SOUSA, 2006), definindo um sistema de anotação com objetivo de permitir a análise linguística por ferramentas computacionais e também de preservar informações filológicas fundamentais no *corpus* Tycho Brahe.

A técnica adotada nesse sistema de anotação consiste em codificar todo o texto com etiquetas XML para as estruturas variantes, possibilitando assim o controle e o mapeamento das intervenções realizadas nos documentos. Dessa maneira, os textos podem ser recuperados de várias formas – em sua forma original ou com as edições realizadas. A inserção das etiquetas também é feita automaticamente pela ferramenta E-Dictor. Embora haja exibição do conteúdo visual do arquivo com as etiquetas inseridas, a interface do software visa evitar o contato direto entre o usuário (editor do texto) e a estrutura XML gerada (PAIXÃO DE SOUSA, 2006). A figura 15 mostra a interface gráfica do E-Dictor que permite a realização destas edições evitando o contato do editor com a estrutura XML subjacente.

As interferências editoriais são realizadas no texto com o objetivo de torná-los legíveis para análises linguísticas automáticas. Assim, faz-se a correção das segmentações vocabulares que não respeitam as fronteiras lexicais, fazendo-se necessária a junção de segmentos de palavras que aparecem separados no texto original e separar/segmentar vocábulos que aparecem escritos de maneira amalgamada (emendada); o desenvolvimento das abreviaturas, a interpreção de trechos de difícil leitura, e a uniformização da grafia dos textos, como em uma edição interpretativa tradicional, também fazem parte dos tipos de intervenção necessária para a elaboração da melhor versão do texto para ser submetida as ferramentas de anotação automática (*tagger* e *parser*).

O sistema adotado para estas interferências funciona "como uma **anotação em camadas** sucessivas: ainda depois da aplicação de novas informações num texto de base, é possível distinguir as diferentes camadas do texto". Desta maneira, nas etapas seguintes é possível recuperar o texto em diferentes versões, segundo a camada de edição escolhida. É possível, por exemplo, recuperar uma versão com todas as uniformizações grafemáticas, sem as modernizações de grafia (PAIXÃO DE SOUSA, 2007a).

Figura 15 - Anotação de elementos do texto no *Corpus* Tycho Brahe.



Fonte: Paixão de Sousa, 2007a.

O arquivo XML com as anotações de edições gerado pelo E-Dictor não é exibido para o usuário na ferramenta, mas é salvo na estrutura de arquivos do computador. A anotação de edição em XML é realizada identificando todas as interferências do editor sobre o texto original com elementos <e>, e os itens originais correspondentes com elementos <o>. O tipo de edição é identificado através da propriedade "t" dos elementos <e> e os tipos possíveis são listados no quadro 15. A numeração dos elementos, identificados pela propriedade "id" são atribuídas automaticamente pela ferramenta. A figura 16 mostra um exemplo de anotação de edição realizada pelo E-Dictor em um texto do Corpus Tycho Brahe (PAIXÃO DE SOUSA; KEPLER; FARIA, 2010).

Quadro 15 - Tipos de edição possíveis para o *corpus* Tycho Brahe e representação na anotação XML.

Tipo de edição	**Atributo de <e>**	**Exemplo**
uniformização grafemática	t="gra"	<e t="gra">serviço</e><o>ſeruiço</o>
separação de vocábulos	t="seg"	<e t="seg"> e legitimos </e><o> elegitimos</o>
junção de vocábulos	t="jun"	<e t="jun">que ſe</e><o>queſe</o>
expansão de abreviatura	t="exp"	<e t="exp">Vossa Mercê</e><o>V.M.</o>
uniformização de pontuação	t="punc"	<e t="punc"> >> </e><o> " </o>
modernização de grafia	t="mod"	<e t="mod">inclita</e><o>inclita</o>
Correções	t="cor"	<e t="cor">depois</e><o>deqois</o>

Fonte: Adaptado de Paixão de Sousa (2007a), atualizado com base em Paixão de Souza; Kepler; Faria (2010) .

Figura 16 - Trecho de anotação de edições XML gerada pelo E-Dictor

Fonte: Paixão de Souza; Kepler; Faria, 2010.

5.3 Anotação morfossintática no CTB

O conjunto de etiquetas morfossintáticas para o português foi proposto por Britto; Finger e Galves (1998) e utilizado no desenvolvimento do *tagger* híbrido de Finger (2000) e depois pelo *tagger* probabilístico de Kepler e Finger (2006). Tal conjunto foi projetado com o intuito de ser compatível com um tratamento computacional do *corpus*, mais especificamente com o treinamento do etiquetador morfossintático automático. Um requisito para reduzir a complexidade computacional era manter o conjunto de etiquetas reduzido, mas este preceito deveria ser também compatível com a necessidade de abarcar a rica morfologia do português (GALVES; BRITTO, 2008).

O trabalho trouxe como resultado um sistema de anotação morfossintática em sub-níveis, formado por dois grupos básicos de etiquetas. O grupo das etiquetas categoriais é utilizado para classificação do item lexical segundo a classe da palavra a que pertence. O grupo das etiquetas flexionais é utilizado articulando-as às categorias por meio de diacríticos, podendo ser de natureza verbal, designadores de informações modo-temporais ou não-verbal, indicadoras de traços flexionais de gênero e número. Assim, a etiqueta que o sistema propõe é composta por uma parte principal que indica a parte da classe POS ao qual o item lexical pertence, podendo ser acompanhada ou não de uma parte secundária, que especifica um subgrupo de determinada classe, ou vários traços flexionais carregados pelo item. O sinal diacrítico "-" conecta partes primárias e secundárias uns com os outros, enquanto que "+" combina as classes POS quando se aplica mais do que um, como em contrações por exemplo. (1), (2) e (3) mostram exemplos de trechos anotados morfossintaticamente com o conjunto de etiquetas proposto (GALVES; BRITTO, 2008; BRITTO; FINGER; GALVES, 1998).

.

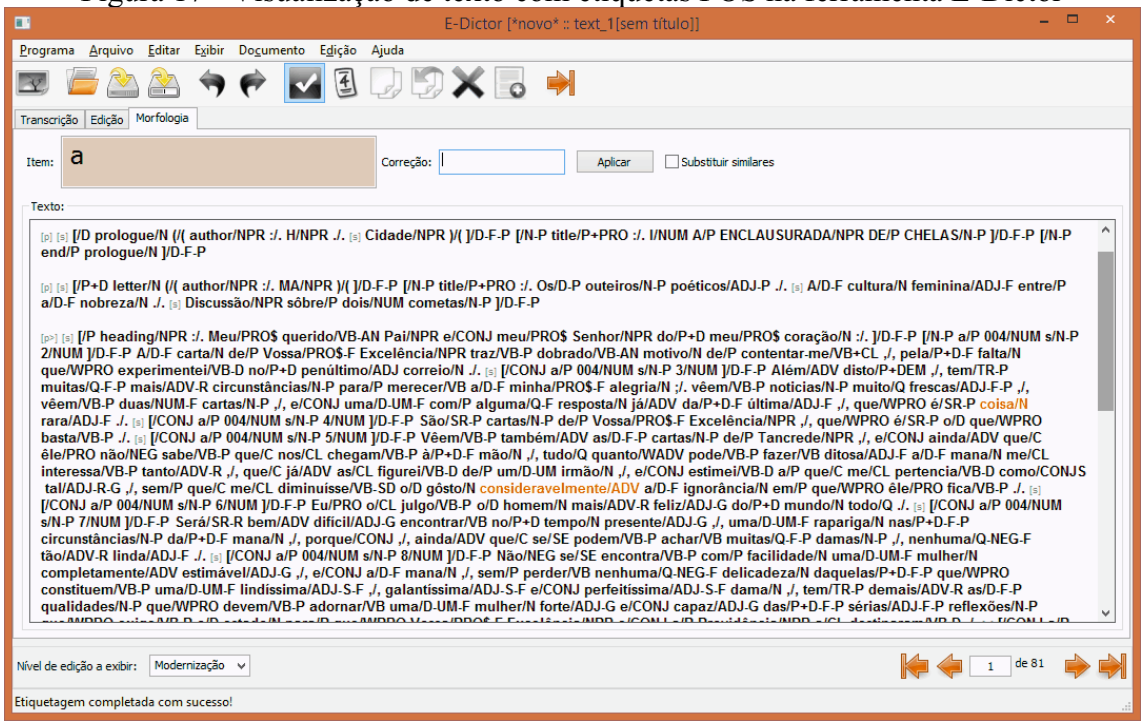
(1) o/D problema vs. um/D-UM problema

(2) os/D-P belos/ADJ-P campos/N-P

(3) Perto/ADV da/P+D-F Cidade/NPR principal/ADJ-G da/P+D-F Lusitânia/NPR está/ET-P uma/D-UM-F graciosa/ADJ-F Aldeia/NPR

Com o desenvolvimento do E-Dictor, o algoritmo de etiquetagem morfossintática no formato POS foi embutido na programação da ferramenta, tornando este processo transparente para o usuário. O texto anotado no formato POS serve como entrada para uma transformação na anotação XML. O E-Dictor realiza a conversão para etiquetas na linguagem XML e as mantém no mesmo arquivo com as edições. O texto com etiquetas POS pode ser visualizado no E-Dictor e um arquivo correspondente pode ser salvo no computador do usuário. A figura 17 mostra uma visualização de um texto com etiquetas POS na ferramenta.

Figura 17 - Visualização de texto com etiquetas POS na ferramenta E-Dictor



Fonte: Paixão de Souza; Kepler; Faria, 2010.

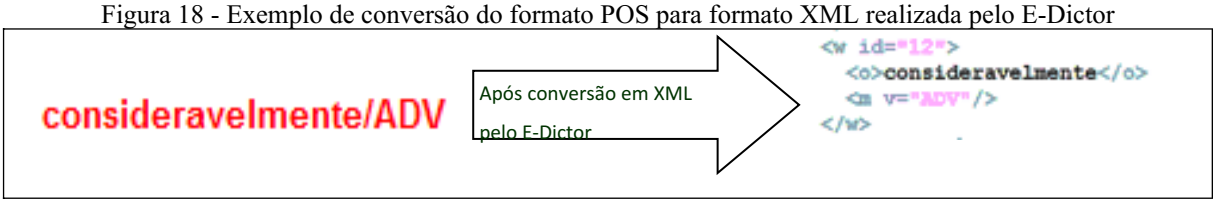
A identificação de informação morfossintática em XML se dá pela marcação do item lexical com o elemento <m>. A propriedade "v" marca o valor da categoria lexical. O quadro 16 mostra um trecho da anotação gerada pelo E-Dictor para um texto do *corpus* DOViC. As categorias lexicais recebem os mesmos valores das etiquetas no formato POS. Por exemplo, um item lexical etiquetado como advérbio no formato POS é concatenado à etiqueta /ADV. No formato XML gerado pelo E-Dictor, este item seria anotado com a seguinte marcação: <m

v="ADV"/>. A figura 18 ilustra este exemplo com um trecho de texto pertencente ao Corpus DOViC.

Quadro 16 - Trecho de texto do Corpus DOViC com anotação morfossintática em XML gerada pelo E-Dictor

```
<w id="22">
   <o>descrever</o>
   <m v="VB"/>
 </w>
 <w id="23">
   <o>oseu</o>
   <e t="seg">o seu</e>
   <m v="D">o</m>
   <m v="PRO$">seu</m>
 </w>
 <w id="24">
   <o>desenvolvimento</o>
   <m v="N"/>
 </w>
 <w id="25">
   <o>histÃ³rico</o>
   <m v="ADJ"/>
 </w>
```

Figura 18 - Exemplo de conversão do formato POS para formato XML realizada pelo E-Dictor



5.4 Anotação sintática no CTB

A versão atual do programa E-Dictor (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 de texto (.txt) anotado no formato POS, com as etiquetas morfossintáticas, e gera como saída um outro arquivo texto no formato *Penn TreeBank* (seção 3.5.3). O *parser* foi desenvolvido pela Universidade da Pensilvânia e seu treinamento para o português brasileiro foi feito por pesquisadores da Unicamp. A figura 19 mostra o trecho de um texto do *Corpus* Tycho Brahe anotado sintaticamente pelo *parser* citado.

Figura 19 - Trecho de um texto do *Corpus* Tycho Brahe anotado sintaticamente pelo *parser* da Pensilvânia

```
( (CODE <P_03>))

( (IP-MAT (NP-SBJ *pro*)
          (NP-VOC (NPR Senhor))
          (, :)
          (VB Ofereço)
          (PP (P a)
              (NP (PRO$-F Vossa) (NPR Majestade)))
          (NP-ACC (D-F-P as)
                  (NPR-P Reflexões)
                  (PP (P sobre)
                      (NP (D-F a)
                          (N vaidade)
                          (PP (P de@)
                              (NP (D-P @os) (N-P homens))))))
          (. ;)) (ID A_001_PSD,03.1))

( (IP-MAT (NP-SBJ (DEM isto))
          (SR-P é)
          (NP-ACC (D o)
                  (ADJ mesmo)
                  (CP-CMP (WNP-4 0)
                          (C que)
                          (IP-SUB (NP-ACC *T*-4)
                                  (NP-SBJ *exp*-5)
```

5.5 *Corpora* com metodologia de anotação baseada no *Corpus* Tycho Brahe

Diversos *corpora* utilizam a metodologia de anotação empregada no Corpus Tycho Brahe, seja a adoção apenas do sistema de anotação sintática, que é embasado no Penn-Helsinki Parsed Corpus of Middle English (PPCME) , ou também a adoção do sistema de anotação de edições elaborado por Paixão de Souza (2007a), da ferramenta E-Dictor, que possui embutido o etiquetador morfossintático e sistema de etiquetas desenvolvidos para aquele *corpus*.

Os seguintes *corpora* são compilados ou anotados nos mesmos moldes do Tycho Brahe e portanto, podem beneficiar-se do compartilhamento de recursos computacionais desenvolvidos com embasamento nesta metodologia de compilação de *corpus*:

- ***Corpus* Eletrônico de Documentos Históricos do Sertão (CE-DOHS)** - O *corpus* é formado por cartas pessoais brasileiras do período 1808 a 2000. A maioria do acervo tem origem na grande área do semi-árido baiano, mas também há documentos provenientes de outras áreas da Bahia e diversas regiões do Brasil. O *corpus* segue uma tecnologia de edição inspirada no CTB e as versões são disponibilizadas baseando-se em critérios estabelecidos pelo Projeto Para a História do Português Brasileiro (PHPB) (CE-DOHS, 2010).
- **Parsed Historical *Corpus* (IcePaHC) -** *Corpus* da língua islandesa, composto por textos que cobrem a história da Islândia do século XII até o presente, englobando cerca

de 100.000 palavras por cada século. Os textos são anotados sintaticamente no formato *Penn TreeBank* (RÖGNVALDSSON; INGASON; SIGURDSSON, 2011).

- **Penn Parsed *Corpora* of Historical English** - Penn Parsed *Corpora* of Historical English é um conjunto de três *corpora*: a segunda edição do Penn-Helsinki Parsed Corpus of Middle English (PPCME2), o Penn-Helsinki Parsed *Corpus* of Early Modern English (PPCEME), e o Penn Parsed Corpus of Modern British English (PPCMBE). Os textos estão disponíveis em três formas: texto simples, textos etiquetados no formato POS e anotação sintática no formato *Penn TreeBank*. Os *corpora* foram projetados para serem usados por estudiosos da história do Inglês, especialmente a sintaxe histórica da língua. Estão disponíveis para pesquisadores individuais, grupos de pesquisa e bibliotecas, mediante aceitação de licença de uso (KROCH; TAYLOR, 2000; KROCH; SANTORINI; DIERTANI, 2004; KROCH; DIERTANI, 2010).
- ***Corpus* Dialectal para o Estudo da Sintaxe (CORDIAL-SIN) -** O *Corpus Dialectal para o Estudo da Sintaxe* (CORDIAL-SIN) é um *corpus* oral, compilado com o objetivo de investigar a variação sintática dialetal do português europeu . Seu arquivo sonoro contém cerca de 4.500 horas de gravações, obtidas em mais de 200 localidades de Portugal. A transcrição dos dados de fala resultam num *corpus* com cerca de 600.000 palavras (CORDIAL-SIN, 2014).

5.6 Buscas automáticas no *Corpus* Tycho Brahe

As buscas automáticas no *Corpus* Tycho Brahe são realizadas com auxílio da ferramenta *Corpus Search*. Os textos do *corpus* estão disponíveis para download no site do projeto Tycho Brahe. Para ter acesso ao *download* do *corpus*, o usuário deve fazer um cadastro no portal. As versões disponíveis são de textos transcritos como no original, ou textos com anotação morfossintática no formato POS, ou com anotação sintática no formato *Penn TreeBank*.

Com o *download* dos textos do *corpus*, o pesquisador pode realizar buscas morfossintáticas nos arquivos POS com o conhecimento de alguma linguagem de programação ou ferramenta que faça buscas em textos. Ferramentas genéricas de buscas em textos fornecem recursos limitados como localizar uma *string* dentro de um texto. A realização de buscas mais elaboradas como por exemplo, sentenças contendo verbos seguidos de advérbios, demandam ferramentas específicas para pesquisas linguísticas. Buscas sintáticas podem ser realizadas com o download do *Corpus Search* ou outra ferramenta e o conhecimento da linguagem de consulta desses softwares.

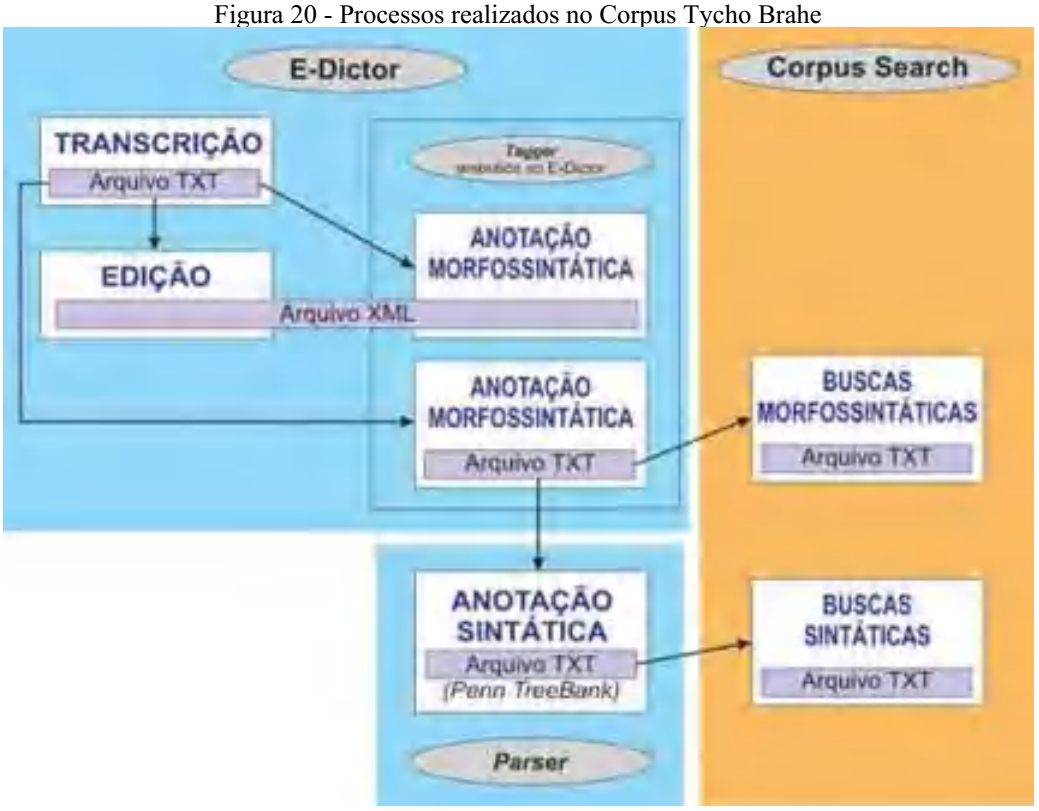
Outra maneira de realizar buscas no *corpus* Tycho Brahe é através da interface gráfica disponibilizada no site. As buscas são realizadas com a ferramenta *Corpus Search*, mas de maneira transparente para o usuário. Usando a interface do site, não é preciso que instale a ferramenta em seu computador nem tenha conhecimento de como usá-la. No entanto, as buscas são restritas a uma estrutura linear, realizadas por classes de palavras. Através da interface do site, não é possível realizar pesquisas na estrutura sintática dos textos. Além da consulta gráfica, o site também disponibiliza a opção de consulta manual, o que dá maior flexibilidade caso o usuário tenha conhecimento da linguagem utilizada para consultas na ferramenta *Corpus Search*.

Os textos com a anotação morfossintática em XML realizada pelo E-Dictor não são disponibilizados no site do Tycho Brahe. Há possibilidade de obter o arquivo no formato XML com anotações de edições através da opção de exibição do código-fonte do navegador. Também é possível a visualização do texto na versão original ou modernizada.

5.7 Resumo dos processos com textos no *corpus* Tycho Brahe

A figura 20 ilustra um resumo dos processos realizados com os textos do Corpus Tycho Brahe. Os retângulos brancos representam os processos realizados. O texto em azul indica o nome do processo e o texto no retângulo interno ao processo representa o formato da saída gerada. Os retângulos maiores em azul ou laranja representam a ferramenta computacional utilizada. A cor laranja foi utilizada para diferenciar a ferramenta de busca de outro tipo de ferramenta. As setas indicam a interação entre os processos, sendo que a direção indica qual arquivo fornece entrada para outro processo.

Figura 20 - Processos realizados no Corpus Tycho Brahe



6 O *CORPUS* DIGITAL DOVIC

O *Corpus* DOViC (*Corpus* de Documentos Oitocentistas de Vitória da Conquista) é um *corpus* digital de documentos manuscritos do século XIX que está sendo compilado no âmbito do projeto "Memória conquistense: implementação de um *corpus* digital" (NAMIUTI, 2013) em parceria com os projetos: (i) "Corpora Digitais Para a História do Português Brasileiro – região Sudoeste da Bahia: Aliança PHPB – Tycho Brahe" (SANTOS; NAMIUTI, 2010); (ii) "Sintaxe diacrônica em *corpus* eletrônico: do português pré-clássico às variantes modernas" (NAMIUTI, 2010); e, (iii) "Memória conquistense: recuperação de documentos oitocentistas para a implementação de um corpus digital" (SANTOS; NAMIUTI, 2009).

Os documentos manuscritos que compõem o *corpus* estão sob a guarda do Fórum de Vitória da Conquista-Bahia e foram catalogados e arquivados pelos colaboradores dos projetos parceiros supra-citados.

Dentre os documentos que compõe o banco de textos do corpus DOViC, encontram-se documentos avulsos e livros de notas que pertencem a uma série que vai de 1 a 21 e datam desde 1841 (livro 1) a 1888 (livro 21) - todos provenientes da Imperial Villa da Victoria (atual cidade de Vitória da Conquista) (NAMIUTI, 2013; SANTOS; NAMIUTI, 2009).

Os textos notariais componentes dos livros são de natureza variada: Cartas de alforria; Testamentos; Procurações; Matrículas de escravos; Escrituras de imóveis; Atas de eleições municipais. No período citado, vários negros de diversas etnias e de diversos lugares da África foram trazidos como escravos para a região. Produziu-se na época vasta documentação manuscrita relacionada à escravidão, que é mantida nos cartórios da região. A compilação do *corpus* DOViC demonstra-se, portanto, relevante não apenas para o estudo da língua mas também para a preservação do patrimônio histórico da cidade. Tais documentos, além de serem testemunhos da história da língua portuguesa no Brasil, registram a história social da região e, por extensão, do Brasil (NAMIUTI, 2013).

A disponibilização de documentos históricos para fins de pesquisa requer que tais documentos possam ser facilmente acessados e analisados. Como o *corpus* DOViC se trata de um *corpus* de documentos manuscritos que apresentam complexidades relacionadas ao acesso, à forma, à fragilidade e à raridade, existem etapas anteriores à transcrição e compilação do corpus que precisam ser consideradas. Para lidar com tais complexidades, a alternativa utilizada é a de tornar o documento histórico um Documento Digital (DD) através de técnicas profissionais da fotografia, realizando, desta forma, a transposição material do Documento Físico (DF) para o meio digital reproduzindo fielmente o documento original para ser acessado

de forma imagética (SANTOS, 2010; NAMIUTI; SANTOS; LEITE, 2011; NAMIUTI ET. AL, 2013).

Ao considerar o uso da fotografia enquanto um meio científico de transposição do texto em papel para o digital, Santos (2010), Namiuti, Santos e Leite (2011), Namiuti et. al (2013) e Santos e Brito (2014) procuram enfatizar a necessidade de se colocar na posição de um Pesquisador Formador de Corpora (PFC) e não apenas de um pesquisador pragmático.

O trabalho para a criação do corpus DOViC perpassa as seguintes questões centrais:

> "Qual a viabilidade do uso da fotografia para a captação fidedigna de documentos para compor corpora digitais, visando estudos linguísticos e científicos?", e "Quais são as complexidades intervenientes no processo de digitalização de documentos físicos escritos?", defendendo a hipótese de que, desde que metodicamente controlada em suas fases de captura, catalogação, edição, armazenamento, e leitura, a Fotografia apresenta-se como forma altamente viável e producente de digitalização, permitindo à Linguística, ou outra ciência, acessar imageticamente, de modo confiável, o documento não disponível no local da pesquisa. (NAMIUTI, SANTOS, COSTA, FARIAS, 2013, p. 13)

A Fotografia, enquanto um meio científico de transposição do texto em papel para o meio digital garante a fidedignidade necessária para a pesquisa científica e permite que as etapas subsequentes de pesquisa com o *corpus* sejam possíveis.

Esta etapa anterior à transcrição é desenvolvida utilizando-se do *Método Lapelinc* (SANTOS, 2009), um método específico criado para o corpus DOViC para gerar os documentos originais em formato de imagem digital. Tal método possui etapas como: (i) Registro e Controle das informações sobre a fonte original gerando um catálogo de textos; (ii) Captura fotográfica dupla da imagem do original e registro dos dados da coleta; (iii) Catalogação no Banco de dados DOViC das imagens componentes do documento; (iv) Edição das imagens catalogadas para enviá-las para as etapas de compilação do *corpus* (transcrição, edição, anotação morfossintática, transformações e *parser*).

Para o controle e recuperação das informações relacionadas aos formatos originais, foi projetada dentro do *Método Lapelinc,* a *Mesa Cartesiana* - uma espécie de placa plana, de cor cinza, quadriculada milimetricamente, características que servem para recuperar, no computador, a exata medida do papel no mundo físico. Sobre a mesa o documento original é assentado e em seguida são colocados escalas de tom de cores, informações catalográficas, paginação e sequência. A página do documento pode tanto ser apresentada no computador com todas essas informações como também de forma recortada, mostrando apenas a parte manuscrita.

A figura 21 mostra a fotografia do verso da folha 40 do livro 1, na qual registra-se a primeira parte de um documento manuscrito “A carta de liberdade da cabra de nome Sofia”, componente do *corpus* DOViC, capturado pelo Método Lapelinc. A figura 22 mostra as duas folhas que compõem esse mesmo documento.

Figura 21 - Imagem de documento manuscrito do corpus DOViC em mesa cartesiana.

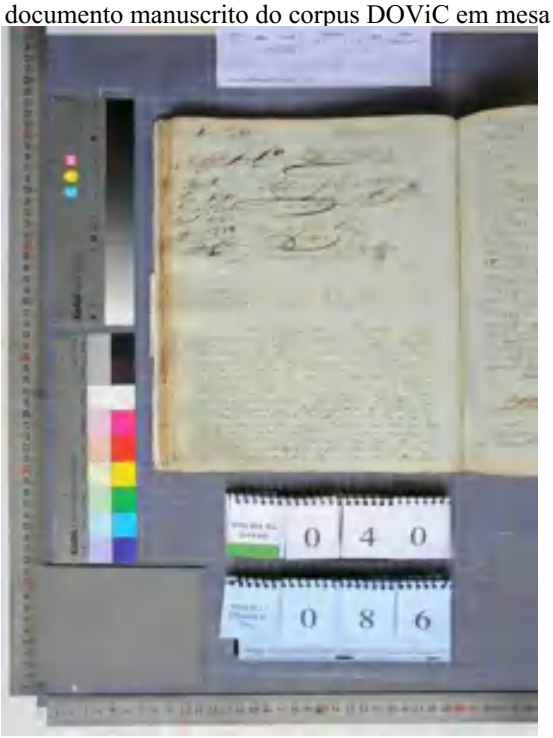


Figura 22 - Imagens de documento manuscrito do *corpus* DOViC.

As informações para a identificação e recuperação dos documentos originais digitais para posterior transcrição e edição são registradas em um catálogo ilustrado. A figura 23 mostra uma entrada do catálogo do DOViC-Beta.

Figura 23 - Imagem do catálogo do Livro 1 do corpus DOViC.

Os textos do corpus DOViC são transcritos, editados e anotados nos mesmos moldes do *Corpus Histórico do Português Tycho Brahe*, utilizando a mesma ferramenta (E-Dictor) e o mesmo esquema de anotação, conforme ilustrado nas figuras 24, 25 e 26.

Figura 24 - Imagem da visão do texto transcrito e editado em XML com a ferramenta E-Dictor.

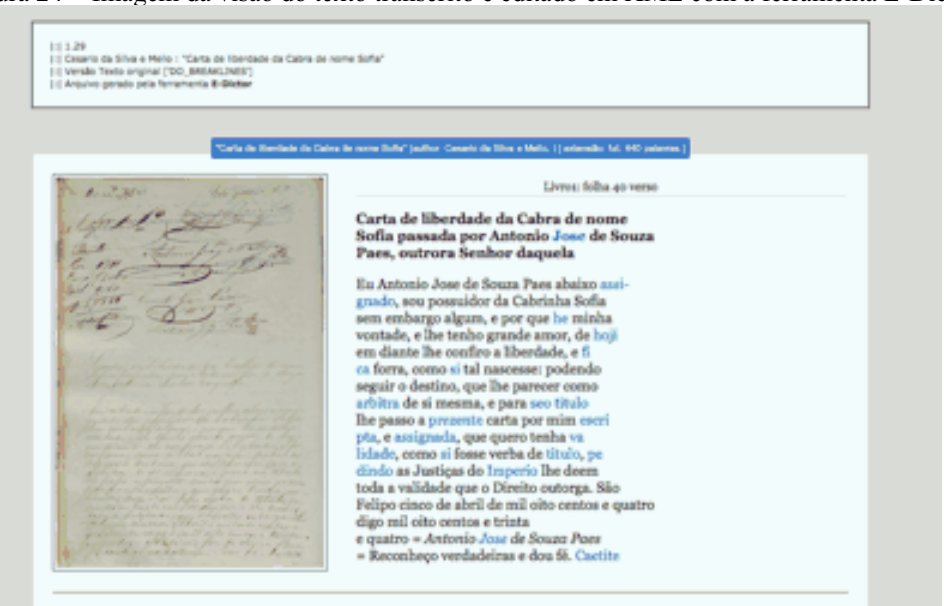


Figura 25 - Imagem da visão modernizada do texto editado em XML com a ferramenta E-Dictor.

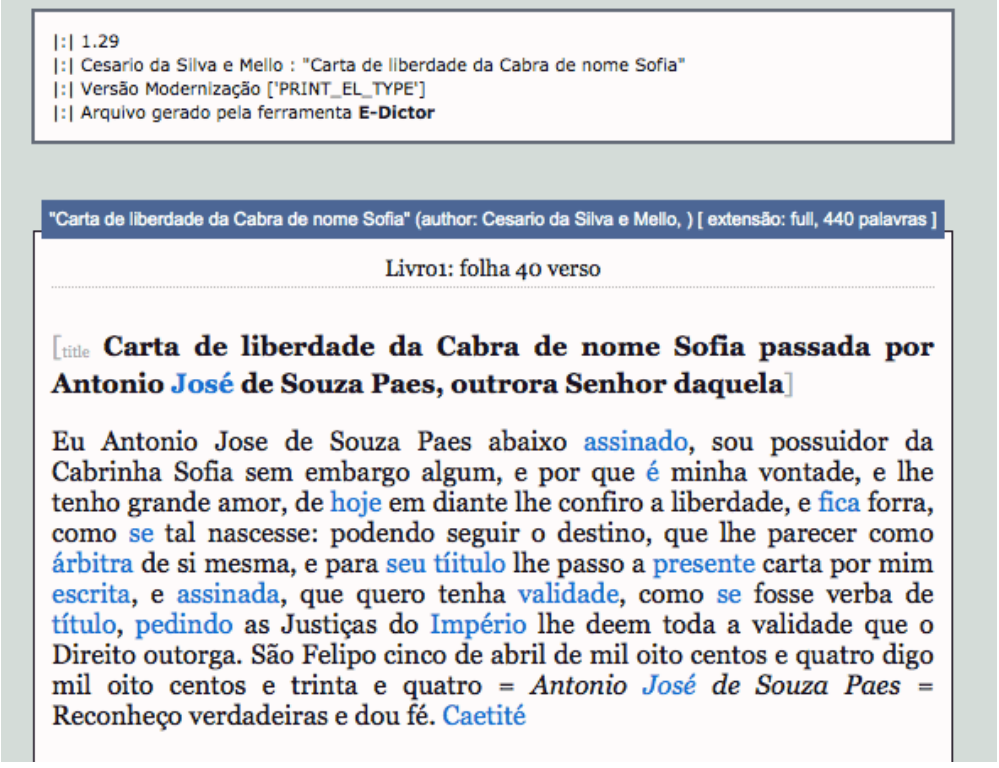
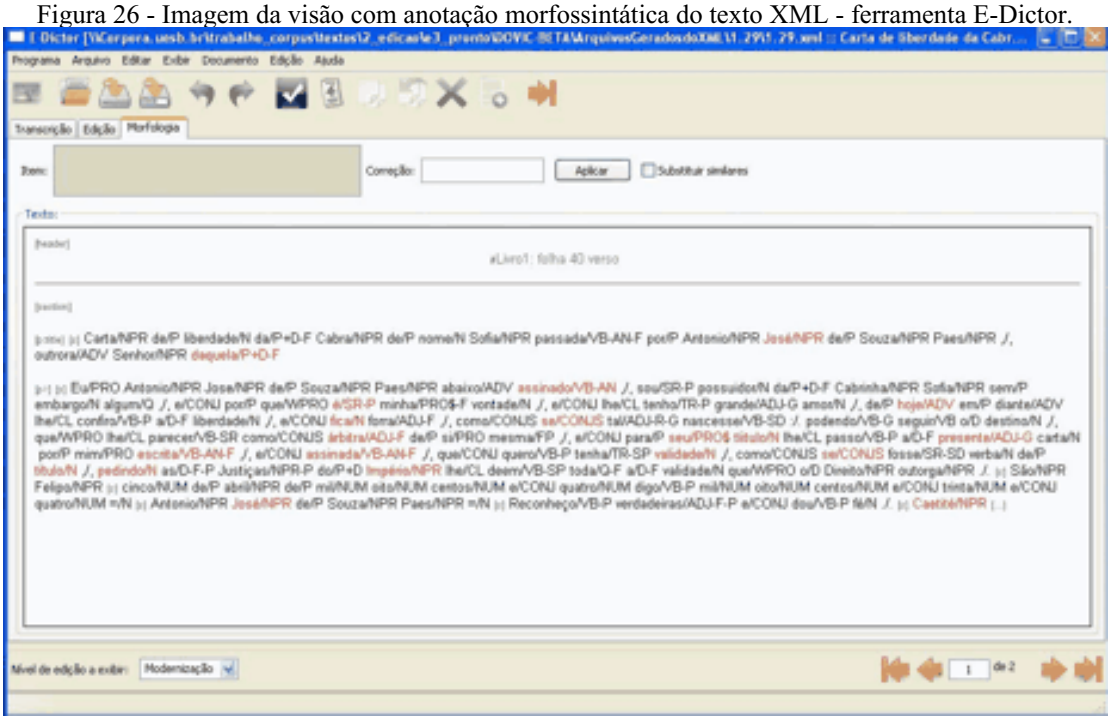


Figura 26 - Imagem da visão com anotação morfossintática do texto XML - ferramenta E-Dictor.



Para realizar buscas sintáticas no *corpus* DOViC, uma ferramenta como o *Corpus Search* pode ser utilizada, já que o *corpus* segue modelo Tycho Brahe com anotação sintática no formato *Penn TreeBank*. No entanto, não existem outras ferramentas, além da que desenvolvemos no âmbito deste mestrado, para pesquisas morfossintáticas nos arquivos XML gerados pelo E-Dictor. As diversas visões possíveis de serem obtidas a partir das edições anotadas no arquivo XML não podem ser realizadas pelos usuários do *Corpus Tycho Brahe*, pois as pesquisas nos textos do *corpus* são aplicadas apenas nos arquivos texto gerados ou pelo E-Dictor (versão POS do texto), ou pelo *parser*, deixando, assim, de explorar ao máximo as potencialidades da anotação em camadas XML.

Atualmente, o *corpus* DOViC é formado por oito livros, os quais são compostos por uma gama de documentos já catalogados. Deste total, foram capturadas sessenta e cinco imagens e transcritos trinta documentos (todos do tipo "carta de alforria"). Destes trinta documentos transcritos, seis passaram pelo processo de edição e anotação morfossintática, utilizando-se da ferramenta E-Dictor. Tais documentos possuem, portanto, uma versão em arquivo de texto com o registro das edições realizadas e a anotação morfossintática em camadas XML. Ainda não há nenhum documento do *corpus* DOViC com anotação sintática, esta será a próxima etapa de implementação da versão beta do *corpus*.

A seguir apresentaremos a ferramenta desenvolvida no âmbito desta pesquisa, visando soluções para a construção, publicação e busca de dados do corpus DOViC: o WebSinC. O projeto desta ferramenta organiza o *pipeline* da construção do *corpus*, permite seu controle e

gerenciamento, além de potencializar a utilização dos recursos permitidos pela anotação em camadas.

7 WEBSINC: FERRAMENTA WEB PARA BUSCAS AUTOMÁTICAS NO *CORPUS* DOVIC

Este capítulo apresenta o software web desenvolvido nesta pesquisa, o qual denominamos WebSinC, construído para disponibilizar o *corpus* DOViC na Internet e fornecer o recurso de buscas automáticas nos textos do *corpus*, tanto por categorias sintáticas quanto por categorias morfossintáticas.

O nome WebSinC foi escolhido por remeter à característica do software como uma ferramenta Web, ao recurso implementado de buscas sintáticas e morfossintáticas com a parte "Sin" e aos projetos de construção de *corpora* aqui contemplados (Memória Conquistense e Tycho Brahe) com a letra "C" na parte final do nome.

O WebSinc provê uma interface gráfica de fácil utilização pelo usuário, que pode ser um pesquisador interessado no *corpus*, ou um administrador do sistema para gerenciá-lo, fazendo *upload* e cadastro de documentos e fotografias, tornando o *corpus* disponível pela 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 linguísticas, 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.

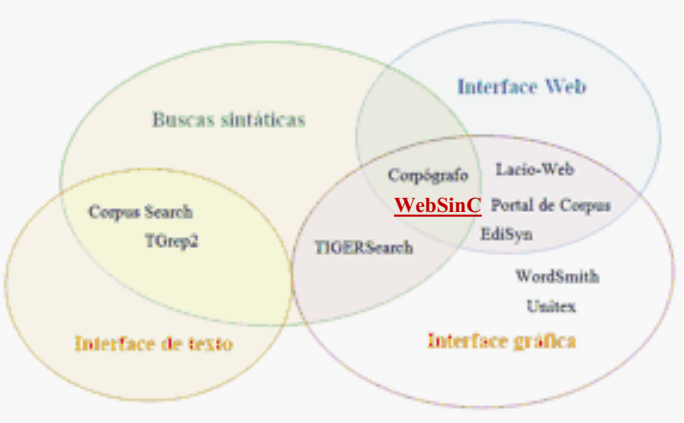
Apresentamos no quadro 17 um comparativo de características da ferramenta WebSinC com as ferramentas já apresentadas no capítulo três. Adicionamos a característica de buscas baseadas no padrão de anotação em XML. O WebSinC é o software na primeira linha em destaque das demais.

Quadro 17- Quadro comparativo entre características de algumas ferramentas de exploração de *corpora*.

	Características											
Ferramentas	**Frequenciador**	**Busca de colocações**	**Concordanciador**	**Busca sintática**	**Suporte à compilação**	***Corpus* do usuário**	**Disponibiliza *corpus***	**Expressões regulares**	**Interface gráfica**	**Interface web**	**Software Gratuito**	**Buscas em XML**
WebSinc	X		X	X			X		X	X	X	X
WordSmith	X		X		X	X			X			
Unitex	X		X			X		X	X		X	
Corpus Search	X		X	X		X		X			X	
Tgrep2	X		X	X		X		X			X	
TigerSearch	X	X	X	X		X			X		X	
Corpógrafo	X		X	X	X	X		X	X	X	X	
Ferramentas Lacio-Web	X		X			X	X	X	X	X	X	
Portal de *Corpus*	X		X		X	X		X	X	X	X	
EdiSyn	X		X						X	X	X	

A figura 27 mostra uma divisão entre as ferramentas abordadas no capítulo três acrescentando a ferramenta WebSinC desenvolvida nesta pesquisa. Considerando as características de interface (web, gráfica e texto) e a funcionalidade de buscas sintáticas, a WebSinC assemelha-se mais à ferramenta Corpógrafo. No entanto, tem as funcionalidades adicionais de disponibilização do *Corpus* na Internet e o recurso de buscas no padrão XML, características ausentes no Corpógrafo e de fundamental importância para o *corpus* DOViC.

Figura 27 - Recursos disponíveis em algumas ferramentas de exploração de *corpora*.



As seções seguintes apresentam alguns artefatos do software gerados nas fases de análise, projeto e implementação da ferramenta.

7.1 Análise e modelagem do software

No levantamento de requisitos foram identificados diversos requisitos a que o software deveria atender. Os requisitos funcionais representam as funcionalidades externamente observáveis do sistema. Para a aplicação desenvolvida, foram identificados dois atores que podem interagir com o sistema: o administrador e o usuário do *Corpus*. Os requisitos funcionais identificados foram:

1. Permitir ao administrador manter o cadastro de documentos do *corpus*, fazendo o *upload* das imagens e textos anotados em XML para o servidor;
2. Permitir ao administrador manter o cadastro de usuários do sistema, podendo estes ser pesquisadores do laboratório Lapelinc (Labotatório de Pesquisa em Linguística de *Corpus*);
3. Permitir ao administrador converter um arquivo com anotação sintática no formato *Penn TreeBank* para o formato XML, a fim de possibilitar buscas automáticas baseadas em categorias sintáticas;
4. Permitir ao administrador a impressão de relatórios com informações diversas como: relatório de textos já armazenados no Banco de Dados, relatório de textos com imagem de capa, relatório de textos filtrados por autor, data, editor, etc.;
5. Permitir ao administrador a manutenção do cadastro de cidades, estados e países;
6. Permitir ao administrador a manutenção do cadastro de tipos de documentos, autores, materiais de capa e forro e cores de capas;
7. Permitir ao administrador a manutenção do cadastro de locais de depósito;

8. Permitir ao administrador a manutenção do cadastro de tipos de colaboradores;
9. Permitir ao administrador a manutenção do cadastro de colaboradores;
10. Permitir ao usuário cadastrar-se no sistema para ter acesso ao *Corpus;*
11. Permitir ao usuário visualizar as imagens dos manuscritos do *corpus;*
12. Permitir ao usuário visualizar os textos transcritos do *corpus;*
13. Permitir ao usuário visualizar as informações gerais, a ficha catalográfica e o léxico de edições dos textos do *corpus;*
14. Permitir ao usuário visualizar os textos do *corpus* nas versões editadas com ou sem modernização de grafia;
15. Permitir ao usuário realizar pesquisas nos textos com base em filtros por autor, data, tipo de documento, etc.;
16. Permitir ao usuário realizar o *download* dos arquivos transcritos do *Corpus* no formato TXT;
17. Permitir ao usuário realizar o *download* dos arquivos anotados no formato XML (anotação morfossintática e de edições);
18. Permitir ao usuário realizar o *download* dos arquivos anotados sintaticamente no formato XML;
19. Permitir ao usuário realizar buscas morfossintáticas no textos do *corpus;*
20. Permitir ao usuário realizar buscas sintáticas no textos do *corpus;*

As figuras 28 e 29 mostram os diagramas de casos de uso elaborados para o software na fase de análise.

Figura 28 - Diagrama de casos de uso para o ator administrador.

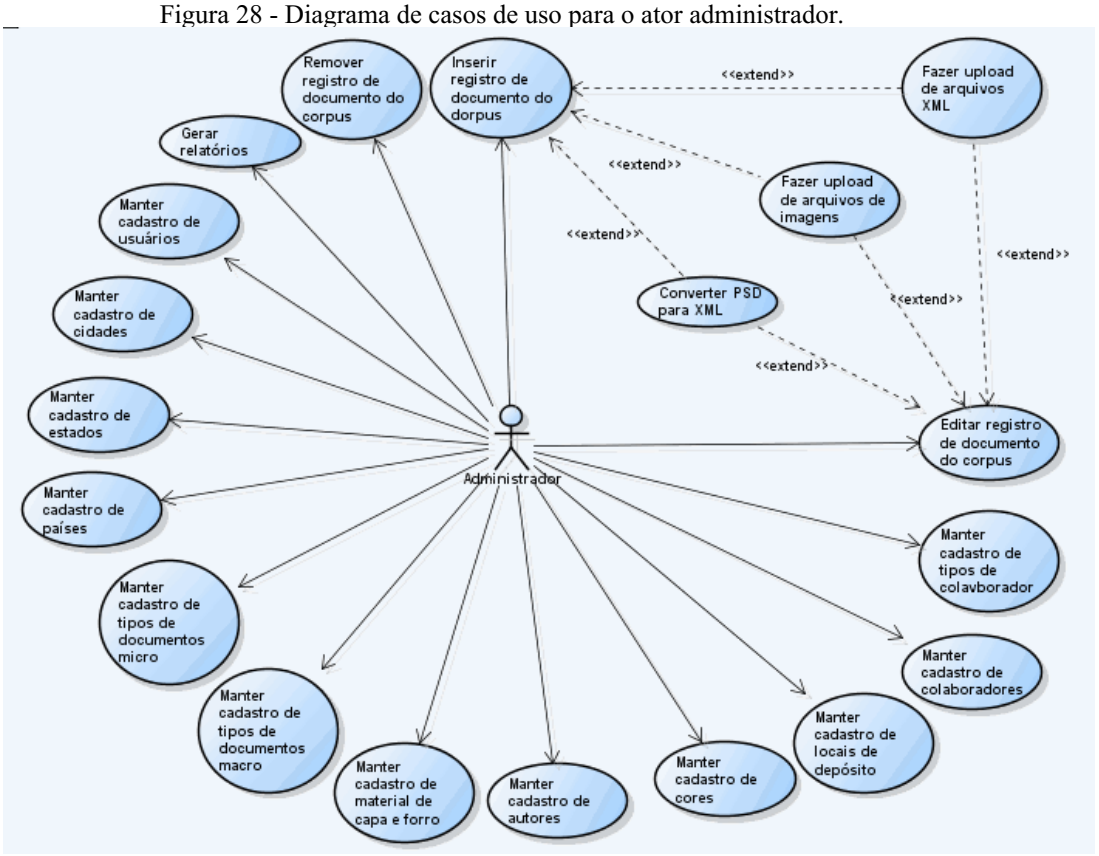
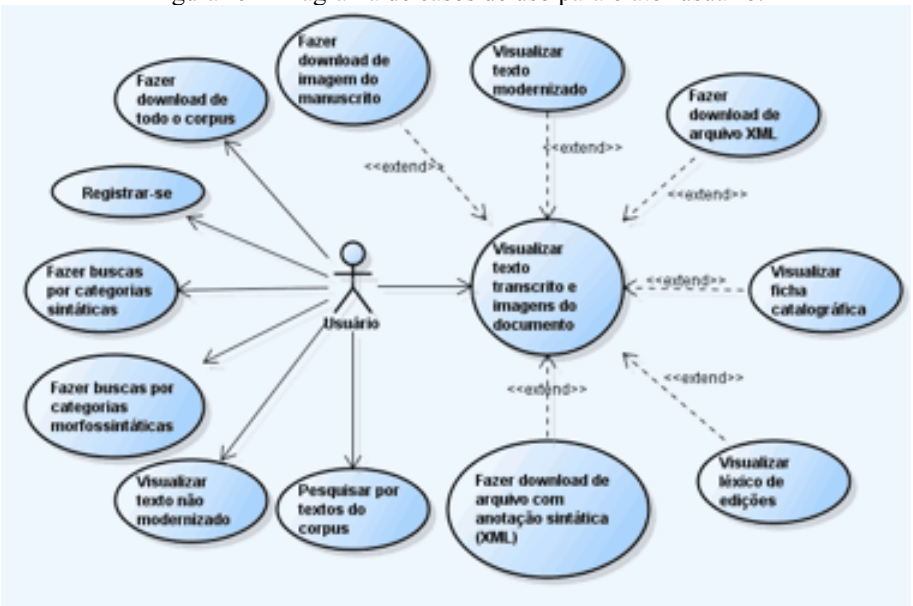


Figura 29 - Diagrama de casos de uso para o ator usuário.



As elipses no diagrama representam casos de uso e as linhas contínuas ilustram uma comunicação entre o ator e os mesmos. A direção da seta denota o sentido das informações. As linhas tracejadas representam relacionamentos de comunicação entre casos de uso que indicam comportamento opcional (representado pelo estereótipo *extend*). Por exemplo, ao visualizar um

texto transcrito e suas imagens, o usuário pode ou não realizar o caso de uso "Visualizar léxico de edições".

Requisitos não funcionais expressam restrições que o software deve atender mas que não são funcionalidades externamente observáveis. Os requisitos não funcionais identificados foram:

1. O software deve ser Web, uma vez que o *corpus* deve estar disponível na Internet;
2. O software deve ser acessível nos principais navegadores: Internet Explorer, Firefox, Crhome, Safari e Opera.

7.2 Projeto do software

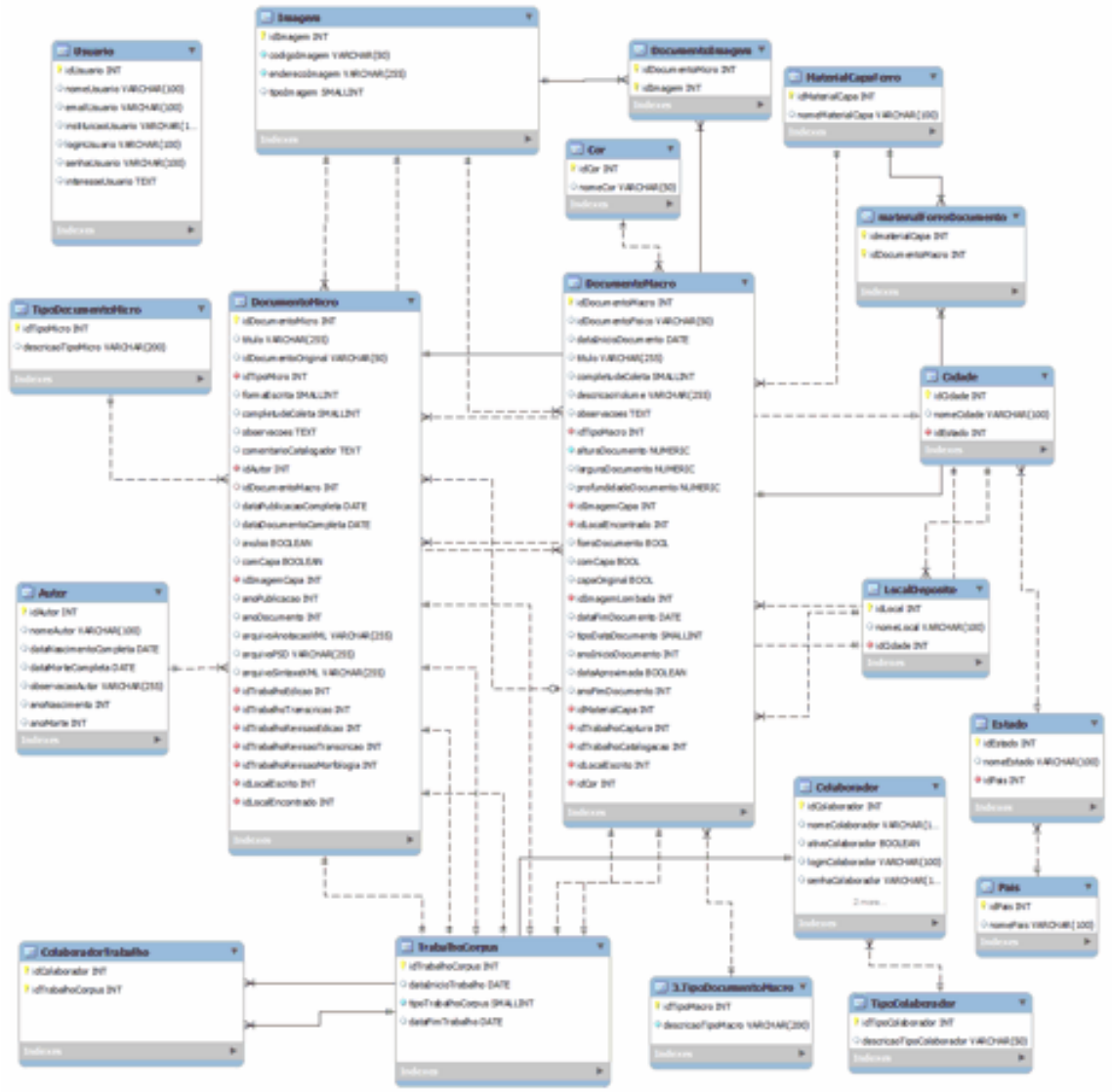
A figura 30 mostra o modelo lógico de banco de dados elaborado para o software. Cada retângulo corresponde a uma tabela no banco de dados, com os respectivos dados que serão armazenados. As linhas entre os retângulos representam relacionamentos entre as tabelas.

O diagrama da figura 31 apresenta os nove pacotes de classes de software desenvolvidos, cada um agrupando um conjunto de classes relacionadas.

Segue a descrição de cada pacote representado pelo diagrama:

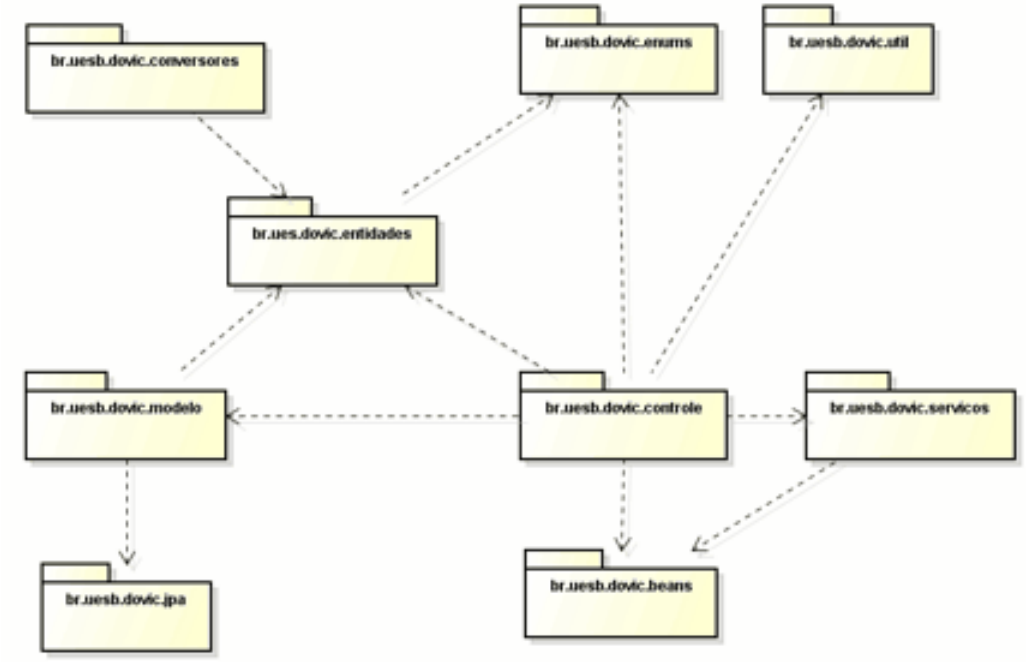
- *br.uesb.dovic.entidades* - agrupa as classes relacionadas às entidades que serão armazenadas em banco de dados, como Autor, Documento, Cidade, Estado,País, Imagem, Tipo de documento, Usuário, etc.
- *br.uesb.dovic.beans*- agrupa as classes relacionadas às entidades que fazem parte do domínio do sistema mas que não serão armazenadas em banco de dados.
- *br.uesb.dovic.modelo* – reune as classes responsáveis pelo serviço de persistência das classes do pacote *br.uesb.dovic.entidade*.
- *br.uesb.dovic.jpa* – utilizado para conter as classes de persistência de dados para o *framework* utilizado no projeto, chamado *Hibernate*.
- *br.uesb.dovic.enums* – contém as classes que implementam tipos enumerados. Foram implementadas nesse pacote classes para as funções sintáticas e morfossintáticas, tipos de imagem (capa, contracapa, lombada, frente, verso), tipos de trabalho com o *corpus* (catalogação, edição, trasncrição, revisão, etc.), dentre outras.

Figura 30 - Modelo lógico de banco de dados do WebSinC



- *br.uesb.dovic.controle* – apresentam as classes responsáveis pela comunicação entre a interface gráfica e as outras classes, coordenando a execução das funcionalidades do sistema.
- *br.uesb.dovic.util* – reúne as classes utilitárias desenvolvidas para o sistema, como mensagens de erro.
- *br.uesb.dovic.servicos* – reúne as classes responsáveis pelo serviço de buscas em XML implementadas com XQuery, tratamento de arquivos, e fornecimento das etiquetas para listagem na interface gráfica.
- *br.uesb.dovic.conversores* – reúne as classes responsáveis pela conversão dos tipos de dados das classes em um formato legível para a interface gráfica.

Figura 31 - Diagrama de pacotes do software WebSinc



As setas apresentadas pelo diagrama da figura 31 modelam a dependência entre os diversos pacotes do software, ou seja, quais recursos de um pacote são utilizados por outro.

7.3 Implementação do software

O software implementado permite o gerenciamento dos documentos do *corpus*, desde o cadastro das informações até o *upload* de arquivos e imagens, a disponibilização dos documentos na Internet para o público em geral e a realização de buscas automáticas.

A figura 32 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*.

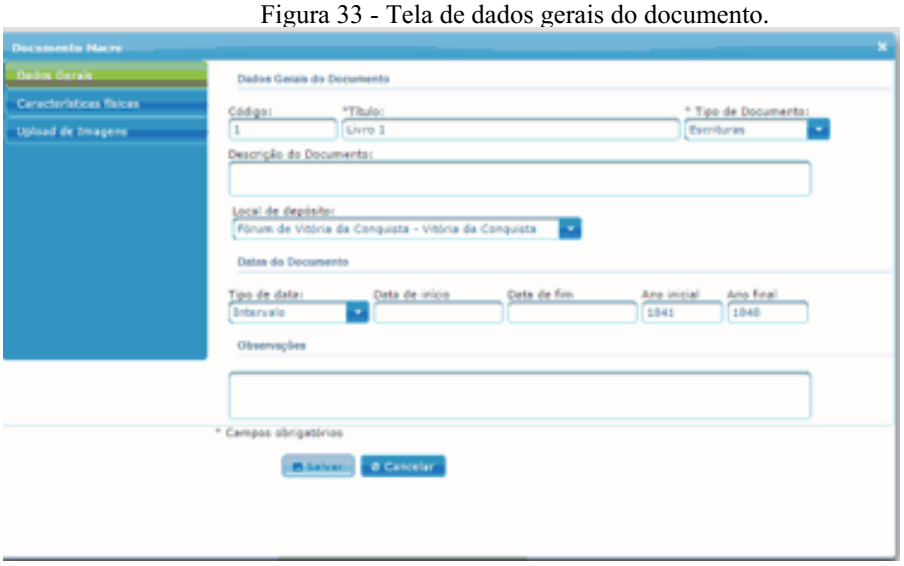
Figura 32 - Tela de login do WebSinC.

7.3.1 Funcionalidades de cadastro e *upload* de arquivos e imagens

A inserção de registros de documentos do *corpus* pode ser feita preenchendo os dados solicitados na tela de cadastro. Tais informações são distribuídas em três abas na tela, exibidas nas figuras 33 a 35. Na aba "Dados Gerais", dados como código, título, tipo de documento, local de depósito, entre outros são solicitados. Os documentos são divididos nos tipos que denominamos "macro" e "micro". O documento macro é uma coleção de vários outros documentos, podendo ser um livro composto de cartas, por exemplo. Já o documento micro é cada documento individual, que pode ou não estar inserido num documento macro. Quando o documento micro não faz parte de um documento macro, o consideramos como um documento avulso. A inserção dos registros pode ser tanto para os livros (considerados documentos macro não avulsos) quanto para os documentos avulsos (documentos micro).

É importante ressaltar que apenas um usuário previamente cadastrado com perfil de administrador, fazendo uso de login e senha, terá acesso a esta funcionalidade de cadastro e *upload* de arquivos e imagens.

Figura 33 - Tela de dados gerais do documento.

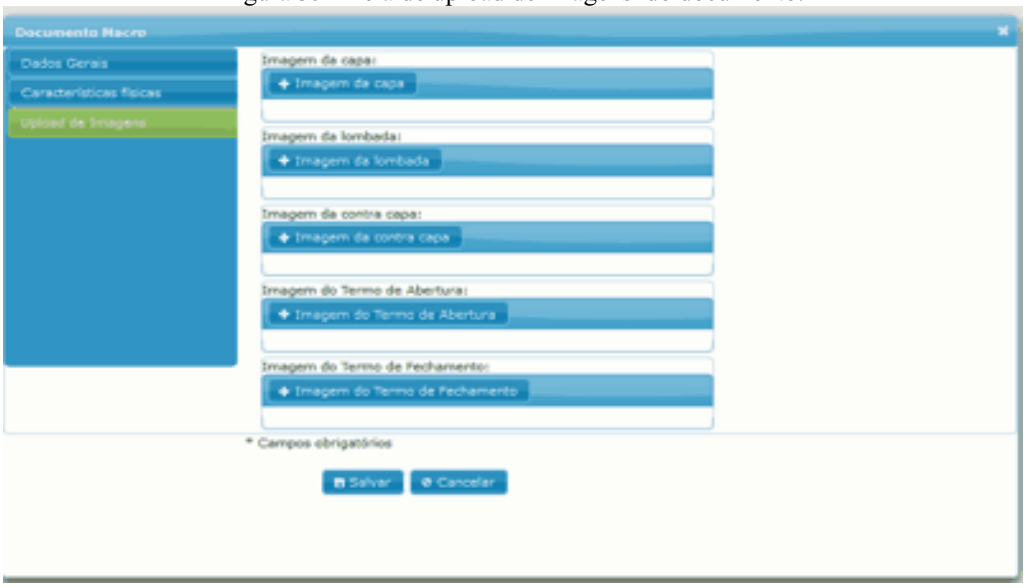


Na aba "Características físicas", dados como altura, largura, profundidade e informações sobre a capa e o forro podem ser fornecidas. Estes dados não são de preenchimento obrigatório.

Na aba "*upload* de imagens", o administrador pode fazer a seleção de arquivos de imagens em seu computador e estes serão copiados e armazenados no servidor. As imagens que poderão ser armazenadas e posteriormente recuperadas e exibidas são: imagem da capa, lombada, contra capa, termo de abertura e termo de fechamento. Todos os arquivos são de preenchimento facultativo, uma vez que nem todos os documentos do *corpus* possuem estas cinco imagens disponíveis.

Figura 34 - Tela de características físicas do documento.

Figura 35 - Tela de upload de imagens do documento.



O software WebSinc também provê funcionalidades de manutenção de todos os cadastros listados nos requisitos funcionais (seção 6.1). A figura 36 mostra a tela para manutenção do cadastro de autores de documentos. Todas as outras telas seguem a mesma interface gráfica, com a exibição dos registros já cadastrados numa tabela, com colunas incluindo atalhos para ações de edição e exclusão do registro.

Figura 36 - Tela de manutenção de cadastro de material de capa e forro.

Os documentos micro que compõem os documentos macro são inseridos mantendo uma relação de dependência com os mesmos. Apenas após o cadastro do documento macro, os documentos componentes poderão ser inseridos. A figura 37 mostra a tela de documentos macro do *corpus* DOViC cadastrados em forma de tabela, com colunas contendo atalhos para as ações de registro do trabalho com o *corpus* (informações acerca da catalogação, captura, transcrição e edição), de registro dos manuscritos (documentos micro) ligados a cada documento macro, ações de edição e de remoção.

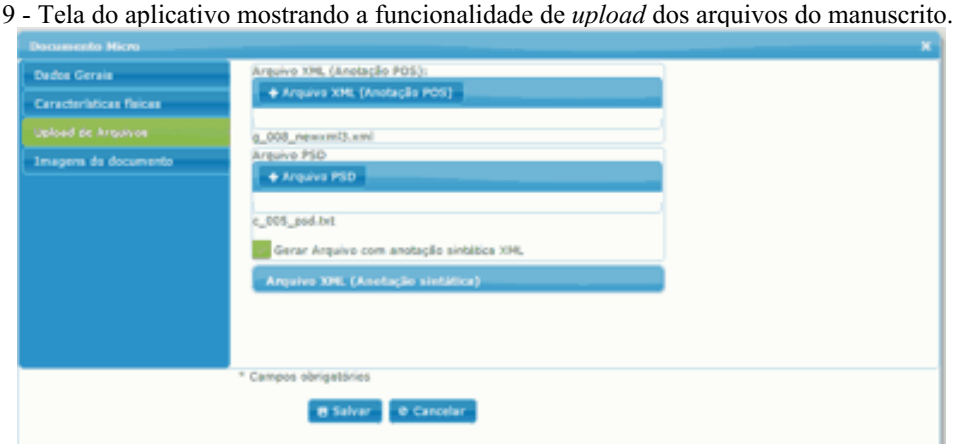
Figura 37 - Tela exibindo a listagem de documentos já cadastrados.

Através do atalho para inserção de manuscritos do documento macro, uma tela com os manuscritos já inseridos e relacionados a ele é exibida em forma de tabela, contendo colunas com atalhos para as ações de edição e remoção. Um novo manuscrito pode ser inserindo clicando-se no botão "Novo". A figura 38 mostra uma tela de exemplo com manuscritos cadastrados para o livro 1.

Figura 38 - Tela exibindo manuscritos cadastrados para o livro 1

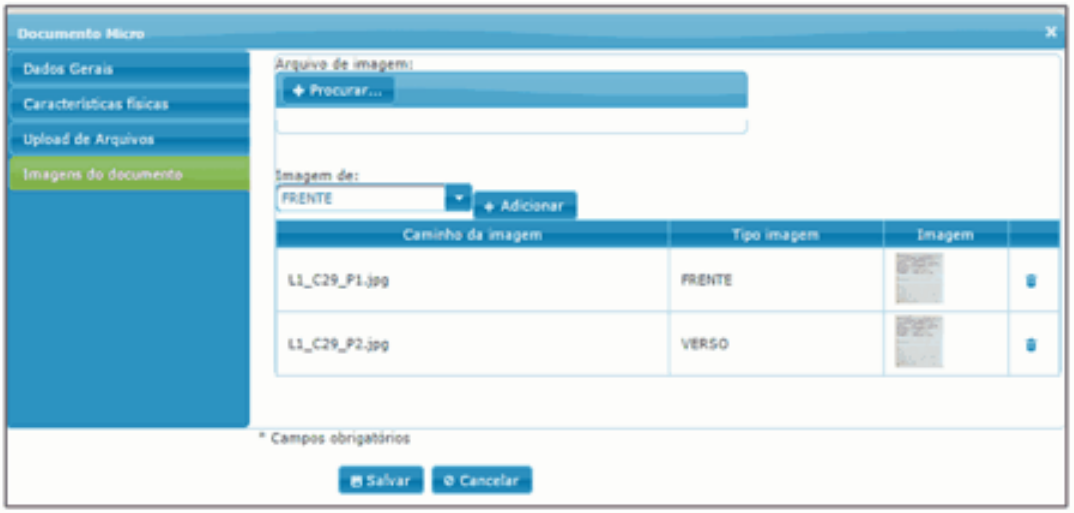
A inserção ou edição do manuscrito (documento micro) exibe uma tela para registro das informações do documento. Os dados referentes ao manuscrito são distribuídos em quatro abas: "Dados Gerais", "Características Físicas", "*Upload* de arquivos" e "Imagens do documento". Na aba "*Upload* de arquivos" é possível ao administrador selecionar em seu computador o arquivo XML gerado pelo E-Dictor para este documento e o arquivo no formato *Penn TreeBank*, chamado arquivo PSD. Tais arquivos serão copiados e armazenados no servidor e serão usados na realização das buscas automáticas. A figura 39 exibe a tela com a funcionalidade de *upload* dos arquivos.

Figura 39 - Tela do aplicativo mostrando a funcionalidade de *upload* dos arquivos do manuscrito.



Na aba "Imagens do documento", o administrador irá selecionar os arquivos de imagens com a fotografia do documento. É possível inserir mais de uma imagem para cada manuscrito, já que um documento pode ser composto de várias páginas físicas. A indicação de ser uma imagem de "Frente" ou "Verso" também deve ser realizada. Conforme as imagens vão sendo inseridas, são instantaneamente visualizadas em formato de tabela na mesma tela. A figura 40 mostra a tela com esta funcionalidade.

Figura 40 - Tela exibindo a funcionalidade de inserção de imagens do manuscrito.



.

7.3.2 Funcionalidade de disponibilização dos documentos do *corpus* para o público

Os documentos do *corpus* DOViC serão disponibilizados na Internet pelo aplicativo WebSinC de maneira controlada. Para ter acesso ao *corpus*, um usuário precisa realizar um cadastro e marcar a opção de leitura e aceite do termo de compromisso.

Por se tratar de um *corpus* de documentos históricos manuscritos, as fotografias dos documentos originais do DOViC também são disponibilizadas através da ferramenta. As

principais características dos manuscritos, como o texto transcrito, as imagens dos originais e as informações gerais (título, local de depósito, data, gênero, informações sobre edição, etc.) são armazenados no banco de dados projetado para o sistema.

7.3.3 Implementação do recurso de buscas automáticas no software

O software possui um mecanismo de buscas automáticas baseadas em categorias sintáticas ou morfossintáticas. O arquivo de entrada para realização das buscas morfossintáticas é o arquivo XML gerado pelo E-Dictor. Para pesquisas desse tipo, o algoritmo implementado sempre irá buscar pelos valores dos atributos "m" contidos nas etiquetas <w> do arquivo fonte. Para as pesquisas baseadas nas categorias sintáticas, o software converte um arquivo no formato *Penn TreeBank* para uma estrutura XML correspondente e assim utiliza as mesmas tecnologias para XML nos dois tipos de busca. O formato PTB é usado apenas como fonte de entrada para essa conversão, o que dispensa completamente o uso da ferramenta de busca *Corpus Search*.

A decisão em usar XML para realização de buscas sintáticas se deu baseada principalmente por razões de independência tecnológica e reutilização. Um esquema de anotação sintática utilizando XML traz a vantagem de utilizar um padrão aberto para interoperabilidade e intercâmbio de dados. Utilizando formatos específicos para o esquema de anotação sintática, as tecnologias para recuperação da informação dificilmente são reutilizadas. Para cada tipo de anotação, são necessárias ferramentas de busca restritas àquela anotação em questão. Como o aplicativo provê uma interface gráfica para que o usuário realize as buscas, a tecnologia utilizada para ele torna-se transparente, uma vez que não terá conhecimento do que está realmente sendo utilizado na busca.

Com o uso de XML em *corpora* digitais, as buscas sintáticas tornam-se independentes de tecnologia específica, passando a utilizar tecnologias padrão. Em se tratando do *corpus* DOViC, a vantagem é a reaplicação da mesma tecnologia que será utilizada para as pesquisas morfossintáticas. Como a anotação morfossintática e de edições dos textos do *corpus* já é feita em XML, as buscas nestes arquivos terão que ser feitas obrigatoriamente utilizando tecnologias para XML. Assim, a mesma tecnologia pode então ser reutilizada, dispensando o uso do software *Corpus Search* ou similar.

7.3.3.1 Implementação do recurso de buscas sintáticas

As buscas automáticas por categorias sintáticas podem ser realizadas no WebSinC através de uma interface gráfica. Não é necessário que o usuário possua conhecimento de determinada linguagem de programação, linguagem de consulta ou sintaxe de comandos. As buscas podem ser feitas pela manipulação dos componentes gráficos na interface. Não é necessário também que o usuário conheça o conjunto de etiquetas da anotação do *corpus* pesquisado.

A figura 41 mostra a tela do WebSinC para o recurso de buscas sintáticas. A consulta para a busca é construída graficamente 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*. A figura 42 mostra a tela exibindo a lista de itens/sintagmas. O usuário pode selecionar um item 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. A figura 43 também mostra o uso desse recurso. Após a seleção do item, o mesmo é inserido automaticamente no bloco. A figura 43 ilustra um bloco contendo algumas categorias de classes e sintagmas selecionados.

Figura 41 - Tela do WebSinC para o recurso de buscas sintáticas.

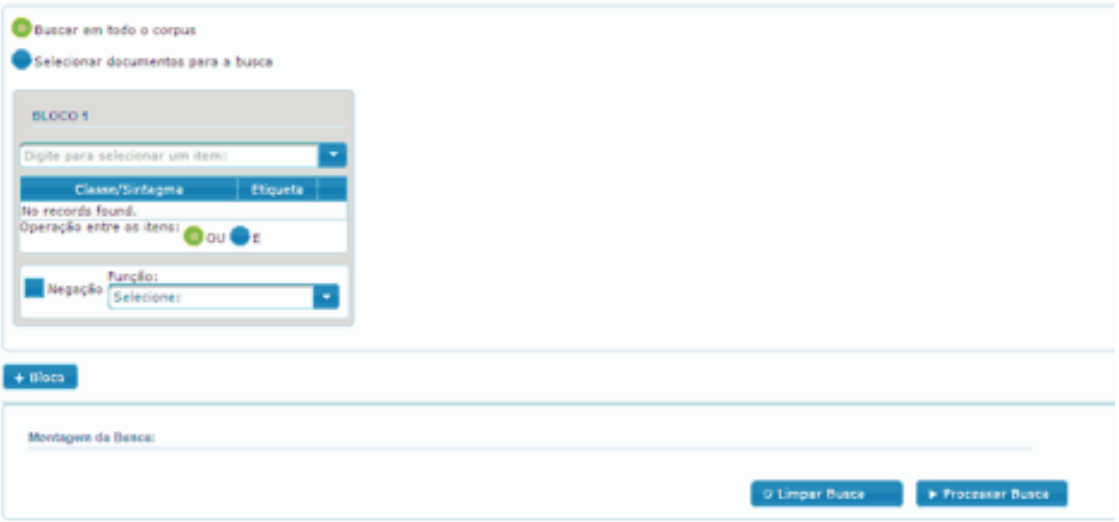


Figura 42 - Lista para seleção de sintagmas no WebSinC.

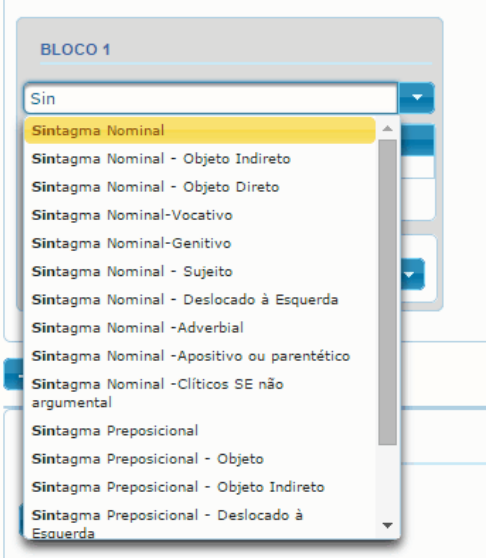
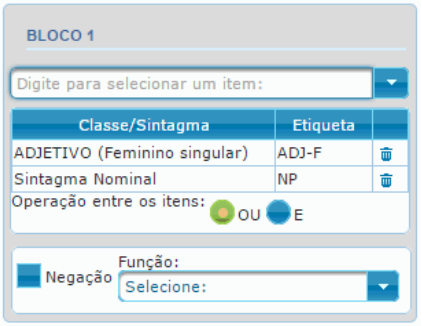
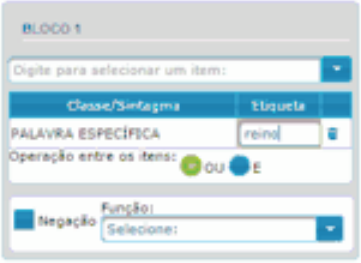


Figura 43 - Bloco com itens selecionados.



Além de classes lexicais e/ou sintagmas, o usuário pode selecionar para a busca uma determinanda palavra, que correponde a um nó folha na árvore sintática. Na interface do WebSinC, ao selecionar uma palavra específica, a etiqueta não é visualizada, mas um campo de texto é exibido, para que o usuário digite a palavra desejada. A figura 44 mostra esse recurso, dando como exemplo a seleção da palavra "reino".

Figura 44 - Seleção de palavra específica para busca.



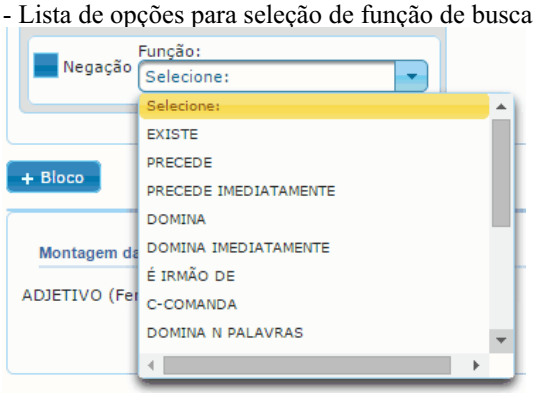
Após a montagem do bloco, o usuário deve selecionar uma função de busca. Há funções que requerem apenas um argumento, como é o caso da função "Existência". As outras funções requerem dois ou mais argumentos, e ao selecioná-las, os componentes da interface são exibidos dinamicamente para seleção deles. As funções de busca sintática implementadas no WebSinC são:

1. Existência - Função que requer apenas um argumento. Retorna as sentenças em que os argumentos selecionados existem em qualquer posição.
2. Precedência - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento precedendo o segundo na árvore sintática.
3. Precedência imediata - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento precedendo imediatamente o segundo na árvore sintática.
4. Dominância - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando o segundo na árvore sintática, mesmo que não seja imediatamente.
5. Dominância imediata - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando imediatamente o segundo na árvore sintática.
6. Irmandade - Função que requer dois argumentos. Retorna as sentenças onde o primeiro e o segundo argumentos possuem o mesmo pai na árvore sintática.
7. C-Comando - Função que requer dois argumentos. Retorna as sentenças onde o primeiro argumento c-comanda o segundo na árvore sintática.
8. Dominância de N palavras - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando N palavras na árvore sintática.
9. Dominância de mais de N palavras - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando mais de N palavras na árvore sintática.

10. Dominância de menos de N palavras - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando menos de N palavras na árvore sintática.
11. Dominância imediata como primeiro filho - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando imediatamente o segundo como primeiro filho (filho mais à esquerda) na árvore sintática.
12. Dominância imediata como último filho - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando imediatamente o segundo como último filho (filho mais à direita) na árvore sintática.
13. Dominância imediata como N-ésimo filho - Função que requer três argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando imediatamente o segundo como filho na posição N (terceiro argumento) na árvore sintática.
14. Dominância imediata como único filho - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando imediatamente o segundo como único filho na árvore sintática.
15. Dominância imediata de N filhos - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando imediatamente N (segundo argumento) filhos na árvore sintática.
16. Dominância imediata de menos de N filhos - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando imediatamente menos de N (segundo argumento) filhos na árvore sintática.
17. Dominância imediata de mais de N filhos - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento dominando imediatamente mais de N (segundo argumento) filhos na árvore sintática.

A figura 45 mostra o componente gráfico na tela do WebSinC exibindo a lista de opções com as funções de busca para seleção. Ao lado da lista há uma opção que pode ser marcada para negação da função. Assim, se o usuário por exemplo, desejar buscar por sentenças com classes ou sintagmas que NÃO dominam N palavras, além de selecionar a opção "DOMINA N PALAVRAS" na lista, deve-se também marcar a opção "Negação".

Figura 45 - Lista de opções para seleção de função de busca sintática.



Se uma função de busca que requer dois ou mais argumentos for selecionada, os componentes gráficos correspondentes serão gerados dinamicamente na tela. A figura 46 mostra a tela gerada após a seleção da função "DOMINA IMEDIATAMENTE COMO N-ÉSIMO FILHO", que requer três argumentos. Um segundo bloco foi gerado à direita do primeiro para seleção dos itens ou sintagmas dominados pelos primeiros. Abaixo da lista de funções, um campo de texto foi gerado para receber o argumento N, correspondente ao número de palavras. É possível observar ainda na figura que a construção da descrição da consulta também é gerada dinamicamente a cada escolha do usuário. Na descrição da consulta, a função é exibida com destaque na cor vermelha e operações lógicas recebem destaque na cor azul. A figura 47 mostra a tela com lista de sentenças retornadas como resultado desta consulta aplicada ao texto de Pero Magalhães de Gandavo, do corpus Tycho Brahe.

Na figura 46, pode ser observada a existência de uma caixa de seleção de função também para o bloco 2. Assim, se uma função que seleciona mais argumentos for escolhida, um terceiro bloco será gerado dinamicamente na tela. Os itens do bloco 2 serão os últimos argumentos da função escolhida para o bloco 1 e serão os primeiros argumentos da função escolhida no bloco 2, encadeados por uma operação lógica "E" implícita. A geração dinâmica desses blocos no WebSinC limita-se ao número de quatro. Portanto, até quatro blocos podem ser gerados dinamicamente e a figura 48 mostra um exemplo. A interpretação para esta consulta é: Um *sintagma nominal sujeito* **precede** *verbo "estar" no presente,* **E** *verbo "estar" no presente* **precede imediatamente** um *verbo no gerúndio,* **E** o *verbo no gerúndio* por sua vez, **precede** um *nome próprio no singular*.

Figura 46 - Tela com montagem dinâmica de blocos.

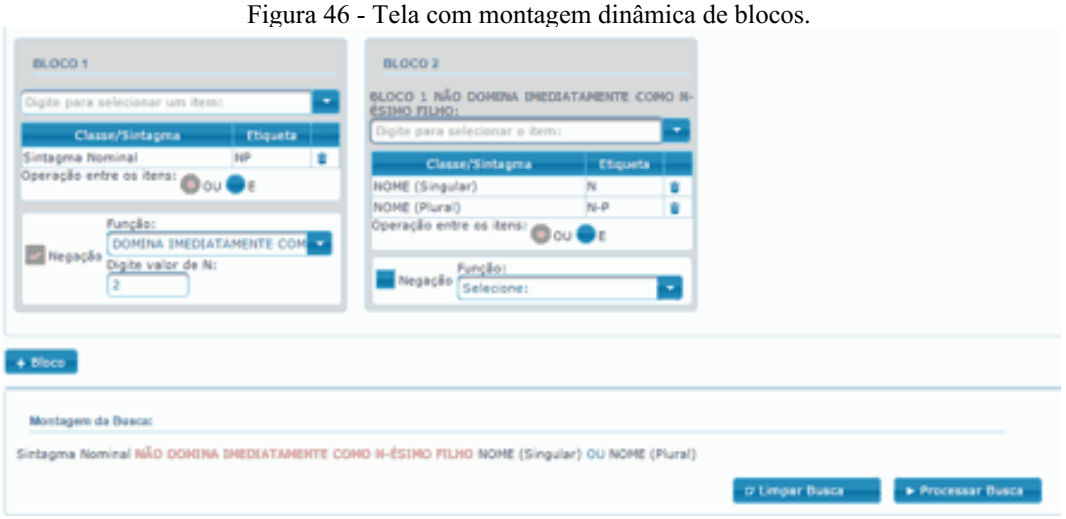


Figura 47 - Tela com sentenças de resultado da busca.

Id	Sentença
1	E assim se punham de joelhos
2	e dura até de madrugada :
3	e fica a terra toda coberta de névoa , por respeito de ter muitos arvoredos que chamam a si todos estes humores .
4	Outro muito grande cinqüenta léguas d@ @este para Oriente sai também a@ @o Norte , a que chamam rio d@ @o Maranhão .
5	D@ @o outro não descobriram coisa alguma ,
6	e assim se não sabe até agora de onde procedem ambos .
7	dentro é muito fundo e limpo ,
8	e vem sair d@ @aí uma légua :
9	e quando há cheias arrebenta por cima
10	e arrasa toda a terra .
11	Duzentas e setenta léguas por ele acima , está edificada uma cidade povoada de Castelhanos , que se chama Ascenção .
12	Até aqui se navega por ele , e ainda d@ @aí por diante muitas léguas .
13	Junto d@ @elas havia muitos Índios , quando os Portugueses começaram de as povoar :
14	e mataram muitos d@ @eles :

Figura 48 - Blocos gerados dinamicamente para busca sintática.

Blocos com itens sendo argumentos independentes também podem ser gerados, limitando-se ao número de um. Ao clicar no botão "+Bloco", um novo bloco será gerado, mas

seus itens não são argumentos do bloco anterior. Os parâmetros escolhidos neste novo bloco podem gerar mais um outro, o que leva a um total de até seis blocos na montagem gráfica das buscas do WebSinC. A figura 49 mostra essa configuração e a interpretação para busca da figura é: Um *sintagma nominal sujeito* **precede** *verbo "estar" no presente* **E** *verbo estar no presente* **precede imediatamente** um *verbo no gerúndio* **E** o *verbo no gerúndio* **precede** um *nome próprio no singular* **E** um *NP sujeito* **domina** um *nome no singular*. Cabe ressaltar que os blocos cinco e seis são independentes em relação aos blocos de um a quatro, mas estabelece uma relação entre si.

Figura 49 - Busca sintática montada com seis blocos.

Depois que o usuário montar a busca graficamente, deverá clicar no botão "Processar Busca". Os resultados da busca serão exibidos em outra janela.

O arquivo pesquisado na busca sintática do WebSinC é um arquivo XML com representação da estrutura sintagmática, proveniente da conversão do arquivo anotado no formato *Penn TreeBank* em XML. Para essa transformação, foi implementado no software um algoritmo que recebe como entrada um arquivo no primeiro formato e gera um arquivo de saída XML correspondente. O programa não implementa a função de *parser*, e portanto, o arquivo de entrada deve ser um documento *Penn TreeBank* bem formado (NAMIUTI; COSTA, 2014) .

Para o arquivo de saída, foram usados os mesmos nomes de rótulos para nomear as *tags*, com exceção do nó raiz e de rótulos com caracteres não aceitos pela linguagem XML. Como o arquivo de entrada não possui um elemento raiz, foi inserido no arquivo de saída a *tag* <DOCUMENT> como raiz do documento. Nós com o sinal de pontuação "." no formato *Penn TreeBank* foram mapeados para tag*s* <POINT>. Nós com o símbolo "," foram mapeados para

tag*s* <COMMA>. Houve a necessidade de substituir o caracter “$” pelo caracter “S”. Assim, rótulos como “PRO$” foram mapeados para etiquetas “PROS”. Os demais nomes das tag*s* para o documento XML permaneceram os mesmos utilizados no formato *Penn TreeBank*. Assim, cada nó do arquivo de entrada é mapeado numa *tag* XML com mesmo nome.

Cada nó folha (nó sem filho), que corresponde a cada palavra da sentença, é gerado na saída como conteúdo textual de um elemento <LEAF>. Os elementos <LEAF> receberam um atributo "w", cujo valor pode ser "yes" ou "no" para indicar se tal conteúdo textual corresponde a uma palavra ou não. Isto foi necessário para que fosse possível separar o conteúdo textual que deve estar visível nos resultados do que não deve estar. Categorias como sujeito vazio são representadas no formato *Penn TreeBank* pela *string* *pro* e corresponde a um nó folha. Vestígios de movimentos na árvore também são representados como nó folha, geralmente da forma *T*-n, onde n é um número natural. Não é desejável que estas *strings* sejam visualizadas nas sentenças de resultado. Outro motivo para adoção do atributo foi a busca por domínio de palavras. Nesse tipo de busca, tais *strings* não devem ser contadas como palavras, mesmo que sejam nós do tipo folha, mapeados portanto para elementos <LEAF>. Outro atributo do elemento <LEAF> é "v", cujo conteúdo corresponde ao valor textual das folhas. A redundância desta informação foi inserida por questões práticas, com objetivo de facilitar os algoritmos de buscas.

Todos os elementos XML gerados pelo algoritmo também possum um atributo "id", cujo valor recebe um número inteiro sequencial. Assim, o elemento raiz <DOCUMENTO> possui id="1", e os outros ementos subsequentes recebem um número na sequência. Tal atributo é utilizado nas buscas em XQuery.

O quadro 18 mostra um trecho de um arquivo do *corpus* Tycho Brahe com anotação sintática *Penn TreeBank* e a figura 50 mostra a visualização do arquivo de saída em XML gerado pelo WebSinc. O arquivo na figura está sendo exibido pelo navegador Firefox para uma melhor visualização, já que navegadores fornecem uma melhor exibição das relações hierárquicas existentes num documento XML.

Quadro 18- Trecho de arquivo do corpus Tycho Brahe com anotação Penn TreeBank.

```
( (CODE <P_6>))

( (IP-MAT (IP-GER (VB-G REINANDO)
                  (NP-SBJ (D aquele)
                          (ADJP (Q muito)
                                (ADJ católico)
                                (CONJP (CONJ e)
                                       (ADJX (ADJ-S sereníssimo))))
                          (NPR Príncipe)
                          (NP-PRN (NPR el-Rei) (NPR Dom) (NPR MANUEL))))
```

Para a realização das buscas o WebSinC utilizou-se de expressões na linguagem XQuery. Para cada função sintática ou morfossintática, uma expressão XQuery foi utilizada. Os argumentos das funções são alterados dinamicamente no programa fazendo a junção da expressão XQuery com a linguagem de programação utilizada. As funções XQuery utilizadas para cada função de busca implementada estão no apêncice A.

Figura 50- Visualização da estrutura hierárquica de anotação com XML no navegador Firefox

7.3.3.2 Implementação do recurso de buscas morfossintáticas

As buscas automáticas por categorias morfossintáticas no WebSinC também podem ser 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. O limite do número de blocos para este tipo de busca também é seis. A figura 51 mostra um exemplo da tela de busca morfossintática com um exemplo montado.

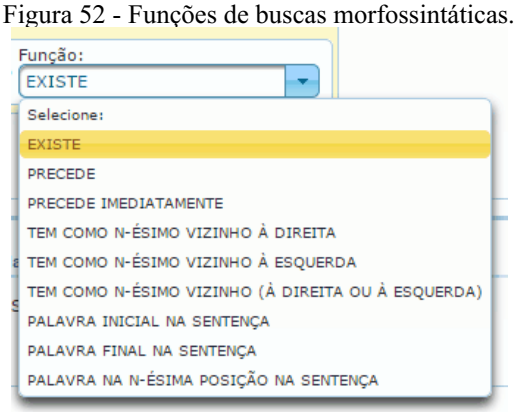
Figura 51 - Blocos gerados para busca morfossintática.

Após a montagem do bloco, o usuário deve selecionar uma função de busca. Assim como nas buscas sintáticas, 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 também são exibidos dinamicamente para seleção dos argumentos. As funções de busca morfossintática implementadas no WebSinC são descritas a seguir e a figura 52 mostra a exibição da lista de funções na interface gráfica.

1. Existência - Função que requer apenas um argumento. Retorna as sentenças em que os argumentos selecionados existem em qualquer posição.
2. Precedência - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento precedendo o segundo na sequência linear da sentença.

3. Precedência imediata - Função que requer dois argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento precedendo imediatamente o segundo na sequência linear da sentença.
4. Vizinhança à direita - Função que requer três argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento tendo o segundo argumento como N-ésimo vizinho à direita dele.
5. Vizinhança à esquerda - Função que requer três argumentos. Retorna as sentenças onde há ocorrência do primeiro argumento tendo o segundo argumento como N-ésimo vizinho à esquerda dele.
6. Início da sentença - Função que requer apenas um argumento. Retorna as sentenças em que os argumentos selecionados ocorrem na posição inicial delas.
7. Fim da sentança - Função que requer apenas um argumento. Retorna as sentenças em que os argumentos selecionados ocorrem na posição final delas.
8. Posição N da sentança - Função que requer dois argumentos. Retorna as sentenças em que o argumento selecionado ocorre na posição N delas.
9.

Figura 52 - Funções de buscas morfossintáticas.



Se uma função de busca que requer dois ou mais argumentos for selecionada, os componentes gráficos correspondentes serão gerados dinamicamente na tela. A figura 53 mostra a tela gerada após a seleção da função "PRECEDE IMEDIATAMENTE ", que requer dois argumentos. O exemplo da figura busca por sentenças onde há clíticos (CL ou SE) que precedem imediatamente um verbo flexionado (VB-*). Aplicamos esta busca no arquivo correspondente à carta de Alforria da cabra de nome Sofia, do *corpus* DOViC, e o resultado da busca não retornou nenhuma sentença, conforme mostra a figura 54.

Figura 53 - Exemplo de busca morfossintática.

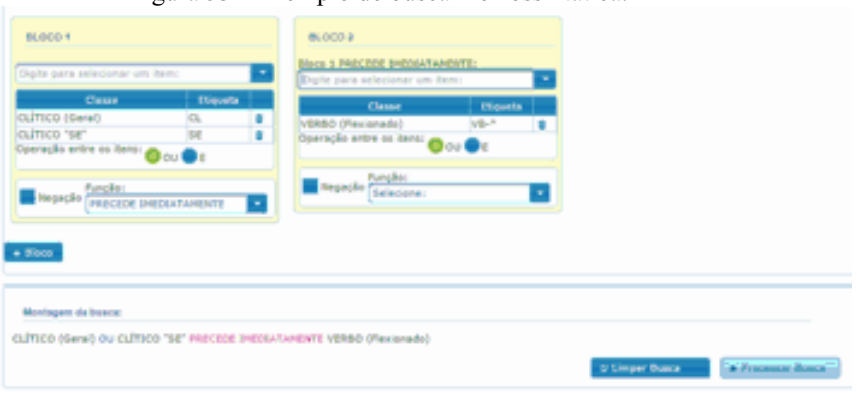


Figura 54 - Resultado de busca morfossintática.

O arquivo pesquisado na busca morfossintática do WebSinC é o arquivo XML gerado pelo E-Dictor, com anotações da estrutura dos textos, informações morfossintáticas e anotações de edições. As buscas são feitas pelos elementos <m> com o atributo "v" no arquivo XML, que indicam a anotação da categoria POS.

8 RESULTADOS

Neste capítulo tratamos dos resultados produzidos pela ferramenta WebSinC, desde a disponibilização dos textos do corpus DOViC na Internet aos resultados de buscas morfossintáticas e sintáticas. Apresentamos os testes realizados para as funções de buscas com seus respectivos resultados, metodologia utilizada e dicussões.

8.1 Disponibilização do *Corpus* DOViC na Internet

A disponibilização dos textos do *corpus* DOViC é um dos objetivos do trabalho desenvolvido. Após o cadastro de informações e *upload* dos arquivos do *corpus* no banco de dados pelo administrador através do WebSinC, os textos podem ser disponibilizados ao público. A figura 55 mostra a tela do aplicativo exibindo o catálogo visual de um livro de escrituras que compõe o *corpus* DOViC.

8.2 Recuperação de diferentes versões do texto em XML

Para visualizar o texto em versões diferentes, o aplicativo exibe as abas "Texto modernizado" e "Texto não modernizado". O léxico de edições também pode ser visualizado na última aba.

Tanto o texto transcrito quanto as versões com ou sem modernização de grafia são gerados a partir de um único arquivo fonte, o arquivo XML gerado pela ferramenta E-Dictor. O aplicativo Web recupera a informação sobre a fonte no banco de dados e a partir de um algoritmo com a linguagem XQuery faz as transformações necessárias.

Figura 55 - Tela exibindo dados de livro de escrituras do *corpus* DOViC.

A figura 56 mostra uma tela com o texto original da carta de alforria do *corpus*. Informações como título, gênero, data do texto original, localização e dados da captura e edição das imagens do texto original, previamente cadastrados, podem ser exibidos pelo WebSinC. As imagens editadas dos manuscritos originais são exibidas juntamente com o texto transcrito. A figura 57 mostra o texto na versão modernizada e na figura 58 é exibido o léxico de edições.

Figura 56 - Tela do aplicativo exibindo informações e texto transcrito de carta de alforria do *corpus* DOViC.

Figura 57 - Tela do aplicativo exibindo texto modernizado.

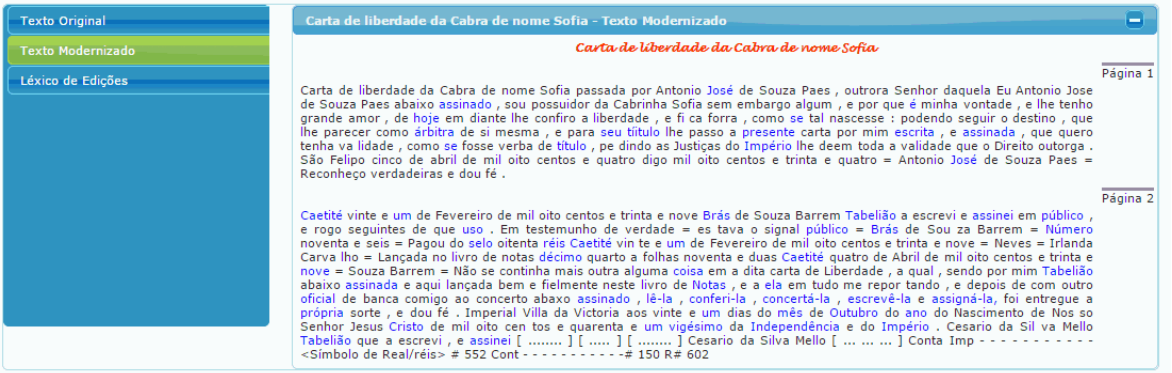
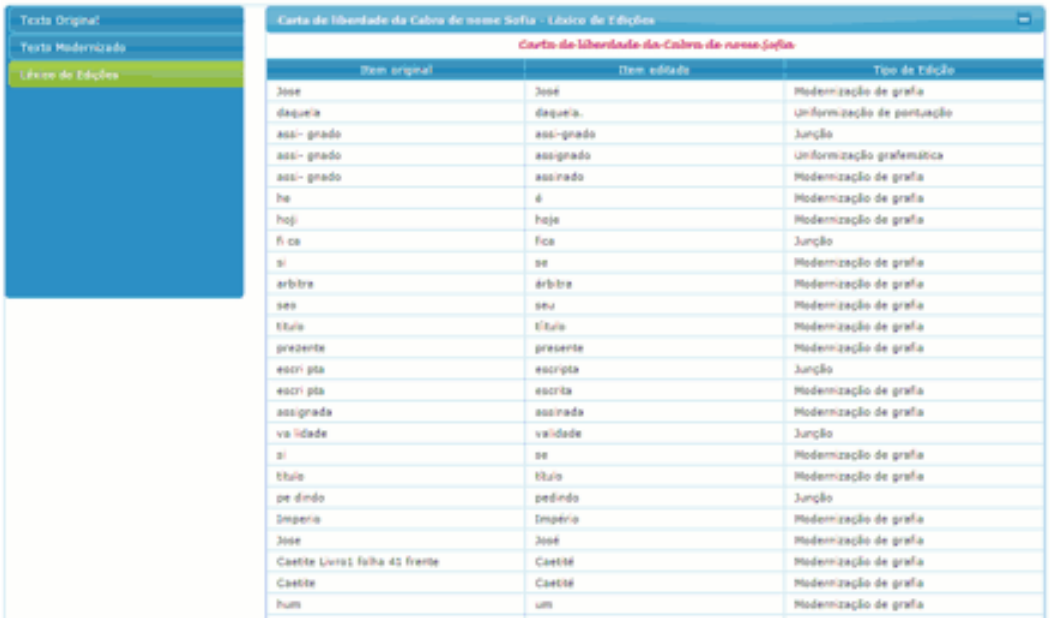


Figura 58 - Tela do aplicativo exibindo léxico de edições.



8.3 Resultados de buscas sintáticas

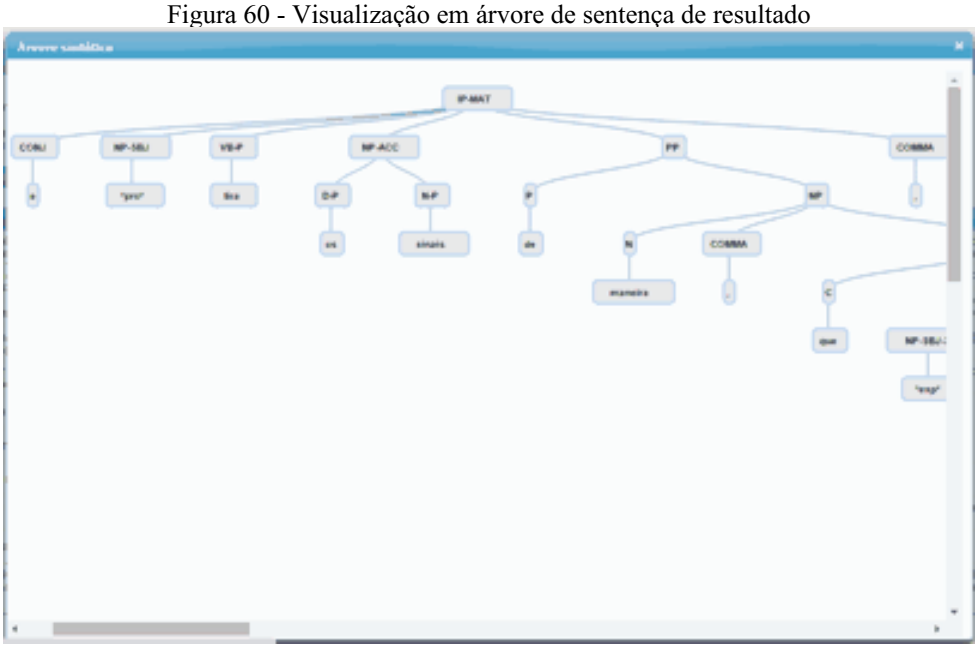
Para realização das buscas sintáticas no WebSinc, é necessária a transformação do arquivo PSD gerado pelo *parser* em um arquivo XML com representação da estrutura sintagmática. Tal conversão é realizada pelo WebSinC e foi abordada na seção 6.3.3.1. Sendo assim, o arquivo XML para buscas sintáticas se trata de um outro arquivo diferente e não relacionado ao arquivo com anotações morfossintáticas e de edições do texto gerado pelo E-Dictor.

As buscas no WebSinc são montadas graficamente conforme mostrado na seção anterior. 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.

A figura 59 mostra como exemplo a tela de resultado para uma busca sintática realizada com o documento "História da Província de Santa Cruz", do *corpus* Tycho Brahe. A busca procurou por sentenças onde um *NP sujeito* e um *sintagma adverbial* tem como irmão na árvore um *verbo estar no tempo passado*. A figura 60 mostra a representação gráfica arbórea de uma das sentenças do resultado da busca.

Figura 59 - Tela de resultado de busca sintática

Figura 60 - Visualização em árvore de sentença de resultado



8.3.1 Aplicação das buscas sintáticas em pesquisas linguísticas

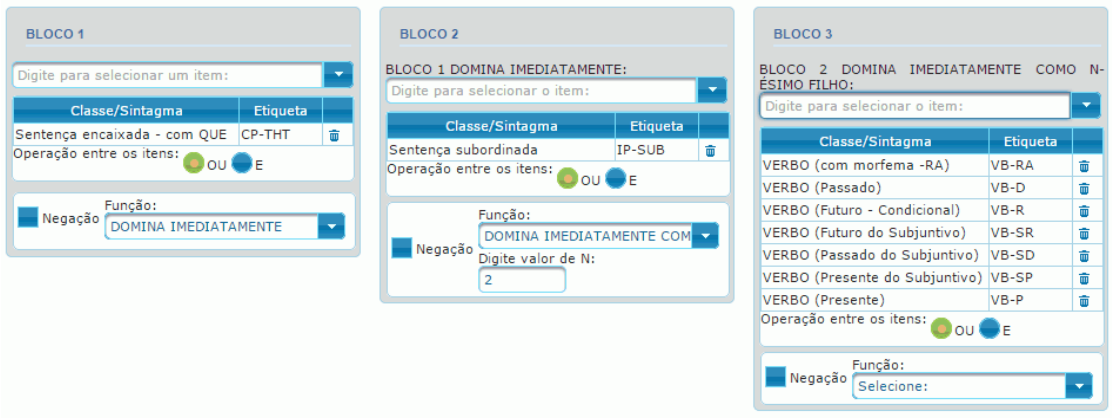
Defendemos que a ferramenta WebSinC pode contribuir com os estudos gramaticais da língua portuguesa, mais especificamente com o avanço das pesquisas em sintaxe. Nossa hipótese baseia-se no fato de que muitas pesquisas necessitam realizar buscas automáticas em textos de *corpora* para obtenção de grande volume de dados. Existem diversas outras ferramentas disponíveis, mas a maioria requer o aprendizado de uma linguagem de comando, como o *Corpus Search*, ou de uma linguagem de programação, como a linguagem Perl. Especificamente para o *corpus* DOViC e outros que utilizam a mesma tecnologia de anotação do *Corpus* Tycho Brahe, as buscas nos textos anotados morfossintaticamente em XML com uso do E-Dictor requerem o aprendizado de linguagem de consulta para XML, como XQuery. Para o linguista não familiarizado com tais tecnologias, o aprendizado das mesmas pode protelar ou restringir a obtenção dos dados. A ferramenta WebSinC pode contribuir com tais pesquisas uma vez que permite realizar buscas em *corpora* através de uma interface gráfica que reduz a necessidade de domínio de linguagens de programação e busca por parte do usuário linguísta.

Baseando-nos nesta premissa, apresentamos nesta seção exemplos de pesquisas na área da sintaxe que se utilizaram do recurso de buscas automáticas em *corpora* e poderiam ser auxiliadas pela ferramenta.

Tomemos como primeiro exemplo a pesquisa realizada por Antonelli (2011), que teve como objetivo investigar a relação entre a sintaxe da posição do verbo e as propriedades de fronteamento de sintagmas na história do Português Europeu. Para esse trabalho de pesquisa o

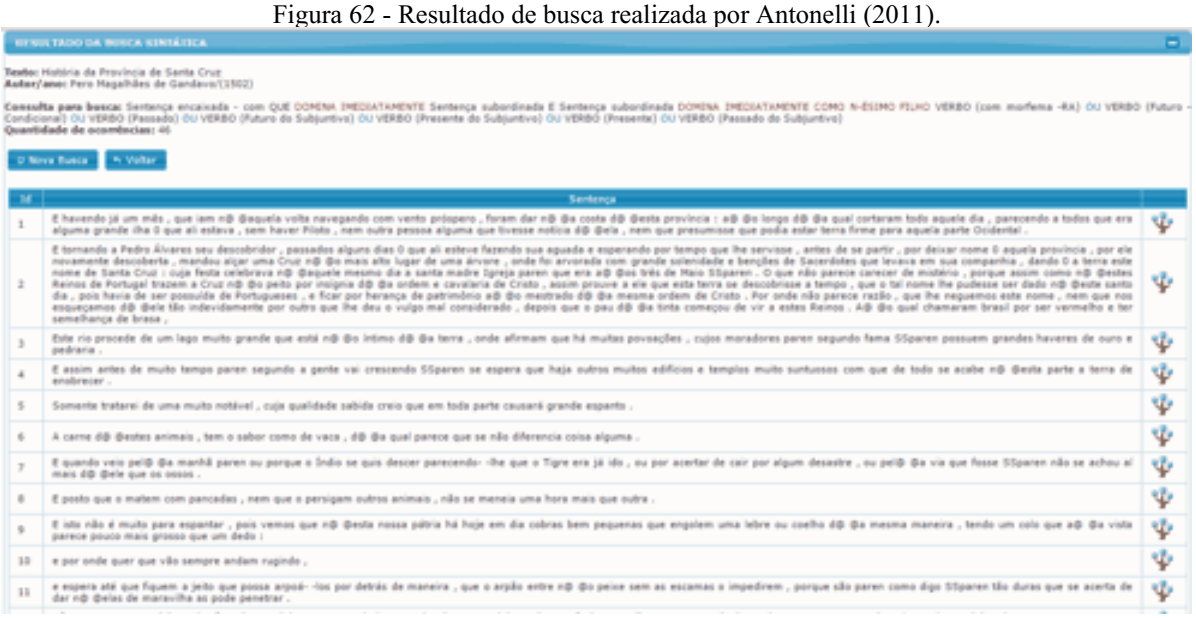
autor fez buscas automáticas em orações finitas, focalizando as sentenças declarativas. Nas orações dependentes, as buscas foram feitas em orações complemento introduzidas pelo complementaizador "que". O trabalho procurou descrever as possíveis ordens de palavras nos contextos sintáticos investigados, dando destaque principalmente a aspectos da posição linear do verbo em relação a outros constituintes. Um exemplo de busca realizada, portanto, foi de sequências empregadas em orações principais, em que o verbo aparece lineramente em primeira, segunda ou terceira posição. A ferramenta WebSinC poderia auxiliar nesta pesquisa realizando buscas por sentença completiva (CP-THT) que domine imediatamente uma sentença subordinada (IP-SUB), que por sua vez domine imediatamente como primeiro, segundo ou terceiro filho um verbo flexionado no presente, passado ou futuro, do indicativo ou subjuntivo, ou um verbo com o morfema "ra". A configuração desta busca considerando a dominância imediata do verbo como segundo filho (N=2) é exibida na figura 61.

Figura 61 - Configuração gráfica de busca sintática realizada por Antonelli (2011).



Aplicando esta busca realizada por Antonelli (2011) sobre o arquivo correspondente ao texto de Pero Magalhães de Gandavo (1502), obtemos 46 sentenças de resultado para este texto, conforme mostra a figura 62.

Figura 62 - Resultado de busca realizada por Antonelli (2011).



Outro exemplo é a pesquisa publicada em Namiuti (2011), cujo objetivo foi investigar a relação entre o fenômeno do fronteamento e da interpolação na diacronia do português. Para tanto, a autora buscou, com o auxílio da ferramenta *Corpus Search*, às várias ordenações de constituintes em sete textos sintaticamente anotados do *Corpus* Anotado do Português Histórico – Tycho Brahe, mais especificamente, em textos representativos dos século XVI, XVII, XVII e XIX. Dentre as buscas realizadas nesse trabalho, reproduzimos uma com a ferramenta WebSinc responsável por retornar sentenças onde há sentença subordinada (IP-SUB) que domine um sintagma nominal objeto direto (NP-ACC) que precede verbo (VB*). A configuração desta busca no WebSinC pode ser vista na figura 63.

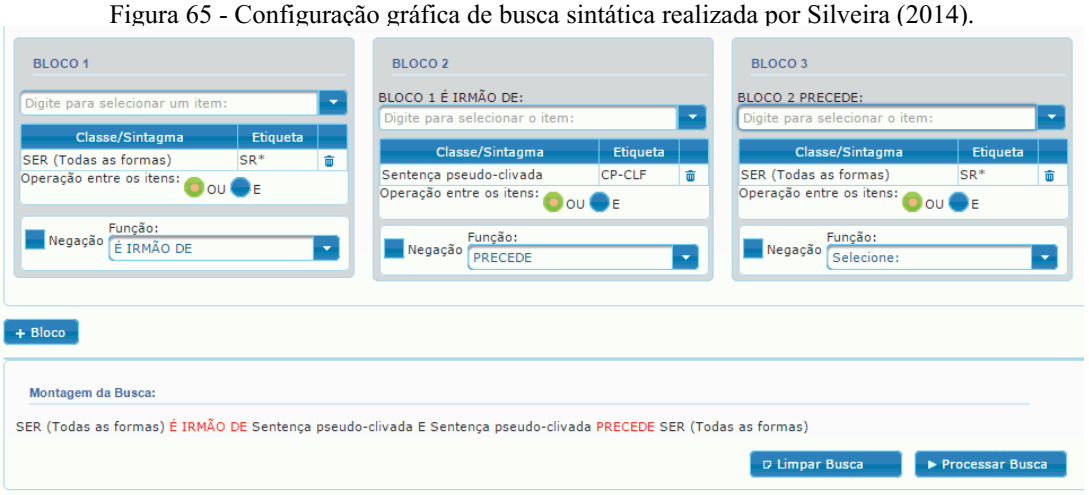
Figura 63 - Configuração gráfica de busca sintática realizada por Namiuti (2011).

Aplicando esta busca realizada por Namiuti (2011) sobre o arquivo correspondente ao texto de Pero Magalhães de Gandavo (1502), obtemos 4 sentenças de resultado para este texto, conforme mostra a figura 64.

Figura 64 - Resultado de busca realizada por Namiuti (2011).

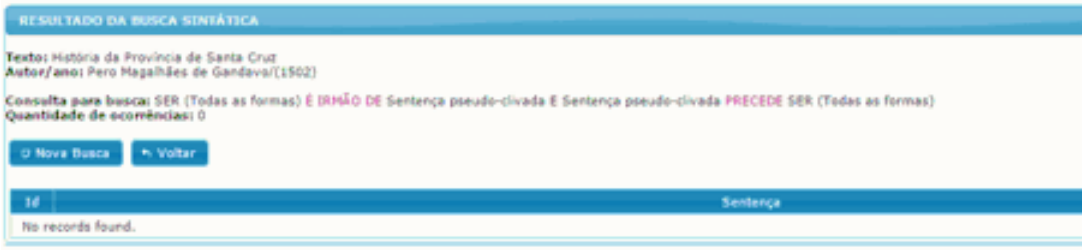
A pesquisa de Silveira (2014) investigou as sentenças utilizadas para focalizar constituintes sintáticos, as Clivadas e Pseudo-clivadas, sob o ponto de vista da Gramática Gerativa. As buscas foram realizadas com o *Corpus Search* em textos do *Corpus* Tycho Brahe com anotação sintática. Dentre as buscas realizadas pela autora, citamos como exemplo a busca por Pseudo-Clivadas - sentenças com um verbo *ser* flexionado (SR*) tendo como irmão um CP do tipo clivado (CP-CLF), precedendo o verbo ser. A configuração desta busca no WebSinC é exibida na figura 65.

Figura 65 - Configuração gráfica de busca sintática realizada por Silveira (2014).



Aplicando esta busca realizada por Silveira (2014) sobre o arquivo correspondente ao texto de Pero Magalhães de Gandavo (1502), não obtemos nenhuma sentença de resultado para este texto, conforme mostra a figura 66.

Figura 66 - Resultado de busca sintática realizada por Silveira (2014).



Diversos trabalhos de pesquisa na área da sintaxe diacrônica realizaram buscas automáticas em *corpora* anotados nos mesmos moldes do CTB utilizando-se da ferramenta *Corpus Search*, para os quais a ferramenta WebSinC teria aplicabilidade. Citamos aqui alguns trabalhos que utilizaram o corpus Tycho Brahe:

- Gravina (2014), cuja pesquisa teve objetivo de investigar a relação entre o sujeito nulo parcial e a inversão do sujeito na diacronia do português brasileiro. Realizou buscas no *corpus* com uso do *Corpus Search* a fim de identificar o número de sujeitos nulos nos textos. As buscas realizadas no trabalho utilizaram *Corpus Search.*
- Gomes do Santos (2013), cuja pesquisa teve objetivo de "mapear as ocorrências de NPs acusativos não clíticos em diferentes ambientes sintáticos em textos do século XII ao XIX, com o intuito de delinear as mudanças gramaticais ocorridas". As buscas da pesquisa foram realizadas em textos sintaticamente anotados do *corpus* Tycho Brahe com uso da ferramenta *Corpus Search*. As buscas realizadas no trabalho utilizaram *Corpus Search.*
- Andrade (2010), cuja pesquisa teve como objetivo investigar “a subida de clíticos no português europeu dos séculos XVI a XX”. As buscas realizadas no trabalho utilizaram *Corpus Search.*
- Lopes (2010), cuja pesquisa teve como objetivo investigar “a ênclise em orações dependentes na história do Português Europeu”. As buscas realizadas no trabalho utilizaram *Corpus Search.*

8.4 Resultados de buscas morfossintáticas

Com o WebSinC, as buscas morfossintáticas no corpus DOViC podem ser feitas no arquivo XML gerado pelo E-Dictor, não havendo necessidade de transformação para um arquivo POS e consequente perda de informação. O arquivo XML que o E-Dictor produz possui

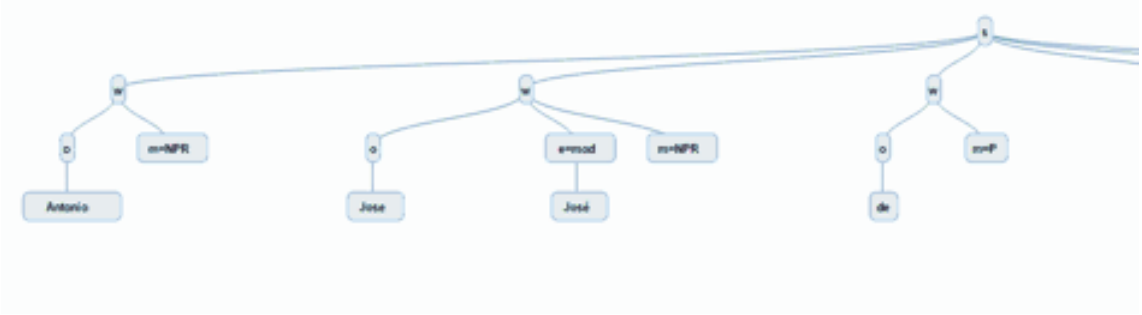
anotações da estrutura dos textos, anotações morfossintáticas e de edições. As buscas são feitas pelos elementos <m> com o atributo "v" dentro do arquivo, que indicam a anotação da categoria POS. O WebSinC fornece o resultado das buscas trazendo as sentenças na forma original do texto, possibilitando também exibi-lo na forma modernizada. Isto se dá em virtude do arquivo XML conter tanto as anotações POS quantos as anotações de edições. Com a ferramenta *Corpus Search*, o resultado de buscas morfossintáticas é exibido apenas na forma modernizada e não há possibilidade de exibi-lo na forma original do texto. Isto deve-se ao fato de que o arquivo lido pelo *Corpus Search* para esse tipo de busca é o arquivo POS gerado pelo *tagger*, que recebe como entrada o arquivo na forma modernizada. Assim, há perda de informação nas buscas morfossintáticas feitas no arquivo não XML com a ferramenta *Corpus Search.*

As buscas morfossintáticas no WebSinc são montadas graficamente assim como demonstrado no capítulo anterior. Ao processar a consulta, o resultado da busca é exibido da mesma forma que o resultado das buscas sintáticas. As sentenças são exibidas numa tabela, com cada sentença numa linha e também um atalho para visualização da sentença no formato gráfico de árvore. Todos os nós folha na árvore gerada para os arquivos de anotação morfossintática têm a mesma altura, uma vez que não existe a estrutura hierárquica de constituintes, mas apenas a hierarquia de parágrafos, sentenças e palavras.

A figura 67 mostra como exemplo a tela de resultado para uma busca morfossintática realizada com o documento "Carta de Liberdade da Cabra de nome Sofia", do *corpus* DOViC. A busca procurou por sentenças onde um nome no singular existe. A figura 68 mostra a visualização de uma das sentenças do resultado em forma de árvore, conforme estrutura hierárquica do arquivo XML gerado pelo E-Dictor.

Figura 67 - Tela de resultado de busca morfossintática.

Figura 68 - Visualização em árvore de sentença de resultado.



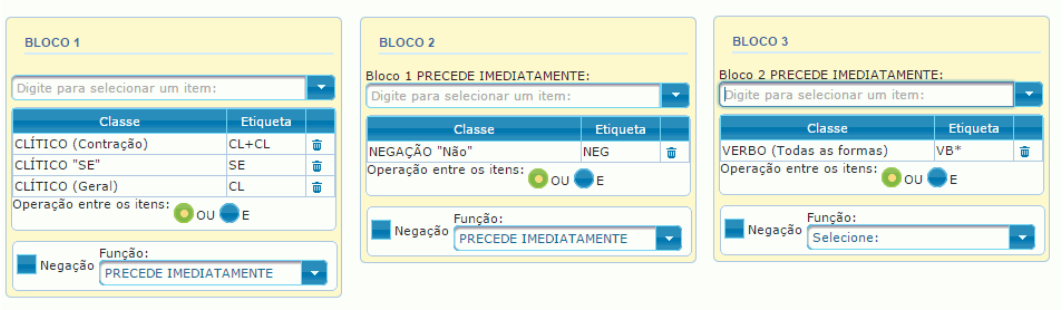
8.4.1 Aplicação das buscas morfossintáticas em pesquisas linguísticas

A pesquisa de Namiuti (2008) realizou buscas em textos anotados morfossintaticamente, usando scripts na linguagem Perl. Dentre as buscas realizadas nesse trabalho, citamos como exemplo algumas buscas morfossintáticas que poderiam ser realizadas com a ferramenta WebSinc:

1. Buscas por adjacências Clítico-Negação-Verbo (CL ou SE ou CL+CL precede imediatamente NEG, que precede imediatamente VB*)
2. Buscas por adjacências Negação-Clítico-Verbo (NEG precede imediatamente CL ou SE ou CL+CL, que precede imediatamente VB*)
3. Buscas por próclise com adjacência clítico-Verbo (CL precede VB*)

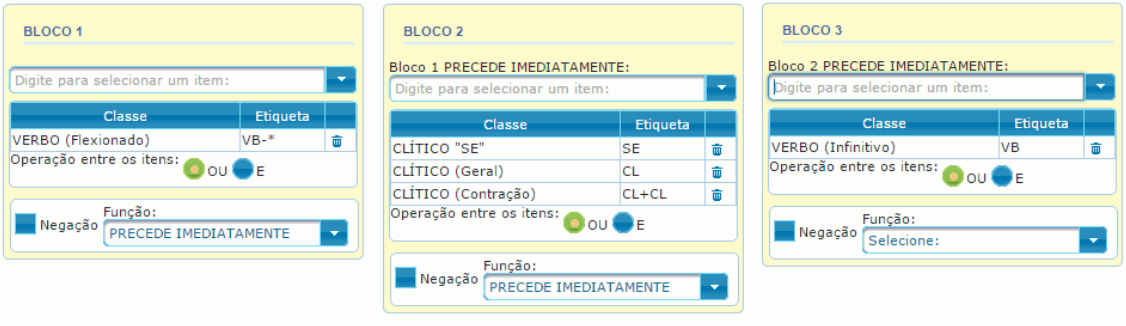
A busca do exemplo 1 poderia ser realizada com a configuração de busca morfossintática no WebSinc mostrada na figura 69.

Figura 69 - Configuração gráfica de busca mofossintática realizada por Namiuti (2008).



Lourençato (2001) em seu trabalho em sintaxe diacrônica investigou a colocação de clíticos em orações infinitivas introduzidas por preposição no português clássico. Dentre as buscas realizadas para esse trabalho, que poderiam ser realizadas com a ferramenta WebSinc, ilustraremos, como exemplo, a busca por verbos flexionados seguidos de próclise, seguida de verbo no infinitivo. A figura 70 mostra a configuração no WebSinc para essa busca.

Figura 70 - Configuração gráfica de busca mofossintática realizada por Lourençato (2001).



Aplicando as buscas realizadas por Lourençato (2001) e Namiuti (2208) sobre o arquivo XML correspondente à carta de Alforria da cabra de nome Sofia (1845), não obtemos nenhuma sentença de resultado para este texto.

Outros trabalhos de pesquisa realizaram buscas morfossintáticas em *corpora* anotados nos mesmos moldes do CTB, para os quais a ferramenta WebSinC também teria aplicabilidade:

- Trannin (2010), cuja pesquisa analisou os complementos infinitivos selecionados por verbos causativos na história do Português Europeu. As buscas foram realizadas em texto anotados morfossintaicamente com o uso do *Corpus Search*.
- Floripi (2008) cuja pesquisa teve como objetivo descrever e analisar a variação do uso do determinante em estruturas com sintagmas nominais possessivos, dentro de uma perspectiva diacrônica. As buscas realizadas no trabalho utilizaram scripts na linguagem Perl.

- Cavalcante (2006) cujo trabalho investiga “o uso de 'se' com infinitivo na história do Português: do Português Clássico ao Português Europeu e Português Brasileiro modernos”. As buscas realizadas no trabalho utilizaram scripts na linguagem Perl.
- Godoy (2006), que também estuda a colocação de clíticos em orações infinitivas introduzidas por preposição no português clássico. As buscas realizadas no trabalho utilizaram scripts na linguagem Perl.
- Paixão de Sousa (2004), cujo trabalho investiga a sintaxe dos clíticos e a posição do sujeito na História do Português. As buscas realizadas no trabalho utilizaram scripts na linguagem Perl.
- Menezes (2003), cujo trabalho investiga a colocação de clíticos em orações coordenadas no português clássico. Na pesquisa foram realizadas buscas em arquivos do CTB anotados morfossintaticamente. As buscas realizadas no trabalho utilizaram scripts na linguagem Perl.

8.5 Avaliação do resultados das buscas

Existem diversos métodos para avaliação dos resultados produzidos pelo sistema. Hirschman e Mani (2003) dizem que a saída pode ser avaliada por si mesma, pode ser comparada com outras saídas, ou contrastada com o resultado esperado para determinada entrada.

Para avaliar os resultados produzidos pelas buscas sintáticas, foi utilizado o método de comparação da saída do WebSinC com a saída produzida por outra ferramenta, o *Corpus Search*. Para cada função sintática implementada pela ferramenta WebSinC, foram feitos testes utilizando as duas ferramentas. Os resultados foram comparados verificando o número total de ocorrências para a busca em cada ferramenta e a igualdade das sentenças retornadas. O resultado esperado dos testes é que as buscas retornem as mesmas sentenças em ambas as ferramentas para todas as buscas realizadas e que os totais de ocorrências contabilizados por cada uma também sejam idênticos.

A avaliação foi feita através da observação de cada sentença trazida como resultado. Para fins práticos, as expressões de busca escolhidas para os testes foram as que retornavam até trinta sentenças como resultado.

8.5.1 Testes das buscas sintáticas

O arquivo utilizado para os testes das buscas sintáticas foi o arquivo g_008_psd.txt correspondente ao texto "História da Província de Santa Cruz", escrito em 1502 por Pero Magalhães Gandavo, do *corpus* Tycho Brahe, escolhido aleatoriamente entre os diversos arquivos disponíveis no site do projeto do *corpus*. O teste para buscas sintáticas não foi realizado em textos do *corpus* DOViC devido à não completude de anotação sintática para nenhum documento deste *corpus* até o desenvolvimento desta pesquisa. No entanto, como o *corpus* DOViC utiliza a mesma metodologia de anotação, consideramos os mesmos testes como válidos para o *corpus* DOViC.

Para as buscas na ferramenta WebSinC, esse arquivo PSD do Tycho Brahe serviu como entrada para o algoritmo que transforma a anotação *Penn TreeBank* em XML. Portanto, as buscas sintáticas realizadas nos testes com o WebSinC foram feitas com o arquivo g_008_xml.xml que a ferramenta gerou, correspondente ao arquivo g_008_psd.txt do Tycho Brahe. Cada função de busca implementada na ferramenta WebSinc foi testada. Para cada função, foram elaborados quatro testes, de maneira que fossem testados : a função com apenas um operando em cada bloco, a função com pelo menos dois operandos em cada bloco usando a operação lógica OU entre eles, a função com pelo menos dois operandos em cada bloco usando a operação lógica E entre eles, e a função com o operador lógico de negação. Para a maioria dos testes com o operador de negação, foi acrescentada mais uma cláusula com operador lógico E, com objetivo de restringir o resultado a um número menor ou igual a trinta sentenças, já que apenas com a negação os testes trouxeram resultados na ordem de centenas de sentenças. A tabela 1 mostra estes testes realizados com os respectivos resultados.

Tabela 1 - Resultados dos testes realizados nas buscas sintáticas

Função: EXISTÊNCIA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
01	Verbo estar no infinitivo **existe**	ET exists	9	9	Sim
02	Adjetivo superlativo feminino no singular **OU** Adjetivo superlativo masculino no singular **existem**	(ADJ-S-F exists) OR (ADJ-S exists)	2	2	Sim
03	Clítico "Se" **E** Negação "não" existem	(SE exists) AND (NEG exists)	2	2	Sim
04	NP Sujeito **NÃO existe** e sintagma adverbial **existe**	NOT(NP-SBJ exists) AND (ADVP exists)	12	12	Sim
Função: PRECEDÊNCIA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
05	Verbo estar no infinitivo **precede** clítico	ET precedes CL	6	6	Sim
06	Verbo estar no infinitivo **OU** verbo ter no infinitivo **precedem** clítico	(TR precedes CL) OR (ET precedes CL)	11	11	Sim
07	Verbo ter no infinitivo **E** verbo no infinitivo **precedem** clítico	(TR precedes CL) AND (VB precedes CL)	3	3	Sim
08	Verbo estar no infinitivo existe e **NÃO precede** clítico	NOT(ET precedes CL) AND (ET exists)	3	3	Sim
Função: PRECEDÊNCIA IMEDIATA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
09	Projeção adverbial **precede imediatamente** um nome prórpio no singular	ADJP iprecedes NPR	2	2	Sim
10	Projeção adverbial **OU** determinante definido feminino no singular **precedem imediatamente** um nome próprio no singular	(ADJP iprecedes [1]NPR) OR (D-F iprecedes [2]NPR)	21	21	Sim
11	Projeção adverbial **E** determinante definido masculino no singular **precedem imediatamente** um nome próprio no singular	(ADJP iprecedes [1]NPR) AND (D iprecedes [2]NPR)	1	1	Sim
12	Adjetivo feminino no singular existe e **NÃO precede imediatamente** um nome prórpior no singular e NP genitivo existe	NOT(ADJ-F iprecedes NPR) AND (NP-GEN exists) AND (ADJ-F exists)	8	8	Sim
Função: DOMINÂNCIA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
13	Sentença encaixada com "que" **domina** Adjetivo gênero duplo no singular	(CP-THT dominates ADJ -G)	9	9	Sim
14	Sintagma nominal **OU** sintagma preposicional **dominam** sintagma nominal genitivo	(NP dominates NP-GEN) OR (PP dominates NP-GEN)	24	24	Sim
15	Sentença encaixada com "que" **domina** Adjetivo gênero duplo no singular **E** Sentença subordinada	(CP-THT dominates ADJ-G) AND (CP-THT dominates IP-SUB)	9	9	Sim
16	Sentença no gerúndio existe e **NÃO domina** um NP sujeito	NOT(IP-GER dominates NP-SBJ) AND (IP-GER exists)	27	27	Sim

Função: DOMINÂNCIA IMEDIATA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências**	**Total de ocorrências WebSinC**	**Resultado traz as**

			Corpus Search		mesmas sentenças
17	Sentença subordinada **domina** imediatamente verbo ter no passado do subjuntivo	(IP-SUB idominates TR-SD)	1	1	Sim
18	Sintagma nominal **OU** Sintagma preposicional **dominam** imediatamente um número cardinal	(NP idominates NUM) OR (PP idominates NUM)	26	26	Sim
19	Sintagma nominal **domina** imediatamente um número cardinal **E** um determinante feminino no plural	([1]NP idominates NUM) AND ([2]NP idominates D-F-P)	7	8	Sim
		(NP idominates NUM) AND (NP idominates D-F-P)	1		
20	Sintagma nominal **NÃO domina** imediatamente um nome próprio no plural e um nome próprio no plural existe	NOT(NP idominates NPR-P) AND (NPR-P exists)	19	19	Sim
Função: IRMANDADE					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
21	NP sujeito **é irmão** de verbo estar no passado	NP-SBJ hassister ET-D	15	15	Sim
22	NP sujeito **OU** NP objeto direto **são irmãos** de verbo estar no passado	(NP-SBJ hassister ET-D) OR (NP-ACC hassister ET-D)	15	15	Sim
23	NP sujeito **E** Sintagma adverbial **são irmãos** de verbo estar no passado	(NP-SBJ hassister ET-D) AND(ADVP hassister ET-D)	9	9	Sim
24	Sentença relativa adjungida existe e **NÃO é irmão** de sintagman preposicional	(NOT(CP-CAR hassister PP) AND (CP-CAR exists))	5	5	Sim
Função: C-COMANDO					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
25	NP **c-comanda** um clítico	NP ccommands CL	9	9	Sim
	NP **c-comanda** um clítico **OU** NP objeto direto	(NP ccommands CL) OR (NP ccommands NP-ACC)	24	24	Sim
26	NP **c-comanda** um clítico **E** NP objeto direto	(NP ccommands CL) AND (NP ccommands NP-ACC)	7	7	Sim
27	NP sujeito **NÃO c-comanda** NP e small clause existe	NOT(NP-SBJ ccommands NP) AND (IP-SMC exists)	17	17	Sim
Função: DOMINÂNCIA N PALAVRAS					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
28	NP sujeito **domina** 8 palavras	NP-SBJ domswords 8	14	14	Sim
29	NP sujeito **OU** NP objeto direto **dominam** 12 palavras	(NP-SBJ domswords 12) OR (NP-ACC domswords 12)	14	14	Sim
30	NP sujeito **E** NP objeto direto **dominam** 3 palavras	(NP-SBJ domswords 3) AND (NP-ACC domswords 3)	9	9	Sim
31	Projeção adjetival **NÃO domina** 3 palavras e NP sujeito domina 12 palavras	(NOT(ADJP domswords 3) AND (NP-SBJ domswords 12))	6	6	Sim
Função: DOMINÂNCIA DE MAIS DE N PALAVRAS					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**

32	NP sujeito **domina mais** de 20 palavras	NP-SBJ domswords> 20	15	15	Sim
33	NP sujeito **OU** NP objeto direto **dominam mais** de 35 palavras	(NP-SBJ domswords> 35) OR (NP-ACC domswords> 35)	24	24	Sim
34	NP sujeito **E** NP objeto direto **dominam mais** de 20 palavras	(NP-SBJ domswords> 20) AND (NP-ACC domswords> 20)	2	2	Sim
35	Projeção adjetival **NÃO domina mais** de 3 palavras e NP sujeito domina 12 palavras	(NOT(ADJP domswords> 3) AND (NP-SBJ domswords 12))	3	3	Sim
Função: DOMINÂNCIA DE MENOS DE N PALAVRAS					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
36	NP apositivo ou parentético **domina menos** de 3 palavras	NP-PRN domswords< 3	15	15	Sim
37	NP apositivo ou parentético **OU** sentença no gerúndio **dominam menos** de 2 palavras	(NP-PRN domswords< 2) OR (IP-GER domswords< 2)	4	4	Sim
38	NP apositivo ou parentético **E** sintagma nominal **dominam menos** de 3 palavras	(NP-PRN domswords< 3) AND (NP domswords< 3)	14	14	Sim
39	Projeção adjetival **NÃO domina menos** de 3 palavras e NP sujeito domina 12 palavras	(NOT(ADJP domswords< 3) AND (NP-SBJ domswords 12))	4	4	Sim
Função: DOMINÂNCIA IMEDIATA COMO PRIMEIRO FILHO					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
40	Sentença infinitiva **domina imediatamente como primeiro filho** um verbo haver no presente	IP-INF idomsfirst HV	7	7	Sim
41	Sentença infinitiva **domina imediatamente como primeiro filho** um verbo haver no presente **OU** verbo condicional no futuro	(IP-INF idomsfirst HV) OR (IP-INF idomsfirst VB-R)	7	7	Sim
42	Sentença infinitiva **E** Sentença subordinada **dominam imediatamente como primeiro filho** um verbo haver no presente	(IP-INF idomsfirst HV) AND (IP-SUB idomsfirst HV)	0	0	-
43	Sintagma conjuncional **NÃO domina imediatamente como primeiro filho** uma conjunção negativa e conjunção negativa existe	NOT(CONJP idomsfirst CONJ-NEG) AND (CONJ-NEG exists)	16	16	Sim
Função: DOMINÂNCIA IMEDIATA COMO ÚLTIMO FILHO					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
44	Sentença infinitiva **domina imediatamente como último filho** um verbo haver no presente	IP-INF idomslast HV	1	1	Sim
45	Sentença infinitiva **domina imediatamente como último filho** um verbo haver no presente **OU** verbo condicional no futuro	(IP-INF idomslast HV) OR (IP-INF idomslast VB-R)	1	1	Sim
46	Sentença no gerúndio **E** Sentença subordinada **dominam imediatamente como último filho** um sintagma preposicional	(IP-GER idomslast PP) AND (IP-SUB idomslast PP)	0	0	-
47	Sentença relativa adjungida existe e **NÃO domina imediatamente como último filho** uma sentença subordinada	NOT(CP-CAR idomslast IP-SUB) AND (CP-CAR exists)	1	1	Sim

Função: DOMINÂNCIA IMEDIATA COMO N-ÉSIMO FILHO					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
48	Sentença infinitiva **domina imediatamente como segundo filho** um verbo haver no presente	IP-INF idomsnumber 2 HV	12	12	Sim
49	Sentença infinitiva **domina imediatamente como segundo filho** um verbo haver no presente **OU** verbo condicional no futuro	(IP-INF idomsnumber 2 HV) OR (IP-INF idomsnumber 2 VB-R)	12	12	Sim
50	Sintagma nominal **domina imediatamente como primeiro filho** um determinante defindo feminino singular **E** um outro NP **domina imediatamente como primeiro filho** um nome próprio no singular e um número cardinal existe	([1]NP idomsnumber 1 D-F) AND ([2]NP idomsnumber 1 NPR) AND (NUM exists)	11	11	Sim
51	Sentença relativa adjungida existe e **NÃO domina imediatamente como segundo filho** uma sentença subordinada	NOT(CP-CAR idomsnumber 2 IP-SUB) AND (CP-CAR exists)	1	1	Sim
Função: DOMINÂNCIA IMEDIATA COMO ÚNICO FILHO					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
52	NP sujeito **domina imediatamente como único filho** um nome próprio no singular	(NP-SBJ idomsonly NPR)	4	4	Sim
53	NP sujeito **domina imediatamente como único filho** um nome próprio no singular **OU** um nome próprio no plural	([1]NP-SBJ idomsonly NPR) OR ([2]NP-SBJ idomsonly NPR-P)	5	5	Sim
54	NP **domina imediatamente como único filho** um elemento relativo no singular **E** um nome próprio no singular	([1]NP idomsonly WPRO) AND ([2]NP idomsonly NPR)	7	7	Sim
55	NP objeto direto **NÃO domina imediatamente como único filho** um nome e sentença relativa adjungida existe	NOT(NP-ACC idomsonly N) AND (CP-CAR exists)	14	14	Sim
Função: DOMINÂNCIA IMEDIATA DE N FILHOS					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
56	NP sujeito **domina imediatamente** 6 filhos	NP-SBJ idomstotal 6	2	2	Sim
57	NP sujeito **OU** NP **dominam imediatamente** 6 filhos	(NP idomstotal 6) OR (NP-SBJ idomstotal 6)	4	4	Sim
58	NP sujeito **E** NP **dominam imediatamente** 4 filhos	(NP idomstotal 4) AND (NP-SBJ idomstotal 4)	13	13	Sim
59	NP **NÃO domina imediatamente** 4 filhos e sentença relativa adjungida existe	NOT(NP idomstotal 4) AND (CP-CAR exists)	10	10	Sim
Função: DOMINÂNCIA IMEDIATA DE MENOS DE N FILHOS					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
60	Claúsula adverbial **domina imediatamente menos** de 2 filhos	(CP-ADV idomstotal< 2)	1	1	Sim
61	Claúsula adverbial **OU** sentença comparativa **dominam imediatamente menos** de 2 filhos	(CP-ADV idomstotal< 2) OR (CP-CMP idomstotal< 2)	30	30	Sim
62	Sintagma nominal **E** sentença comparativa **dominam imediatamente menos** de 2 filhos	(NP idomstotal< 2) AND (CP-CMP idomstotal< 2)	21	21	Sim
63	Sentença comparativa **NÃO domina imediatamente menos** de 2 filhos e sentença relativa adjungida existe	NOT(CP-CMP idomstotal< 2) AND (CP-CAR exists)	13	13	Sim

Função: DOMINÂNCIA IMEDIATA DE MAIS DE N FILHOS					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
64	NP sujeito **domina imediatamente mais** de 4 filhos	NP-SBJ idomstotal> 4	12	12	Sim
65	NP sujeito **OU** NP **dominam imediatamente mais** de 5 filhos	(NP idomstotal>5) OR (NP-SBJ idomstotal> 5)	5	5	Sim
66	NP sujeito **E** NP **dominam imediatamente mais** de 3 filhos	(NP idomstotal> 3) AND (NP-SBJ idomstotal> 3)	20	20	Sim
67	Sentença comparativa **NÃO domina imediatamente mais** de 2 filhos e sentença relativa adjungida existe	NOT(CP-CMP idomstotal> 2) AND (CP-CAR exists)	15	15	Sim

Uma vez testadas todas as funções individualmente e com os operadores lógicos E, OU e NÃO, foram elaborados mais quatro casos de teste para verificação da operação E entre os blocos, aumentando a cobertura dos testes. Casos com dois blocos já foram contemplados nos testes da tabela 1. Portanto, foram acrescentados testes com três, quatro e cinco blocos, com escolha de funções e parâmetros aleatórios, com a restrição de que o resultado não exceda trinta sentenças. Os resultados destes testes são elencados pela tabela 2.

Tabela 2 - Resultados dos testes entre blocos realizados nas buscas sintáticas

Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
01	TESTE COM BLOCOS 1, 2 e 3: Sintagma preposicional domina imediatamente um sintagma noninal **E** um sintagma nominal domina imediatamente um sintagma genitivo	(PP idominates NP) AND (NP idominates NP -GEN)	18	18	Sim
02	TESTE COM BLOCOS 1, 2, 3 e 4: Sintagma preposicional domina imediatamente um sintagma noninal **E** um sintagma nominal domina imediatamente um sintagma genitivo **E** sintagma no genitivo é irmão de um nome no plural	(PP idominates [1]NP) AND ([2]NP idominates [3]NP-GEN) AND ([4]NP-GEN hassister N-P)	2	2	Sim
03	TESTE COM BLOCOS 1, 2, 3, 4 e 5: Sintagma preposicional domina imediatamente um sintagma noninal **E** um sintagma nominal domina imediatamente um sintagma genitivo **E** sintagma no genitivo é irmão de um nome no plural **E** um determinante definido masculino plural existe	(PP idominates [1]NP) AND ([2]NP idominates [3]NP-GEN) AND ([4]NP-GEN hassister N-P) AND (D-P exists)	2	2	Sim

8.5.1.1 Discussão dos testes para buscas sintáticas

Para todos os testes realizados, as sentenças retornadas nas buscas com o WebSinC foram iguais às retornadas nas buscas com o *Corpus Search*.

No teste n° 19 da tabela 1, foram necessárias duas consultas na ferramenta *Corpus Search* para que o resultado fosse correspondente. Quando a mesma etiqueta é usada numa

consulta na expressão de busca, o *Corpus Search* faz a busca pelo mesmo nó. Assim, a expressão (**NP idominates NUM) AND (NP idominates D-F-P)** buscará por sentenças em que um NP domine imediatamente um número cardinal e esse mesmo nó NP domine também um determinante feminino no plural. Para considerar um outro NP, deve-se colocar índices numéricos entre colchetes precedendo a etiqueta. Sendo assim, a expressão **([1]NP idominates NUM) AND ([2]NP idominates D-F-P)** buscará por sentenças em que um NP domine imediatamente um número cardinal e um outro nó NP diferente do primeiro domine um determinante feminino no plural. Já o WebSinC busca por qualquer NP que domine imediatamente um número cardinal e também um determinante feminino no plural na mesma sentença. O NP pode ser o mesmo, ou pode ser outro. Por isso, o resultado da busca no WebSinC traz oito sentenças, que correspondem à soma das sentenças trazidas pelas buscas no *Corpus Search*.

A tabela 3 resume os dados referentes aos testes sintáticos realizados e aqui discutidos. Dezessete funções sintáticas foram implementadas no WebSinc, para as quais foram elaborados e executados 71 testes. Todos estes testes retornaram resultados iguais em ambas as ferramentas, WebSinc e Corpus Search.

Tabela 3 - Resumo dos resultados dos testes sintáticos

Total de funções sintáticas implementadas	17
Número de testes realizados	71
Número de testes com resultados iguais	71
Número de testes com resultados diferentes	0

8.5.2 Testes das buscas morfossintáticas

O arquivo utilizado para os testes das buscas morfossintáticas no WebSinC foi o arquivo "1.29.xml" correspondente ao texto "Carta de alforria da Cabra de nome Sofia" do *corpus* DOViC. Esse texto foi escolhido entre os seis arquivos anotados disponíveis por ser o maior arquivo dentre eles. Para realização das buscas morfossintáticas no *Corpus Search*, o arquivo deve estar no formato de anotação POS. Para tanto, foi usado o arquivo de anotação gerado na ferramenta E-Dictor com as etiquetas morfossintáticas, o qual foi editado manualmente em seguida para adequar-se ao formato exigido pelo *Corpus Search*. Os arquivos 1.29.xml e o arquivo POS editado manualmente encontram-se no anexo A.

A tabela 4 mostra os testes realizados com os respectivos resultados.

Tabela 4 - Resultados dos testes realizados nas buscas morfossintáticas.

Função: EXISTÊNCIA

Teste nº	Busca WebSinC	Busca *Corpus Search*	Total de sentenças *Corpus Search*	Total de sentenças WebSinC	Resultado traz as mesmas sentenças
01	Conjunção subordinada **existe**	CONJS exists	1	1	Sim
02	Conjunção **E** Verbo no presente **existem**	(CONJ exists) AND (VB-P exists)	5	5	Sim
03	Conjunção **OU** Verbo no presente **existem**	(CONJ exists) OR (VB-P exists)	12	12	Sim
04	Nome próprio no singular **NÃO existe**	NOT (NPR exists)	4	3	Sim
Função: PRECEDÊNCIA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de sentenças *Corpus Search***	**Total de sentenças WebSinC**	**Resultado traz as mesmas sentenças**
05	Nome próprio no singular **precede** advérbio	NPR precedes ADV	2	2	Sim
06	Adjetivo feminino plural **E** conjunção **precedem** nome singular	(ADJ-F-P precedes N) AND (CONJ precedes N)	1	1	Sim
07	Adjetivo feminino plural precede **OU** conjunção **precedem** nome singular	(ADJ-F-P precedes N) OR (CONJ precedes N)	5	5	Sim
08	Número cardinal **NÃO precede** conjunção	NOT (NUM precedes CONJ)	14	13	Sim
Função: PRECEDÊNCIA IMEDIATA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de sentenças *Corpus Search***	**Total de sentenças WebSinC**	**Resultado traz as mesmas sentenças**
09	Nome próprio no singular **precede imediatamente** advérbio	NPR iprecedes ADV	2	2	Sim
10	Nome próprio no singular **E** preposição **precede imediatamente** advérbio	(NPR iprecedes [1]ADV) AND (P iprecedes [2]ADV)	1	1	Sim
11	Nome próprio no singular **OU** preposição **precede imediatamente** advérbio	(NPR iprecedes ADV) OR (P iprecedes ADV)	2	2	Sim
12	Número cardinal **NÃO precede imediatamente** conjunção	NOT (NUM iprecedes CONJ)	14	13	Sim
Função: VIZINHANÇA (À DIREITA OU À ESQUERDA)					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de sentenças *Corpus Search***	**Total de sentenças WebSinC**	**Resultado traz as mesmas sentenças**
13	Nome no singular tem uma preposição como segundo **vizinho**	(N Neighborhood 2 P)	5	2	Não
14	Nome no singular **E** adjetivo gênero duplo singular têm uma preposição como segundo **vizinho**	(N Neighborhood 2 [1]P) AND (ADJ-G Neighborhood 2 [2]P)	1	1	Sim
15	Nome no singular **OU** adjetivo gênero duplo singular têm uma preposição como segundo **vizinho**	(N Neighborhood 2 P) OR (ADJ-G Neighborhood 2 P)	5	2	Não
16	Nome próprio no singular **NÃO** tem outro nome próprio no singular como primeiro **vizinho**	NOT([1]NPR Neighborhood 1 [2]NPR)	12	11	Sim
Função: VIZINHANÇA À DIREITA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de sentenças *Corpus Search***	**Total de sentenças WebSinC**	**Resultado esperado**
17	Nome no singular tem uma preposição como segundo **vizinho à direita**	Não se aplica	-	1	Sim
18	Nome no singular **E** adjetivo gênero duplo singular têm uma	Não se aplica	-	1	Sim

	Preposição como segundo **vizinho à direita**				
19	Nome no singular **OU** adjetivo gênero duplo singular têm uma preposição como segundo **vizinho à direita**	Não se aplica	-	1	Sim
20	Nome próprio no singular **NÃO** tem outro nome próprio no singular como primeiro **vizinho à direita**	Não se aplica	-	11	Sim
Função: VIZINHANÇA À ESQUERDA					
	Busca WebSinC	**Busca *Corpus Search***	**Total de sentenças *Corpus Search***	**Total de sentenças WebSinC**	**Resultado esperado**
21	Preposição tem um nome no singular como segundo **vizinho à esquerda**	Não se aplica	-	1	Sim
22	Preposição têm nome no singular **E** adjetivo gênero duplo singular como segundo **vizinho à esquerda**	Não se aplica	-	1	Sim
23	Nome no singular **OU** adjetivo gênero duplo singular têm uma preposição como segundo **vizinho à direita**	Não se aplica	-	1	Sim
24	Nome próprio no singular existe e **NÃO** tem outro nome próprio no singular como primeiro **vizinho à esquerda**	Não se aplica	-	8	Sim
Função: INÍCIO DA SENTENÇA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de sentenças *Corpus Search***	**Total de sentenças WebSinC**	**Resultado esperado**
25	Nome próprio no singular no **início da sentença**	Não se aplica	-	13	Sim
26	Nome próprio no singular **E** nome no singular no **início da sentença**	Não se aplica	-	0	Sim
27	Nome próprio no singular **OU** preposição no **início da sentença**	Não se aplica	-	14	Sim
28	Nome próprio no singular **NÃO** está **início da sentença**	Não se aplica	-	7	Sim
Função: FIM DA SENTENÇA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de sentenças *Corpus Search***	**Total de sentenças WebSinC**	**Resultado esperado**
29	Nome próprio no singular no **fim da sentença**	Não se aplica	-	7	Sim
30	Nome próprio no singular **E** nome no singular no **fim da sentença**	Não se aplica	-	0	Sim
31	Nome próprio no singular **OU** número cardinal no **fim da sentença**	Não se aplica	-	12	Sim
32	Nome próprio no singular **NÃO** está no **fim da sentença**	Não se aplica	-	13	Sim
Função: POSIÇÃO N NA SENTENÇA					
Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de sentenças *Corpus Search***	**Total de sentenças WebSinC**	**Resultado esperado**
33	Nome próprio no singular na terceira **posição na sentença**	Não se aplica	-	5	Sim
34	Nome próprio no singular **E** nome no singular na terceira **posição na sentença**	Não se aplica	-	0	Sim

35	Nome próprio no singular **OU** preposição na terceira **posição na sentença**	Não se aplica	-	8	Sim
36	Nome próprio no singular **NÃO** está na terceira **posição na sentença**	Não se aplica	-	15	Sim

Uma vez testadas todas as funções individualmente e com os operadores lógicos E, OU e NÃO, foram elaborados mais quatro casos de teste para verificação da operação E entre os blocos, aumentando a cobertura dos testes. Casos com dois blocos já foram contemplados nos testes da tabela 3. Portanto, foram acrescentados testes com três, quatro e cinco blocos, com escolha de funções e parâmetros aleatórios, com a restrição de que o resultado não exceda trinta sentenças. Os resultados destes testes são elencados pela tabela 5.

Tabela 5 - Resultados dos testes entre blocos realizados nas buscas morfossintáticas.

Teste nº	**Busca WebSinC**	**Busca *Corpus Search***	**Total de ocorrências *Corpus Search***	**Total de ocorrências WebSinC**	**Resultado traz as mesmas sentenças**
01	TESTE COM BLOCOS 1, 2 e 3: Verbo no presente precede imediatamente um adjetivo feminino singular **E** um adjetivo feminino singular precede imediatamente uma conjunção.	(VB-P iprecedes ADJ-F-P) AND (ADJ-F-P iprecedes CONJ)	1	1	Sim
02	TESTE COM BLOCOS 1, 2, 3 e 4: Verbo no presente precede imediatamente um adjetivo feminino singular **E** um adjetivo feminino singular precede imediatamente uma conjunção **E** uma conjunção precede imediatamente um verbo no presente.	([1]VB-P iprecedes ADJ-F-P) AND (ADJ-F-P iprecedes CONJ) AND (CONJ iprecedes [2]VB-P)	1	1	Sim
03	TESTE COM BLOCOS 1, 2, 3, 4 e 5: Verbo no presente precede imediatamente um adjetivo feminino singular **E** um adjetivo feminino singular precede imediatamente uma conjunção **E** uma conjunção precede imediatamente um verbo no presente **E** un nome no singular existe	([1]VB-P iprecedes ADJ-F-P) AND (ADJ-F-P iprecedes CONJ) AND (CONJ iprecedes [2]VB-P) AND (N exists)	1	1	Sim

8.5.2.1 Discussão dos testes para buscas morfossintáticas

Para a maioria dos testes, as sentenças retornadas nas buscas com o WebSinC foram iguais às retornadas nas buscas com o *Corpus Search*. No teste nº 04 da tabela 4, o número de sentenças retornadas foi diferente. O *Corpus Search* retornou um total de quatro ocorrências, enquanto que o WebSinC trouxe apenas três. No entanto, as três ocorrências trazidas pelo WebSinC pertencem ao conjunto das quatro trazidas pelo *Corpus Search*, e a outra sentença faltante retorna uma *string* vazia. O mesmo ocorre com os testes de nºs 08, 12 e 16 da tabela 4.

O teste nº 13 da tabela 4 trouxe diferenças nos resultados, tanto na quantidade de ocorrências quanto no conteúdo das sentenças retornadas. Isto se dá devido à interpretação que o *Corpus Search* dá à estrutura do fomato POS para uso da função de vizinhança. Tomemos como exemplo para nossa explicação o trecho do arquivo POS mostrado no quadro 19. Para o *Corpus Search*, a palavra é um nó e a etiqueta também. Assim, no trecho exibido, a palavra "Reconheço" tem como primeiro vizinho a etiqueta "VB-P". A palavra "verdadeiras" é o segundo vizinho, a etiqueta "ADJ-F-P" é o terceiro e assim por diante.

Quadro 19- Trecho de arquivo POS utilizado nos testes com *Corpus* Search.

```
Reconheço/VB-P verdadeiras/ADJ-F-P e/CONJ dou/VB-P fé/N ./PONFP
```

O mesmo trecho é exibido no quadro 20 no formato XML que é usado pelo WebSinC para a consulta. Cada elemento <w> é considerado único. Assim, a palavra não é separada da sua etiqueta e, portanto, a palavra "Reconheço" tem como primeiro vizinho a palavra "verdadeiras". Esta sentença também seria retornada numa consulta por um verbo no presente (etiqueta VB-P) que tem como primeiro vizinho um adjetivo feminino no plural (etiqueta ADJ-F-P). Ou ainda um verbo no presente que tenha como primeiro vizinho a palavra "verdadeiras", porque a palavra e a sua etiqueta correspondente são considerados um único elemento ou nó.

Quadro 20 - Trecho do arquivo XML utilizado nos testes com WebSinC.

```
<w id="153">
   <o>Reconheço</o>
   <m v="VB-P"/>
 </w>
 <w id="154">
   <o>verdadeiras</o>
   <m v="ADJ-F-P"/>
 </w>
 <w id="155">
   <o>e</o>
   <m v="CONJ"/>
 </w>
 <w id="156">
   <o>dou</o>
   <m v="VB-P"/>
 </w>
 <w id="157">
   <o>fé</o>
   <m v="N"/>
 </w>
 <w id="158">
   <o>.</o>
   <m v="."/>
 </w>
```

Os testes de nºs 17 a 24 da tabela 4, que correspondem a funções de vizinhança à direita ou à esquerda não se aplicam ao *Corpus Search*, uma vez que esta ferramenta possui apenas a

função vizinhaça (*neighborhood*), e a consulta é feita considerando como vizinhos tanto elementos que estão à direita quanto os que estão à esquerda. Por esse motivo, os testes foram feitos apenas no WebSinC e as sentenças retornadas correspondem ao resultado esperado.

Os testes de nºs 25 a 36 da tabela 4, que correspondem a funções relacionadas à posição do item na sentença (início, fim, ou posição n) também não se aplicam ao *Corpus Search*, uma vez que esta ferramenta não possui a opção de busca por essas funções. Por esse motivo, os testes foram feitos apenas no WebSinC e as sentenças retornadas correspondem ao resultado esperado.

A tabela 6 resume os dados referentes aos testes morfossintáticos realizados e aqui discutidos. Nove funções morfossintáticas foram implementadas no WebSinc, para as quais foram elaborados e executados 39 testes. Quinze destes testes retornaram resultados iguais em ambas as ferramentas, WebSinc e Corpus Search. Quatro testes retornaram resultados diferentes e vinte testes foram feitos apenas no WebSinc, não sendo possíveis de execução no Corpus Search. Os testes com resultados diferentes ocorreram pelos motivos já explanados nesta seção, a saber: a diferença na interpretação da função de vizinhança pelos dois programas e a contagm de *strings* vazias.

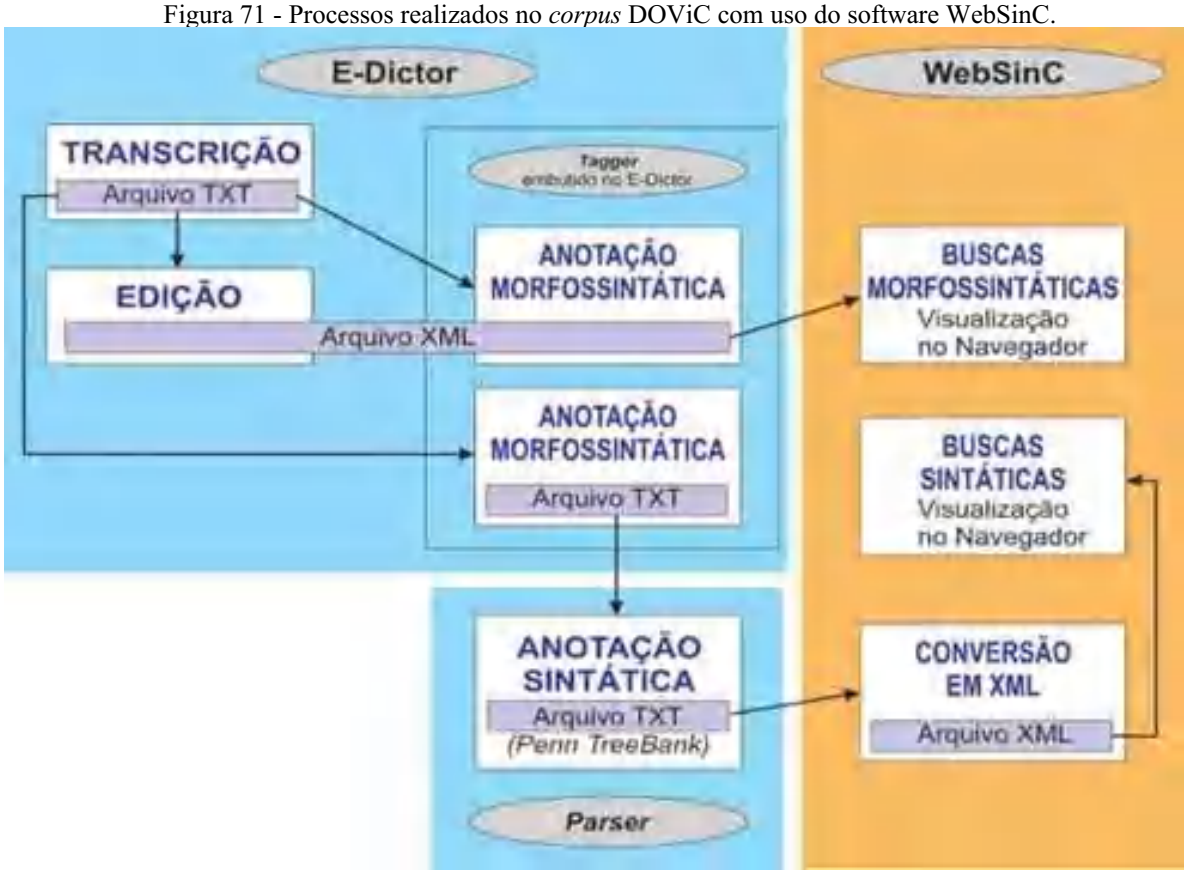
Tabela 6 - Resumo dos resultados dos testes morfossintáticos

Total de funções sintáticas implementadas	9
Número de testes realizados	39
Número de testes com resultados iguais	15
Número de testes com resultados diferentes	4
Número de testes feitos apenas no WebSinc	20

8.6 Mudança no processo de gerenciamento do corpus DOViC

Com o desenvolvimento da ferramenta WebSinC, os arquivos de anotação gerados pelo E-Dictor podem ser utilizados para buscas morfossintáticas. Com a conversão do arquivo de anotação sintática para XML dispensa-se o uso do *Corpus Search* ou outro para buscas sintáticas. O processo de gerenciamento do *corpus* DOViC com o WebSinC passará por mudança, e as atividades relacionadas podem ser visualizadas na figura 71.

Figura 71 - Processos realizados no *corpus* DOViC com uso do software WebSinC.



8.7 Limitações da ferramenta WebSinC

A ferramenta WebSinC tem poder de expressão limitado se comparada a ferramentas de comando textuais. Por se tratar de uma aplicação com interface gráfica, existe uma limitação devido ao tamanho da tela. Teoricamente, infinitos blocos poderiam ser gerados na ferramenta, alinhando-os horizontalmente até o limite de quatro blocos, e depois gerando infinitos conjuntos de quatro blocos no sentido vertical. A barra de rolagem do navegador seria utilizada para visualização dos blocos na tela. No entanto, apesar dessa possibilidade, entendemos que uma grande quantidade de blocos pode tornar a consulta não legível ou compreensível. Na implementação desta pesquisa, a quantidade de blocos por consulta está limitada a seis. Cabe ressaltar que esta limitação se dá pelo nosso entendimento de legibilidade da consulta e não por questões técnicas.

Existem buscas que não podem ser feitas nesta impementação do WebSinC, como por exemplo, buscas com a operação lógica OU entre os blocos. O quadro 21 mostra um exemplo desse tipo de busca, que pode ser feita pelo *Corpus Search*. A expressão busca por sentenças em que haja um NP sujeito (NP-SBJ) que domine 3 palavras OU que haja um NP acusativo (NP-ACC) que domine 4 palavras. No entanto, o resultado desta busca pode ser alcançado no

WebSinC realizando duas buscas separadas, uma por NP sujeito que domine 3 palavras, e outra por NP acusativo que domine 4 palavras, e unindo seus resultados, já que a operação OU corresponde à união dos conjuntos de resultados. De qualquer maneira, cabe ressaltar que esta limitação se deu não por questões técnicas, mas pela restrição do escopo da ferramenta nesta pesquisa, haja vista o horizonte temporal reduzido para o desenvolvimento. A limitação poderá ser eliminada em implementações de versões futuras do WebSinC.

Quadro 21 - Expressão de busca com operação OU entre "blocos".

```
(NP-SBJ domswords 3) OR (NP-ACC domswords 4)
```

Outro tipo de busca que não pode ser realizada nesta versão do WebSinC abrange as buscas que envolvem o mesmo item lexical ou sintagma (uma mesma etiqueta) em mais de uma função na busca, e existe a restrição de que todas as etiquetas idênticas fazem referência ao mesmo item/nó na árvore. O quadro 22 mostra um exemplo desse tipo de busca, que pode ser feita pelo *Corpus Search*. A expressão busca por sentenças em que haja um nome próprio no singular (NPR) que precede um advérbio (ADV) e uma preposição (P) também precede um advérbio (ADV). Como a etiqueta ADV repete-se na expressão de busca, o *Corpus Search* faz a busca como ADV referindo-se ao mesmo advérbio. Sendo assim, nessa busca, o nome próprio e a preposição devem preceder o mesmo advérbio. Se for desejável que dentre os resultados estejam sentenças em que haja este padrão, mas considerando que o advérbio que sucede o nome próprio não é o mesmo advérbio que sucede a preposição, deve-se colocar índices numéricos entre colchetes antes das etiquetas, conforme mostra o quadro 23. No WebSinC, não há como fazer esta distinção, e portanto, são retornados todos os resultados, sem diferenciação do advérbio. Cabe ressaltar, mais uma vez, que a limitação não se deve a questões técnicas, podendo ser eliminada em implementações futuras.

Quadro 22- Expressão de busca com referências à mesma etiqueta.

```
(NPR precedes ADV) AND  (P precedes ADV)
```

Quadro 23 - Expressão de busca com referências a diferentes etiquetas.

```
(NPR precedes [1]ADV) AND  (P precedes [2]ADV)
```

9 CONCLUSÃO E TRABALHOS FUTUROS

Com base nos resultados obtidos neste trabalho, pode-se considerar que o software construído possibilita o gerenciamento e a disponibilização do *corpus* DOViC, permitindo também a realização de buscas automáticas para fins de pesquisa. A interface gráfica construída amplia o leque de usuários uma vez que o uso da ferramenta não demanda o aprendizado de qualquer linguagem de consulta, sistema de anotação ou instalação de algum software pelo usuário pesquisador. A aplicabilidade da ferramenta estende-se a qualquer *corpora* anotado nos mesmos moldes do *corpus* Tycho Brahe e não apenas ao *corpus* DOViC.

Foi possível constatar a corretude dos resultados nos testes das funções de buscas comparando-os com os resultados das buscas executadas em uma ferramenta já testada e utilizada em pesquisas com *corpora* anotados no mundo todo, que é o *Corpus Search*. Os testes das funções de buscas por relações estruturais e exemplos de aplicabilidade da ferramenta em pesquisas linguísticas, como as realizadas por Namiuti (2011), Silveira (2014), Lourençato (2001), Antonelli (2011), entre outros, demonstraram que o WebSinC pode contribuir com pesquisas gramaticais da língua portuguesa e mais especificamente com o avanço de pesquisas em sintaxe.

O uso de XML para anotação sintática evidenciou a vantagem de reutilizar a mesma tecnologia já utilizada para anotações morfológica 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 resultados, e independência tecnológica para grupos de pesquisa interessados em estudo neste *corpus*. O uso da ferramenta *Corpus Search* não foi necessário e dessa maneira, foi possível observar que uma homogeneidade na linguagem de edição, anotação e busca pode contribuir para um maior controle das edições e reutilização de recursos linguísticos.

Como trabalhos futuros sugerimos a implementação de outras funcionalidades na ferramenta WebSinC, especificamente incrementando o poder de expressão nas buscas, com o objetivo de fornecer uma contribuição ainda maior para as pesquisas linguísticas. As funcionalidades que sugerimos são listadas a seguir:

- O aumento da capacidade de buscas da ferramenta, incrementando o número de blocos para a busca gráfica até um limite considerado ainda compreensível pelo usuário.

- O aumento da capacidade de buscas da ferramenta, permitindo a operação lógica OU entre blocos.
- O gerenciamento das buscas realizadas por cada usuário, mantendo-as armazenadas no banco de dados, permitindo que ele mantenha salvas as suas pesquisas e que os administradores do *corpus* possam obter informações a respeito das buscas mais realizadas.
- O destaque com diferentes cores ou diferentes formas para os nós da árvore sintática que correspondem ao padrão pesquisado na busca.

Finalmente, sugerimos o desenvolvimento ou o emprego de um *parser* que faça anotação da estrutura sintática do *corpus* DOViC em XML, sem a necessidade de conversão do formato PTB para este. O produto resultante desta pesquisa não considerou um novo esquema de anotação, como novos nomes de etiquetas, definição de atributos, etc. Assim, um esquema completo de anotação também pode ser projetado, prevendo a sistematização das anotações de edições, morfologia, sintaxe e discurso, baseando-se em padrões existentes mas não deixando de atender às necessidades específicas do *corpus* em questão.

REFERÊNCIAS

ACIOLY, B.M; BEDREGAL, B.R.C. **Introdução à Teoria da Computação**. Linguagens Formais e Computabilidade. 2000.

ALUISIO, M. et al. **The Lacio-Web Project: overview and issues in brazilian portuguese *corpora* creation**. In: *CORPUS* LINGUISTICS 2003, 2003, Lancaster, UK. Proceedings of the *Corpus* Linguistics 2003 Conference: UCREL technical paper number 16. UCREL, Lancaster, UK: Lancaster University, 2003. v. 16.

AMERICAN NATIONAL *CORPUS* (ANC) . **Open Data for language research and education**. 2012. Disponível em:< http://www.anc.org>. Acesso em: 01 nov. 2014.

ANDERSON, S. R. **Where´s morphology**. Linguistic Inquiry, v. 13, 1982.

ANTONELLI, A. **Sintaxe da Posição do Verbo e Mudança Gramatical na História do Português Europeu**. 2011. 248 f. Tese (Doutorado em Linguística) - Universidade Estadual de Campinas, Campinas, 2011.

ASSOCIAÇÃO DAS HUMANIDADES DIGITAIS, 2013. Disponível em: < http://ahdig.org/associacao-das-humanidades-digitais/>. Acesso em: 04 dez. 2014.

BENNET, G. R. **Using *Corpora* in the Language Learning Classroom**: *Corpus* Linguistics for Teachers. Michigan: Michigan ELT, 2010.

BEZERRA, Eduardo. **Princípios de Análise e Projeto de Sistemas com UML**. 2 ed. Rio de Janeiro: Campus, 2007.

BICK, E. **The *parsing* system palavras**: automatic grammatical analysis of portuguese in a constraint grammar framework. 2000. 505 f. Tese (Doutorado em Linguística) - Aarhus University Press, Aarhus, 2000.

BRITISH NATIONAL *CORPUS*. 2009. Disponível em: <http://www.natcorp.ox.ac.uk/>. Acesso em: 05 dez. 2014.

BRITTO, H.; FINGER, M.; GALVES,C. **Computational and linguistic aspects of the construction of the *Tycho Brahe Parsed Corpus of Historical Portuguese.*** São Paulo: Unicamp, 1998. Disponível em:<http://www.tycho.iel.unicamp.br/~tycho/pesquisa/artigos/GALVES_Cetal-Fase1b.pdf> Acesso em: 05 nov. 2014.

BUITELAAR, P. et al. **A Multi-layered, XML-Based Approach to the Integration of Linguistic and Semantic Annotations**. In: PROCEEDINGS OF EACL 2003 WORKSHOP ON LANGUAGE TECHNOLOGY AND THE SEMANTIC WEB. 2003. Disponível em: < http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.8382 >. Acesso em: 05 dez. 2014.

CÂNDIDO JÚNIOR, A.; ALUÍSIO, S.M. **Criação de um ambiente para o processamento de *corpus* de Português Histórico**. USP, 2008. Disponível em:

<http://www.icmc.usp.br/~posgrad/geral/artigos2008/Artigo_Arnaldo_Candido_Junior>. Acesso em: 19 out 2012.

CARROLL, J. ***Parsing***. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

CE-DOHS. ***Corpus*** **Eletrônico de Documentos Históricos do Sertão [CE-DOHS]**. Disponível em: <http://www2.uefs.br/cedohs/apresenta.html 2010>. Acesso em: 02 dez. 2014.

CENTRO DE LINGUÍSTICA DA UNIVERSIDADE DE LISBOA. **CRPC: *Corpus* de Referencia do Português Contemporâneo**. Lisboa, 2014. Disponível em: <http://www.clul.ul.pt/pt/recursos/183-reference-*corpus*-of-contemporary-portuguese-crpc > Acesso em: 4 ago. 2014.

CHOMSKY, N. **Lectures on government and binding**. The Pisa lectures. 7 ed. Berlim; New Yortk: Mouton de Gruyter, 1993.

_____. **Minimalist program**. The MIT Press, 1995. Tradução portuguesa: RAPOSO, E. O programa minimalista. Lisboa: Caminho, 1999.

CORDIAL-SIN. ***Corpus*** **Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)**. 2014. Disponível em:<http://www.clul.ul.pt/pt/recursos/212-cordial-sin-syntax-oriented-*corpus*-of-portuguese-dialects>. Acesso em: 03 dez. 2014.

*CORPUS SEARCH***.** *Corpus Search* **Users Guide**. 2009. Disponível em: <http://*corpus*search.sourceforge.net/CS-manual/Contents.html>. Acesso em: 25 jul. 2013.

DEITEL, H.M.; DEITEL, P.J.; NIETO, T.M.; LIN, T.M.; SHADU, P.V. **XML: Como programar**. Porto Alegre: Bookman, 2005.

DEITEL, H. M. et al. **Perl - Como Programar**. Apresentando CGI e Python. São Paulo: Bookman. 2001.

DEITEL, H.M; DEITEL, P.J. **Java: como programar**. 6.ed. São Paulo: Pearson Prentice Hall, 2005.

DELAMARO, M.E. **Como construir um compilador**. Utilizando ferramentas Java. São Paulo: Novatec, 2004.

ECKART, K. **Aspects of annotations**.In: CLARIN-D User Guide. Universität Stuttgart, 2012. Disponível em: <http://media.dwds.de/clarin/userguide/text/annotation_aspects.xhtml>. Acesso em: 7 ago. 2014.

EDISYN. **EDISYN** Home Page. 2012. Disponível em: < http://www.dialectsyntax.org/wiki/About_Edisyn>. Acesso em: 05 dez 2014.

EISENBACH, A.; EISENBACH, M. **PhpSyntaxTree**: Software para desenho de árvores sintáticas. Disponível em: <http://ironcreek.net/phpsyntaxtree/?>. 2003. Acesso em: 14 out. 2013.

EVANS, D. **Information about *Corpus* building and investigation**: a on-line information pack about *corpus* investigation techniques for the Humanities. Birmingham: Centre for *Corpus* Research/University of Birmingham, 2008. Disponível em: < http://www.birmingham.ac.uk/documents/college-artslaw/*corpus*/intro/unit2.pdf >. Acesso em: 15 jul. 2014.

FINGER, M. **Tagging a morphologically rich language**. In *Proceeding of the first Workshop on Text, Speech and Dialogue (TSD'98)*, pages 39-44, Brno, Czech Republic, 1998.

_____. **Técnicas de otimização da precisão empregadas no etiquetador tycho brahe**. In *V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR2000)*, pages 141-154, Atibaia, Brazil, November 19-22 2000.

FLORIPI, S.A. **Estudo da variação do determinante em sintagmas nominais possessivos do Português Médio ao Português Europeu Moderno.** 2008. 271 f. Tese (Doutorado em Linguística) - Universidade Estadual de Campinas, Campinas, 2008.

FRANCIS, W. N.; KUČERA, H. **Brown *Corpus* manual.** Rhode Island: Department of Linguistics, Brown University, 1979. Disponível em: <http://www.hit.uib.no/icame/brown/bcm.html>. Acesso em: 04 nov. 2014.

GALVES, C. M. C. **Rhythmic Patterns, Parameter Setting and Language Change**. 1998 (Projeto de pesquisa) .

GALVES, C.; BRITTO, H. **A Construção do *Corpus* Anotado do Português Histórico Tycho Brahe** – o sistema de anotação morfológica. 2008. Disponível em: <http://www.tycho.iel.unicamp.br/~tycho/pesquisa/artigos/GALVES_Cetal-Fase1a.pdf>. Acesso em: 5 ago. 2014.

GERBER, R. M.; VASILÉVSKI, V. Um percurso para pesquisas com base em *corpus*. Florianópolis: Editora da UFSC, 2007.

GODOY, M.C. **A colocação dos clíticos no ambiente das orações infinitivas introduzidas por preposições no Português Clássico**. 2006. 53 f. Relatório de Iniciação Científica - Universidade Estadual de Campinas, FAPESB, Campinas, 2006.

GOLDSMITH, J.A. **Segmentation and Morphology**. In: CLARK, A.; FOX, C.; LAPPIN, S. (Editores). The Handbook of Computational Linguistics and Natural Language Processing. Willey-BackWell, 2010.

GOMES DOS SANTOS, C. A. **Complemento-Verbo' vs. 'Verbo-Complemento': uma investigação sobre a estabilização da ordem na diacronia do português**. 2013. 122 f. Dissertação (Mestrado em Linguística) - Universidade Estadual de Campinas, Campinas, 2013.

GONÇALVES, C.A. Iniciação aos estudos morfológicos. Flexão e derivação em português. São Paulo: Contexto, 2011.

GRAVINA, A. P. **Sujeito nulo e ordem VS no português brasileiro: um estudo diacrônico-comparativo baseado em corpus**. 2014. 251 f. Tese (Doutorado em Linguística) - Universidade Estadual de Campinas, Campinas, 2014.

GRISHMAN, R. **TIPSTER Text Architecture Design**. New York University. 1998.

HIRSCHMAN, L.; MANI, I. **Evaluation**. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

HUNSTON, S. **Começando com as palavras pequenas**: Padrões, léxico e sequências semânticas. In: SHEPHERD, T.M.; SARDINHA, T.B.; PINTO, M.V. (Organizadores). Caminhos da Linguística de *Corpus*. Campinas: Mercado de Letras, 2012.

IDE, N. **Encoding Linguistic *Corpora***. In *Proceedings of the Sixth Workshop on Very Large Corpora*, 1998.

IDE, N.; BONHOMME, P.; ROMARY, L. **XCES: An XML-based Encoding Standard for Linguistic *Corpora***. In: INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 2., 2000, Atenas. Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association, 2000.

IDE, N.; ROMARY, L.; CLERGERIE, E. **International Standard for a Linguistic Annotation Framework**. In: WORKSHOP ON SOFTWARE ENGINEERING AND ARCHITECTURE OF LANGUAGE TECHNOLOGY SYSTEMS SEALTS, 2003. Disponível em: < http://clair.eecs.umich.edu/aan/paper.php?paper_id=W03-0804#pdf>. Acesso em: 05 dez. 2014.

JACKSON, P.; MOULINIER, I**. Natural Language Processing for Online Applications**: Text Retrieval, Extraction and Categorization. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2002.

JARGAS, A.M. **Expressões Regulares**. Guia de consulta rápida. São Paulo: Novatec, 2001.

JOHANSSON, S.; STENSTROM, A. (Editores). **English computer corpora**: selected papers and research guide. berlin; New york: Mouton de Gruyter, 1991.

KAPLAN, R.M. **Syntax**. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

KAY, M. **Introduction**. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

KENNEDY, G. **An introduction to *Corpus* linguistics**. London: Longman, 1998.

KEPLER, F. N. ; FINGER, M. **A Part-of-Speech Tagger Based on Variable Length Markov Chains**. In: Concurso de Teses e Dissertações, 2006, Campo Grande, MS. Anais do XXVI Congresso da SBC, 2006.

KÖNIG, E.; LEZIUS, W; VOORMANN, H. **TIGERSearch 2.1. User's Manual.** IMS, University of Stuttgart. 2003. Disponível em: < http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/manual.html>. Acesso em: 05 dez. 2014.

KORTH, H. F.;SILBERSCHATZ, A. ;SUDARSHAN, S. **Sistema de Banco de Dados.** Rio de Janeiro: Elsevier, 2006.

KROCH, A.; TAYLOR, A. **Penn-Helsinki Parsed *Corpus* of Middle English**, second edition. 2000. Disponível em: < http://www.ling.upenn.edu/hist-*corpora*/PPCME2-RELEASE-3/index.html >. Acesso em: 03 dez. 2014.

KROCH, A.; SANTORINI, B.; DIERTANI, A. **Penn-Helsinki Parsed *Corpus* of Early Modern English**. 2004. Disponível em: < http://www.ling.upenn.edu/hist-*corpora*/PPCEME-RELEASE-2/index.html >. Acesso em: 03 dez. 2014.

KROCH, A.; DIERTANI, A. **Penn-Helsinki Parsed *Corpus* of Modern British English**. 2010. Disponível em: < http://www.ling.upenn.edu/hist-*corpora*/PPCMBE-RELEASE-1/index.html >. Acesso em: 03 dez. 2014.

LACIO-WEB. **Compilação de Córpus do Português do Brasil e Implementação de Ferramentas para Análises Linguísticas**. 2004. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/ferramentas.htm> . Acesso em: 4 ago. 2014.

LINGUATECA. **Acesso a corpos de português: Projeto AC/DC**. 2014.
Disponível em: <http://www.linguateca.pt/ACDC/>. Acesso em: 31 jul. 2014.

LOURENÇATO, P.A. **Colocação dos Clíticos em Orações Infinitivas introduzidas por Preposição no Português Clássico**. 2001. 30 f. Relatório de Iniciação Científica - Universidade Estadual de Campinas, FAPESB, Campinas, 2001.

LYONS, J. **Lingua(gem) e Linguística**. Uma introdução. Rio de Janeiro: LTC, 1981.

MAIA, B.; SARMENTO, L. **Corpógrafo - Applications**. In: Third International Workshop on Language Resources for Translation Work Research & Training, Satellite event of LREC 2006 (LR4Trans-III) 28 May 2006, pp. 55-58.

MANNING, C.D.; SCHUTZE, H. **Foundations of Statistical Natural Language Processing**. Massachusetts: The MIT Press, 2000.

MARCUS, M. P.; SANTORINI, B.; MARCINKIEWICZ, M.A. **Building a Large Annotated *Corpus* of English: The *Penn TreeBank***. Computational Linguistics, v.19. 1993.

MARCUS, M.; TAYLOR, A. **The *Penn TreeBank* Project**. Disponível em: <http://www.cis.upenn.edu/~treebank/> 2002. Acesso 14 out. 2013.

McENERY, T. ***Corpus* Linguistics**. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

MENEZES, G. **A Colocação de Clíticos nas Orações Coordenadas do Português Clássico**. 2003. 7 f. Relatório de Iniciação Científica - Universidade Estadual de Campinas, FAPESB, Campinas, 2003.

MENEZES, P. B. **Linguagens Formais e Autômatos**. Porto Alegre: Editora Sagra Luzzato, 2005.

MEGERDOOMIAN, K. **Text mining, *Corpus* building, and testing**. In:FARGHALY, Ali Ahmed Sabry (Ed.). Handbook for language engineers. Standford : CSLI, 2003. pp.14.

MELLO, H.; SOUZA, R. **A linguagem da ciência**: Prospecção de dados baseados em *corpora*. Anais – Seminários Teóricos Interdisciplinares do SEMIOTEC – I STIS. UFMG. 2012. Disponível em: <http://www.periodicos.letras.ufmg.br/index.php/stis/issue/current>. Acesso em: 1 jul. 2014.

MENGEL, A.; LEZIUS, W. **An XML-based representation format for syntactically annotated *corpora***. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=14E13F7984717A2C1EB5E6CB039C4C92?doi=10.1.1.26.6389&rep=rep1&type=pdf>. 2000. Acesso em: 4 ago. 2014.

MIKHEEV, A. **Text Segmentation**. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

MIOTO, C.; SILVA, M.C.F.; LOPES, R. **Novo Manual de Sintaxe**. São Paulo: Contexto, 2013.

MUNIZ, M. et al. **Taming the tiger topic**: **an XCES compliant *corpus* Portal to generate sub*corpus* based on automatic text topic identification**. In: *CORPUS* LINGUISTICS 2007 CONFERENCE, 2007, Birmingham. Proceedings of the *Corpus* Linguistics 2007 Conference. Birmingham: University of Birmingham, 2007. Disponível em: http://ucrel.lancs.ac.uk/publications/CL2007/>. Acesso em: 4 ago. 2014.

NAMIUTI, C. Universidade Estadual de Campinas, Campinas, 2005. Script na linguagem Perl (Código-fonte de software).

_____. **Aspectos da história gramatical do português: interpolação, negação e mudança**. 2008. 331 f. Tese (Doutorado em Linguística) - Universidade Estadual de Campinas, Campinas, 2008.

_____. (Coord.) **Memória Conquistense: implementação de um *corpus* digital**. CNPq 485098/2013-0. UESB, Vitória da Conquista, 2013. (Projeto de Pesquisa).

_____. (Coord.) **Novos meios para antigas fontes: Sintaxe Diacrônica em *corpus* eletrônico do português**. Projeto de Pesquisa. UESB, Vitória da Conquista, 2010.

_____. **Ordem e clíticos: fronteamento e interpolação na diacronia do Português**. In: Anais do VII Congresso Internacional da Abralin, Curitiba 2011. p.923 – 938.

NAMIUTI, C. ; SANTOS, J. V. ; LEITE, C. M. B. **Propostas e Desafios dos Novos Meios das Antigas Fontes: A Preservação da Memória pela Linguística de Corpus**. In: X Colóquio Nacional e II Colóquio Internacional do Museu Pedagógico UESB, 2011, Vitória da Conquista. Anais do X Colóquio Nacional e II Colóquio Internacional do Museu Pedagógico UESB. Vitória da Conquista: UESB, 2011. v. 1. p. 1-11.

NAMIUTI, C. et al. Computação e linguística: importante diálogo para pesquisas e preservação da memória nos novos meios das antigas fontes. **Revista Binacional Brasil Argentina: Diálogo entre as Ciências**, Vitória da Conquista, vol.2, n.1, jul. 2013.

NEDERHOF, M. ; SATTA, A.G. **Theory os *Parsing***. In: CLARK, A.; FOX, C.; LAPPIN, S. (Editores). The Handbook of Computational Linguistics and Natural Language Processing. Willey-BackWell, 2010.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL (NILC). 2014. Disponível em: <http://www.nilc.icmc.usp.br>. Acesso em: 3 ago. 2014.

OTHERO, G.A. Linguística Computacional: Uma breve introdução. **Letras de Hoje**, Porto Alegre v.41, n.2, 2006.

OTHERO, G.A.; MENUZZI, S.M. **Linguística Computacional: teoria & prática**. São Paulo: Parábola Editorial, 2005.

PAIXÃO DE SOUSA, M.C. Memórias do Texto. **Revista Texto Digital**, n.2., 2006. Disponível em: <http://www.textodigital.ufsc.br/num02/paixao.htm>. Acesso em: 5 ago. 2014.

_____. **Sistema de Edições Eletrônicas do *Corpus* Histórico do Português Tycho Brahe**. Fundamentos, Diretrizes e Procedimentos. 2007a. Disponível em: < http://www.tycho.iel.unicamp.br/*corpus*/manual/prep/manual_frameset.html>. Acesso em: 15 nov. 2014.

_____. Digital Text: Conceptual and methodological frontiers. In: ROMERO, D.; SANZ, A. (Org.). **Literatures in the Digital Era: Theory and Praxis**. Cambridge: Cambridge Scholarly, 2007b.

PAIXÃO DE SOUSA, M.C.; KEPLER, F. N.; FARIA, P.P. **E-Dictor**: **novas perspectivas na codificação e edição de *corpora* de textos Históricos**. 2010. In: SHEPHERD, T.M.; SARDINHA, T.B.; PINTO, M.V. (organizadores). Caminhos da Linguística de *Corpus*. Campinas: Mercado de Letras, 2012.

PAIXÃO DE SOUSA, M.C; TRIPPEL, T. **Building a historical *corpus* for Classical Portuguese: some technological aspects**. 2006. Disponível em: <http://www.ime.usp.br/~tycho/participants/psousa/2006/lrec_psousa_trippel.pdf>. Acesso em: 19 out 2012.

PATTERSON, David A. HENNESSY, John L. **Organização e Projeto de Computadores**: A interface hardware/software. Trad.: Daniel Vieira. 3ª Ed. Rio de Janeiro: Elsevier, 2005.

PAUMIER, S. Unitex 3.1 Beta: User Manual. Paris: University of Paris, 2003. Disponível em: < http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf>. Acesso em: jul. 2014.

PEREIRA NETO, A. **PostgreSQL**. Técnicas avançadas: Versões open source: Soluções para desenvolvedores e administradores de Banco de Dados. São Paulo: Editora Érica, 2003.

PINHEIRO, G. M.; ALUISIO, S.M. ***Corpus*** **Nilc**: **descrição e análise crítica com vistas ao projeto Lácio-Web**. São Paulo: USP, 2003. Apresentado no 51º Seminário do Grupo de Estudos Linguísticos do Estado de São Paulo (GEL) em maio 2003, UNITAU/São Paulo. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/downloads/NILC-TR-03-03.zip>. Acesso em: 30 jul. 2014.

PINTO, A.S. **Introdução à utilização do Corpógrafo**: Um pequeno tutorial. 2006. Disponível em: < http://labclup.letras.up.pt/corpografo/docs/tutorial.pdf>. Acesso em: 25 jul. 2014.

PRESSMAN, R. S. **Engenharia de software**. 6 ed. São Paulo: McGraw-Hill, 2006.

PUCSP. **Projeto *Corpus* Brasileiro**. 2014. Disponível em: <http://*corpus*brasileiro.pucsp.br/cb/Inicial.html>. Acesso em: 4 ago. 2014.

RAPOSO, E.P. **Teoria da Gramática à faculdade da Linguagem**. Lisboa: Caminho, 1992.

RESNICK, P.; LIN,J. **Evaluation of NLP Systems**. In: CLARK, A.; FOX, C.; LAPPIN, S. (Editores). The Handbook of Computational Linguistics and Natural Language Processing. Willey-BackWell, 2010.

RIOS, E.; MOREIRA, T. **Teste de software**. 2ª Edição. ed. Rio de Janeiro: Alta Books, 2006.

ROCHA, P. A. ; SANTOS, D. **CETEMPúblico: um *corpus* de grandes dimensões de linguagem jornalística portuguesa**. In: ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DA LINGUA PORTUGUESA ESCRITA E FALADA, 5., 2000, Atibaia, SP. V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000). São Paulo: ICMC/USP, 2000.

RÖGNVALDSSON, E.; INGASON, A.K.; SIGURDSSON, E. **Coping with variation in the icelandic parsed historical *corpus* (ICEPAHC)**. *Language Variation Infrastructure*, Oslo Studies in Language, 2011.

ROHDE, D.L.T. **TGrep2 User Manual**. 2005. Disponível em: <http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf>. Acesso em: 01 dez. 2014.

RUSSEL, Stuart. NORVIG, Peter. **Inteligência Artificial**. 2ª Ed. Rio de Janeiro: Elsevier, 2004.

SANDALO, M.F. Morfologia. In: MUSSALIM, F. BENTES, A.C. **Introdução à linguística**. 9 ed. São paulo: Cortez Editora, 2001.

SANTORINI, B. **Annotation manual for the Penn Historical *Corpora* and the PCEEC**. Disponível em: <http://www.ling.upenn.edu/hist-*corpora*/annotation/index.html>. 2010. Acesso em: 8 out. 2013.

SANTOS, J. V. (Coord.) **Memória Conquistense: recuperação de documentos oitocentistas na implementação de um *corpus* digital**. UESB, Vitória da Conquista, 2009. (Projeto de Pesquisa).

SANTOS, J. V. **Um método de Fotografia técnica documental para formação de *corpora***

digitais de documentos históricos manuscritos. 2013. (No prelo.)

SANTOS, J.V.; BRITO, G. S. **Fotografia técnica de documentos para formação de *corpora* digitais eletrônicos: o método desenvolvido no Lapelinc**. LETRAS & LETRAS, São Paulo, v.30, n.2, 2014, p.421-430.

SANTOS, D. **Disponibilização de *corpora* através da WWW**. In Palmira Marrafa & Maria Antónia Mota (eds.), Linguística Computacional: Investigação Fundamental e Aplicações. Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística (Lisboa, 25-27 de Maio de 1998), Lisboa: Colibri, 1999, pp.323-346.

SARDINHA, T. B. **Linguística de *corpus*: histórico e problemática**. Delta, São Paulo, v.16, n.2, 2000, p.323-367.

_____. **Linguística de *Corpus***. Barueri: Manole, 2004.

_____. **Pesquisa em Lingüística de *Corpus* com WordSmith Tools**. 2006.

SILVA FILHO, A.M. **Programando com XML**. Rio de Janeiro: Elsevier, 2004.

SILVEIRA, D. M. **Clivadas E Pseudo-Clivadas Na Histo ria Do Portugue s: Uma Ana lise Diacro nica Das Estruturas De Foco E Implicac o es Da Grama . tica**. 2014. 273 f. Dissertação (Mestrado em Linguística) - Universidade Estadual de Campinas, Campinas, 2014.

SILVEIRA, F.P. **Integração de ferramentas para compilação e exploração de *corpora***. 2008. 101 f. Dissertação (Mestrado em Ciência da Computação) - Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2008.

SOMMERVILLE, I. **Engenharia de software**. 6 ed. São Paulo: Addison Wesley, 2003.

TANEMBAUM, A.S. **Redes de computadores**. Rio de Janeiro: Elsevier, 2003.

TEXT ENCODING INITIATIVE (TEI). **Text Encoding Initiative**. 2013. Disponível em: <http://www.tei-c.org/index.xml>. Acesso em: 04 nov. 2014.

TRANNIN, J. B. **Aspectos sintáticos do infinitivo com verbos causativos no Português Europeu: uma abordagem diacrônica**. 2010. 144 f. Dissertação (Mestrado em Linguística) - Universidade Estadual de Campinas, Campinas, 2010.

TROST, H. **Morphology**. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

UFRJ. **Para uma história do português do Brasil-RJ**. 2000. Disponível em: <http://www.letras.ufrj.br/phpb-rj/>. Acesso em: 4 ago. 2014.

UNICAMP. **Padrões Rítmicos, Fixação de Parâmetros & Mudança Linguística**. 1998a. Disponível em:<http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/index.html> Acesso em: 31 jul. 2014.

_____. ***Corpus* Histórico Anotado do Português Tycho Brahe**. 1998b. Disponível em: <www.tycho.iel.unicamp.br/~tycho/*corpus*>. Acesso em: 30 jul. 2014.

VAN VALIN JR, R.D. **An Introduction to Syntax**. New York: Cambridge University Press. 2001.

VIEIRA, R.; LIMA, V. L.S. **Lingüística Computacional: princípios e aplicações**. In: Ana Teresa Martins; Díbio Leandro Borges (Org.). SBC - Jornadas de Atualização em Inteligência Artificial (JAIA). Fortaleza, 2001, v. 3, p. 47-86.

VILAÇA, M.L. C. **Pesquisa e ensino: Considerações e reflexões**. Revista e-scrita. Uniabeu, v.1, n.2, 2010.

VOUTILAINEN, A. **Part-of-Speech Tagging**. In: MIKTOV, R. (Editor). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press, 2003.

WALMSLEY, P. **XQuery**. Gravenstein Highway North, Sebastopol: O'Reilly Media, 2007.

W3C. **XML Technology**. 2010. Disponível em: < http://www.w3.org/standards/xml/> Acesso em: 8 out. 2013.

_____. **XQuery**. Disponível em: <http://www.w3.org/XML/Query/> Acesso em: 10 out 2012.

APÊNDICES

APÊNDICE A - EXPRESSÕES XQUERY UTILIZADAS NA IMPLEMETAÇÃO DAS FUNÇÕES DE BUSCA

A expressão XPath que indica o caminho para busca da ocorrência dentro do arquivo XML foi destacada em negrito nos quadros, e os argumentos das funções estão fixos com o intuito de abstrair o uso da linguagem de programação. Em todos os exemplos foi utilizado hipoteticamente o arquivo "arquivo.xml" localizado no diretório "WebContent" do servidor do WebSinC. O quadro 24 mostra toda a consulta XQuery para recuperação das sentenças em buscas sintáticas e nos quadros seguintes, a parte que se repete em todas as consultas foi omitida até o quadro 30, permanecendo apenas a expressão de localização de caminho (restrição na cláusula "where"). Os quadros 41 a 49 mostram as expressões de buscas morfossintáticas. Igualmente, o quadro 41 mostra toda a consulta XQuery para recuperação das sentenças e nos quadros seguintes, a parte que se repete em todas as consultas foi omitida.

Função de busca sintática - Existência

O quadro 24 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças onde existe verbo no gerúndio no arquivo XML (elemento VB-G).

Quadro 24- Expressão XQuery para buscar por sentenças onde existem verbos no gerúndio.

```
for $ipmat in doc('WebContent/arquivo.xml') //DOCUMENTO/(*[starts-
with(name(),'IP-MAT')]|IP-IMP|IP-SUB|FRAG|CP-THT|CONJP)
where exists($ipmat//VB-G)
return string($ipmat)
```

Função de busca sintática - Dominância

O quadro 25 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam sintagmas preposicionais (elementos PP).

Quadro 25 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam sintagmas preposicionais.

```
($ipmat//NP/descendant::PP)
```

Função de busca sintática - Dominância imediata

O quadro 26 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam imediatamente sintagmas preposicionais (elementos PP).

Quadro 26 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam imediatamente sintagmas preposicionais.

```
($ipmat//NP/PP)
```

Função de busca sintática - Irmandade

O quadro 27 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde preposições (elementos P) têm sintagmas nominais (elementos NP) como irmãos na árvore sintática.

Quadro 27 - Expressão XQuery para buscar por sentenças onde preposições têm sintagmas nominais como irmãos na árvore sintática.

```
($ipmat//(P/following-sibling::NP)|(P/preceding-sibling::NP))
```

Função de busca sintática - Dominância imediata como primeiro filho

O quadro 28 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam imediatamente um determinante feminino singular (elementos D-F) como primeiro filho.

Quadro 28 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam imediatamente um determinante definido feminino singular como primeiro filho na árvore sintática.

```
($ipmat//NP/*[1][self::D-F])
```

Função de busca sintática - Dominância imediata como último filho

O quadro 29 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam imediatamente um determinante feminino singular (elementos D-F) como último filho.

Quadro 29 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam imediatamente um determinante definido feminino singular como último filho na árvore sintática.

```
($ipmat//NP/*[last()][self::D-F])
```

Função de busca sintática - Dominância imediata como n-ésimo filho

O quadro 30 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam imediatamente um determinante feminino singular (elementos D-F) como n-ésimo filho. No exemplo do quadro, consideramos n=3, ou seja, a expressão busca por sentenças onde NP têm D-F como terceiro (n=3) filho.

Quadro 30 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam imediatamente um determinante definido feminino singular como terceiro filho na árvore sintática.

```
($ipmat//NP/*[3][self::D-F])
```

Função de busca sintática - Dominância imediata de n filhos

O quadro 31 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam imediatamente N filhos. No exemplo do quadro, consideramos n=3, ou seja, a expressão busca por sentenças onde NP têm exatamente 3 filhos.

Quadro 31 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam imediatamente três filhos na árvore sintática.

```
($ipmat//NP[count(* except (POINT|COMMA))=3])
```

Função de busca sintática - Dominância imediata de menos de n filhos

O quadro 32 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam imediatamente menos de N filhos. No exemplo do quadro, consideramos n=6, ou seja, a expressão busca por sentenças onde NP domina imediatamente menos de 6 filhos.

Quadro 32 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam imediatamente menos de seis filhos na árvore sintática.

```
($ipmat//NP[count(* except (POINT|COMMA))<6])
```

Função de busca sintática - Dominância imediata de mais de n filhos

O quadro 33 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam imediatamente menos de N filhos. No exemplo do quadro, consideramos n=4, ou seja, a expressão busca por sentenças onde NP domina imediatamente mais de 4 filhos.

Quadro 33 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam imediatamente mais de quatro filhos na árvore sintática.

```
($ipmat//NP[count(* except (POINT|COMMA))>4])
```

Função de busca sintática - Dominância imediata como único filho

O quadro 34 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais na função de sujeito (elementos NP-SBJ) dominam imediatamente como único filho um nome próprio no singular (elementos NPR).

Quadro 34 - Expressão XQuery para buscar por sentenças onde sintagmas nominais na função de sujeito dominam imediatamente como único filho um nome próprio no singular na árvore sintática.

```
($ipmat//NP-SBJ/NPR/parent::NP-SBJ[count(*)=1])
```

Função de busca sintática - Dominância de N palavras

O quadro 35 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam N palavras, ou seja, tenham N nós folhas

como descendentes. No exemplo do quadro, consideramos n=3, ou seja, a expressão busca por sentenças onde NP dominam 3 palavras.

Quadro 35 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam três palavras na árvore sintática.

```
($ipmat//NP[count(descendant::LEAF[@W='yes'])=3])
```

Função de busca sintática - Dominância de menos de N palavras

O quadro 36 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam N palavras, ou seja, tenham menos de N nós folhas como descendentes. No exemplo do quadro, consideramos n=6, ou seja, a expressão busca por sentenças onde NP dominam menos de 6 palavras.

Quadro 36 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam menos de seis palavras na árvore sintática.

```
($ipmat//NP[count(descendant::LEAF[@W='yes'])<6])
```

Função de busca sintática - Dominância de mais de N palavras

O quadro 37 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) dominam N palavras, ou seja, tenham mais de N nós folhas como descendentes. No exemplo do quadro, consideramos n=3, ou seja, a expressão busca por sentenças onde NP dominam mais de 3 palavras.

Quadro 37 - Expressão XQuery para buscar por sentenças onde sintagmas nominais dominam mais de três palavras na árvore sintática.

```
($ipmat//NP[count(descendant::LEAF[@W='yes'])>3])
```

Função de busca sintática - Precedência

O quadro 38 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde verbo estar no infinitivo (elementos ET) precedem clítico (elementos CL).

Quadro 38 - Expressão XQuery para buscar por sentenças onde verbo estar no infinitivo precede clítico.

```
(($ipmat//ET/following::ID[1]/LEAF[@v])=($ipmat//ET/following::CL/following::ID[1]/LEAF[@v]))
```

Função de busca sintática - Precedência imediata

O quadro 39 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde uma projeção adverbial (elementos ADJP) precedem imediatamente um nome próprio no singular (elementos NPR).

Quadro 39 - Expressão XQuery para buscar por sentenças uma projeção adverbial precede imediatamente um nome próprio no singular na árvore sintática.

```
((ipmat//ADJP/following-sibling::*[1][self::NPR]) or
(ipmat//ADJP/following-sibling::*[1]/descendant::NPR[position()=1])
```

Função de busca sintática - C-comando

O quadro 40 mostra a expressão XQuery utilizada para realizar uma busca por sentenças onde sintagmas nominais (elementos NP) c-comandam clíticos (elementos CL).

Quadro 40 - Expressão XQuery para buscar por sentenças onde sintagmas nominais c-comandam clíticos na árvore sintática.

```
(ipmat//((NP/following-sibling::CL)|
(NP/preceding-sibling::CL)|
(NP/following-sibling::*/descendant::CL)|
(NP/preceding-sibling::*/descendant::CL)))
```

Função de busca morfossintática - Existência

O quadro 41 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças onde existem verbos no gerúndio no arquivo XML (elementos VB-G).

Quadro 41 - Expressão XQuery para buscar por sentenças onde existem verbos no gerúndio.

```
for $s in doc('WebContent/arquivo.xml')//document/body/text/sc/p/s
let $sentenca:= data($s/w/o)
where ($s/w/m[@v="VB-G"])
return $sentenca
```

Função de busca morfossintática - Palavra no início da sentença

O quadro 42 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças um verbo no gerúndio (elementos VB-G) é a primeira palavra da sentença no arquivo XML.

Quadro 42 - Expressão XQuery para buscar por sentenças onde um verbo no gerúndio é a primeira palavra da sentença.

```
($s/w[1]/m[@v="VB-G"])
```

Função de busca morfossintática - Palavra no fim da sentença

O quadro 43 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças um verbo no gerúndio (elementos VB-G) é a última palavra da sentença no arquivo XML.

Quadro 43 - Expressão XQuery para buscar por sentenças onde um verbo no gerúndio é a última palavra da sentença.

```
($s/w[last()-1]/m[@v="VB-G"])
```

Função de busca morfossintática - Palavra na n-ésima posição da sentença

O quadro 44 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças um verbo no gerúndio (elementos VB-G) é a n-ésima palavra da sentença no arquivo XML. No exemplo do quadro, consideramos n=3, ou seja, a expressão busca por sentenças onde VB-G é a tarceira palavra na sentença.

Quadro 44 - Expressão XQuery para buscar por sentenças onde um verbo no gerúndio é a terceira palavra na sentença.

```
($s/w[3]/m[@v="VB-G"])
```

Função de busca morfossintática - Precedência

O quadro 45 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças um verbo no gerúndio (elementos VB-G) precede um adjetivo feminino no singular (elementos ADJ-F).

Quadro 45 - Expressão XQuery para buscar por sentenças onde um verbo no gerúndio precede um adjeitvo feminino singular.

```
($s/w/ m[@v="VB-G"]/parent::w/following-sibling::w/m[@v="ADJ-F"])
```

Função de busca morfossintática - Precedência imediata

O quadro 46 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças onde um nome (elemento N) precede imediatamente um quantificador (elemento Q).

Quadro 46 - Expressão XQuery para buscar por sentenças onde um nome precede imediatamente um quantificador.

```
($s/w/m[@v='N']/parent::w/following-sibling::w[1]/m[@v='Q'])
```

Função de busca morfossintática - Vizinhança à direita

O quadro 47 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças onde um verbo (elemento VB) possui um adjetivo feminino no singular (elemento ADJ-F) como o n-ésimo vizinho à direita. No exemplo do quadro, consideramos n=1, ou seja, a expressão busca por sentenças onde VB possui um ADJ-F como primeiro vizinho à direita.

Quadro 47 - Expressão XQuery para buscar por sentenças onde um verbo tem um adjetivo feminino singular como primeiro vizinho à direita.

```
($s/w/m[@v='VB']/parent::w/following-sibling::w[1]/m[@v='ADJ-F'])
```

Função de busca morfossintática - Vizinhança à esquerda

O quadro 48 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças onde um verbo (elemento VB) possui um adjetivo feminino no singular (elemento ADJ-F) como o n-ésimo vizinho à esquerda. No exemplo do quadro, consideramos n=2, ou seja, a expressão busca por sentenças onde VB possui um ADJ-F como segundo vizinho à esquerda.

Quadro 48 - Expressão XQuery para buscar por sentenças onde um verbo tem um adjetivo feminino singular como segundo vizinho à esquerda.

```
($s/w/m[@v='VB']/parent::w/preceding-sibling::w[2]/m[@v='ADJ-F'])
```

Função de busca morfossintática - Vizinhança (à esquerda ou à direita)

O quadro 49 mostra a expressão XQuery utilizada pelo WebSinC para realizar uma busca por sentenças onde um verbo (elemento VB) possui um adjetivo feminino no singular (elemento ADJ-F) como o n-ésimo vizinho à esquerda ou à direita. No exemplo do quadro, consideramos n=1, ou seja, a expressão busca por sentenças onde VB possui um ADJ-F como primeiro vizinho à esquerda ou à esquerda.

Quadro 49 - Expressão XQuery para buscar por sentenças onde um verbo tem um adjetivo feminino singular como segundo vizinho à esquerda.

```
($s/w/m[@v='VB']/parent::w/preceding-sibling::w[1]/m[@v='ADJ-F'])|
($s/w/m[@v='VB']/parent::w/following-sibling::w[1]/m[@v='ADJ-F'])
```

ANEXOS

ANEXO A - ARQUIVOS UTILIZADOS NAS BUSCAS MORFOSSINTÁTICAS

Arquivo POS gerado no E-Dictor e editado manualmente para consulta morfossintática no *Corpus Search*

```
#!FORMAT=POS_0

<text>
Eu/PRO Antonio/NPR Jose/NPR de/P Souza/NPR Paes/NPR abaixo/ADV assinado/VB-AN ,/,
sou/SR-P possuidor/N da/P+D-F Cabrinha/NPR Sofia/NPR sem/P embargo/N algum/Q ,/,
e/CONJ por/P que/WPRO é/SR-P minha/PRO$-F vontade/N ,/, e/CONJ lhe/CL tenho/TR-P
grande/ADJ-G amor/N ,/, de/P hoje/ADV em/P diante/ADV lhe/CL confiro/VB-P a/D-F
liberdade/N ,/, e/CONJ fica/N forra/ADJ-F ,/, como/CONJS se/CONJS tal/ADJ-R-G
nascesse/VB-SD :/. podendo/VB-G seguir/VB o/D destino/N ,/, que/WPRO lhe/CL
parecer/VB-SR como/CONJS árbitra/ADJ-F de/P si/PRO mesma/FP ,/, e/CONJ para/P
seu/PRO$ tíitulo/N lhe/CL passo/VB-P a/D-F presente/ADJ-G carta/N por/P mim/PRO
escrita/VB-AN-F ,/, e/CONJ assinada/VB-AN-F ,/, que/CONJ quero/VB-P tenha/TR-SP
validade/N ,/, como/CONJS se/CONJS fosse/SR-SD verba/N de/P título/N ,/, pedindo/N
as/D-F-P Justiças/NPR-P do/P+D Império/NPR lhe/CL deem/VB-SP toda/Q-F a/D-F
validade/N que/WPRO o/D Direito/NPR outorga/NPR ./PONFP

São/NPR Felipo/NPR ./PONFP

cinco/NUM de/P abril/NPR de/P mil/NUM oito/NUM centos/NUM e/CONJ quatro/NUM digo/VB
-P mil/NUM oito/NUM centos/NUM e/CONJ trinta/NUM e/CONJ quatro/NUM =/PONFP

Antonio/NPR José/NPR de/P Souza/NPR Paes/NPR =/PONFP

Reconheço/VB-P verdadeiras/ADJ-F-P e/CONJ dou/VB-P fé/N ./PONFP

Caetité/NPR Caetité/NPR vinte/NUM e/CONJ um/NUM de/P Fevereiro/NPR de/P mil/NUM
oito/NUM centos/NUM e/CONJ trinta/NUM e/CONJ nove/NUM ./PONFP

Brás/NPR de/P Souza/NPR Barrem/NPR Tabelião/NPR a/CL escrevi/VB-D e/CONJ assinei/VB
-D em/P público/ADJ ,/, e/CONJ rogo/VB-P seguintes/ADJ-G-P de/P que/WPRO uso/VB-P
./PONFP

Em/P testemunho/N de/P verdade/N =/. estava/FW o/D signal/N público/ADJ =/PONFP

Brás/NPR de/P Souza/NPR Barrem/NPR =/PONFP

Número/NPR noventa/NUM e/CONJ seis/NUM =/PONFP

Pagou/VB-D do/P+D selo/N oitenta/NUM réis/N-P ./PONFP

Caetité/NPR vinte/NUM e/CONJ um/D-UM de/P Fevereiro/NPR de/P mil/NUM oito/NUM
centos/NUM e/CONJ trinta/NUM e/CONJ nove/NUM =/PONFP

Neves/NPR =/PONFP

Irlanda/NPR Carvalho/VB-D =/PONFP

Lançada/NPR no/P+D livro/N de/P notas/N-P décimo/ADJ quarto/N a/P folhas/N-P
noventa/NUM e/CONJ duas/NUM-F ./PONFP

Caetité/NPR quatro/NUM de/P Abril/NPR de/P mil/NUM oito/NUM centos/NUM e/CONJ
trinta/NUM e/CONJ nove/NUM =/PONFP

Souza/NPR Barrem/NPR =/PONFP
```

```
Não/NEG se/SE continha/VB-D mais/ADV-R outra/OUTRO-F alguma/Q-F coisa/N em/P a/D-F
dita/VB-AN-F carta/N de/P Liberdade/NPR ,/, a/D-F qual/WPRO ,/, sendo/SR-G por/P
mim/PRO Tabelião/NPR abaixo/ADV assinada/VB-AN-F e/CONJ aqui/ADV lançada/VB-AN-F
bem/ADV e/CONJ fielmente/ADV neste/P+D livro/N de/P Notas/NPR-P ,/, e/CONJ a/P
ela/PRO em/P tudo/Q me/CL reportando/VB ,/, e/CONJ depois/ADV de/P com/P
outro/OUTRO oficial/N de/P banca/N comigo/P+PRO ao/P+D concerto/N abaxo/ADV
assinado/VB-AN ,/, lê-la/VB+CL ,/, conferi-la/VB+CL ,/, concertá-la/VB+CL ,/,
escrevê-la/VB+CL e/CONJ assigná-la,/VB+CL foi/SR-D entregue/VB-AN a/P própria/ADJ-F
sorte/N ,/, e/CONJ dou/VB-P fé/N ./PONFP

Imperial/ADJ-G Villa/NPR da/P+D-F Victoria/NPR aos/P+D-P vinte/NUM e/CONJ um/NUM
dias/N-P do/P+D mês/N de/P Outubro/NPR do/P+D ano/N do/P+D Nascimento/NPR de/P
Nosso/PRO$ Senhor/NPR Jesus/NPR Cristo/NPR de/P mil/NUM oito/NUM centos/P+D e/CONJ
quarenta/NUM e/CONJ um/D-UM vigésimo/NPR da/P+D-F Independência/N e/CONJ do/P+D
Império/NPR ./PONFP

Cesario/NPR da/P+D-F Silva/NPR Mello/NPR Tabelião/NPR que/C a/CL escrevi/VB-D ,/,
e/CONJ assinei/VB-D ./PONFP
</text>
```

Arquivo anotado em XML gerado no E-Dictor para consulta morfossintática no WebSinC

```
<?xml version='1.0' encoding='utf-
8'?>
<!-- Para atribuir uma folha de
estilos (XSLT) basta informar o nome
da
     folha no campo 'href' na linha
abaixo -->
<?xml-stylesheet href=""
type="text/xsl"?>
<document>
  <head id="1.29">
    <metadata
generation="original_source">
      <meta>
        <n>1. Genre:</n>
        <v>Carta de Alforria</v>
      </meta>
      <meta>
        <n>2. Author Name:</n>
        <v>Cesario da Silva Melo
(Tabelião)</v>
      </meta>
      <meta>
        <n>3. Author Year of
Birth:</n>
        <v>sd</v>
      </meta>
      <meta>
        <n>4. Original Text
Date:</n>
        <v>1841</v>
      </meta>
      <meta>
        <n>5. Original Text
Editor:</n>
        <v> texto manuscrito</v>
      </meta>
      <meta>
        <n>6. Original Text
Reference:</n>
        <v>Carta de Liberdade da
Cabra de nome Sofia passada por
Antonio Jose de Souza Paes, outrora
Senhor daquela, Livro 1 , folha 101
verso e 102 frente e verso, 1845.
Arquivo do 1° Tabelionato de Notas
Paes (antigo 1° Cartório de Notas do
Forum). Vitória da Conquista, Bahia.
In: SANTOS, Jorge Viana; NAMIUTI-
TEMPONI Cristiane (2014), Corpus de
Documentos Oitocentistas de Vitória
da Conquista - DOViC (versão
BETA)</v>
      </meta>
      <meta>
        <n>7. Original Text
Title:</n>
        <v>Carta de Liberdade da
Cabra de nome Sofia passada por
Antonio Jose de Souza Paes, outrora
Senhor daquela </v>
      </meta>
    </metadata>
    <metadata
generation="edictor_internal">
      <meta>
        <n>Document Name</n>
        <v>1.29</v>
      </meta>
      <meta>
        <n>XML generated by</n>
        <v>E-Dictor-v1.0.b010</v>
      </meta>
      <meta>
        <n>Last Saved Date</n>
        <v>06.05.2014</v>
      </meta>
      <meta>
```

```
        <n>Word Count</n>
        <v>440</v>
      </meta>
    </metadata>
    <metadata
generation="corpus_processing">
      <meta>
        <n>1.Document Title:</n>
        <v>Carta 29</v>
      </meta>
      <meta>
        <n>2. Photograph:</n>
        <v>O corpus Beta DOViC foi
fotografado pelo método Lapelinc
pela seguinte equipe, coordenada
pelo Professor Doutor Jorge Viana
Santos: Giovane Britto; Cecília
ribeiro Souza; Rafael Teixeira;
Silmara de Brito Silva; Ana Paula
dos Reis Couto</v>
      </meta>
      <meta>
        <n>3. Text
Transcription:</n>
        <v>Jorge Viana Santos</v>
      </meta>
      <meta>
        <n>4. Transcription
Revision:</n>
        <v>A Transcrição foi
realizada pela seguinte equipe,
coordenada pela Professora Doutora
Cristiane Namiuti: Cristiane
Namiuti, Paloma Maraisa Oliveira
Carmo, Silmara Brito Silva</v>
      </meta>
      <meta>
        <n>5. Text Edition:</n>
        <v>Silmara de Brito
Silva</v>
      </meta>
      <meta>
        <n>6. Edition Revision:</n>
        <v>Cristiane Namiuti</v>
      </meta>
      <meta>
        <n>7. Morphology
Revision:</n>
        <v>Cristiane Namiuti</v>
      </meta>
      <meta>
        <n>8. Corpus Reference:</n>
        <v>Jorge Viana SANTOS e
Cristiane NAMIUTI-TEMPONI (2014),
Corpus de Documentos Oitocentistas
de Vitória da Conquista - DOViC
Beta.</v>
      </meta>
    </metadata>
  </head>
  <body>
    <text t="full" words="440"
id="text_1" title="Carta de
liberdade da Cabra de nome Sofia"
author="Cesario da Silva e Mello"
year="1841">
      <sc id="sc_1">
        <sce t="header" id="sce_1">
          <te t="pgn" id="te_1">
            <w id="1">
              <o>Livro1: folha 40
verso</o>
            </w>
          </te>
        </sce>
        <p id="p_1" t="title" f="b">
          <s id="s_1">
            <w id="2">
              <o>Carta</o>
              <m v="NPR"/>
            </w>
            <w id="3">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="4">
              <o>liberdade</o>
              <m v="N"/>
            </w>
            <w id="5">
              <o>da</o>
              <m v="P+D-F"/>
            </w>
            <w id="6">
              <o>Cabra</o>
              <m v="NPR"/>
            </w>
            <w id="7">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="8">
              <o>nome<bk t="l"
id="bk_1"/></o>
              <m v="N"/>
            </w>
            <w id="9">
              <o>Sofia</o>
              <m v="NPR"/>
            </w>
            <w id="10">
              <o>passada</o>
              <m v="VB-AN-F"/>
            </w>
            <w id="11">
              <o>por</o>
              <m v="P"/>
            </w>
            <w id="12">
              <o>Antonio</o>
              <m v="NPR"/>
            </w>
            <w id="13">
              <o>Jose</o>
              <e t="mod">José</e>
              <m v="NPR"/>
            </w>
            <w id="14">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="15">
              <o>Souza<bk t="l"
id="bk_2"/></o>
              <m v="NPR"/>
            </w>
```

```
            <w id="16">
              <o>Paes</o>
              <m v="NPR"/>
            </w>
            <w id="17">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="18">
              <o>outrora</o>
              <m v="ADV"/>
            </w>
            <w id="19">
              <o>Senhor</o>
              <m v="NPR"/>
            </w>
            <w id="20">
              <o>daquela</o>
              <e
t="pun">daquela.</e>
              <m v="P+D-F"/>
            </w>
          </s>
        </p>
        <p id="p_2">
          <s id="s_2">
            <w id="21">
              <o>Eu</o>
              <m v="PRO"/>
            </w>
            <w id="22">
              <o>Antonio</o>
              <m v="NPR"/>
            </w>
            <w id="23">
              <o>Jose</o>
              <m v="NPR"/>
            </w>
            <w id="24">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="25">
              <o>Souza</o>
              <m v="NPR"/>
            </w>
            <w id="26">
              <o>Paes</o>
              <m v="NPR"/>
            </w>
            <w id="27">
              <o>abaixo</o>
              <m v="ADV"/>
            </w>
            <w id="28">
              <o>assi-<bk t="l"
id="bk_3"/> gnado</o>
              <e t="jun">assi-
gnado</e>
              <e
t="gra">assignado</e>
              <e
t="mod">assinado</e>
              <m v="VB-AN"/>
            </w>
            <w id="29">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="30">
              <o>sou</o>
              <m v="SR-P"/>
            </w>
            <w id="31">
              <o>possuidor</o>
              <m v="N"/>
            </w>
            <w id="32">
              <o>da</o>
              <m v="P+D-F"/>
            </w>
            <w id="33">
              <o>Cabrinha</o>
              <m v="NPR"/>
            </w>
            <w id="34">
              <o>Sofia<bk t="l"
id="bk_4"/></o>
              <m v="NPR"/>
            </w>
            <w id="35">
              <o>sem</o>
              <m v="P"/>
            </w>
            <w id="36">
              <o>embargo</o>
              <m v="N"/>
            </w>
            <w id="37">
              <o>algum</o>
              <m v="Q"/>
            </w>
            <w id="38">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="39">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="40">
              <o>por</o>
              <m v="P"/>
            </w>
            <w id="41">
              <o>que</o>
              <m v="WPRO"/>
            </w>
            <w id="42">
              <o>he</o>
              <e t="mod">é</e>
              <m v="SR-P"/>
            </w>
            <w id="43">
              <o>minha<bk t="l"
id="bk_5"/></o>
              <m v="PRO$-F"/>
            </w>
            <w id="44">
              <o>vontade</o>
              <m v="N"/>
            </w>
            <w id="45">
              <o>,</o>
              <m v=","/>
            </w>
```

```
            <w id="46">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="47">
              <o>lhe</o>
              <m v="CL"/>
            </w>
            <w id="48">
              <o>tenho</o>
              <m v="TR-P"/>
            </w>
            <w id="49">
              <o>grande</o>
              <m v="ADJ-G"/>
            </w>
            <w id="50">
              <o>amor</o>
              <m v="N"/>
            </w>
            <w id="51">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="52">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="53">
              <o>hoji<bk t="l"
id="bk_6"/></o>
              <e t="mod">hoje</e>
              <m v="ADV"/>
            </w>
            <w id="54">
              <o>em</o>
              <m v="P"/>
            </w>
            <w id="55">
              <o>diante</o>
              <m v="ADV"/>
            </w>
            <w id="56">
              <o>lhe</o>
              <m v="CL"/>
            </w>
            <w id="57">
              <o>confiro</o>
              <m v="VB-P"/>
            </w>
            <w id="58">
              <o>a</o>
              <m v="D-F"/>
            </w>
            <w id="59">
              <o>liberdade</o>
              <m v="N"/>
            </w>
            <w id="60">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="61">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="62">
```

```
              <o>fi<bk t="l"
id="bk_7"/> ca</o>
              <e t="jun">fica</e>
              <m v="N"/>
              <m v="XX"/>
            </w>
            <w id="63">
              <o>forra</o>
              <m v="ADJ-F"/>
            </w>
            <w id="64">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="65">
              <o>como</o>
              <m v="CONJS"/>
            </w>
            <w id="66">
              <o>si</o>
              <e t="mod">se</e>
              <m v="CONJS"/>
            </w>
            <w id="67">
              <o>tal</o>
              <m v="ADJ-R-G"/>
            </w>
            <w id="68">
              <o>nascesse</o>
              <m v="VB-SD"/>
            </w>
            <w id="69">
              <o>:</o>
              <m v="."/>
            </w>
            <w id="70">
              <o>podendo<bk t="l"
id="bk_8"/></o>
              <m v="VB-G"/>
            </w>
            <w id="71">
              <o>seguir</o>
              <m v="VB"/>
            </w>
            <w id="72">
              <o>o</o>
              <m v="D"/>
            </w>
            <w id="73">
              <o>destino</o>
              <m v="N"/>
            </w>
            <w id="74">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="75">
              <o>que</o>
              <m v="WPRO"/>
            </w>
            <w id="76">
              <o>lhe</o>
              <m v="CL"/>
            </w>
            <w id="77">
              <o>parecer</o>
              <m v="VB-SR"/>
            </w>
```

```
            <w id="78">
              <o>como<bk t="l"
id="bk_9"/></o>
              <m v="CONJS"/>
            </w>
            <w id="79">
              <o>arbitra</o>
              <e t="mod">árbitra</e>
              <m v="ADJ-F"/>
            </w>
            <w id="80">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="81">
              <o>si</o>
              <m v="PRO"/>
            </w>
            <w id="82">
              <o>mesma</o>
              <m v="FP"/>
            </w>
            <w id="83">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="84">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="85">
              <o>para</o>
              <m v="P"/>
            </w>
            <w id="86">
              <o>seo</o>
              <e t="mod">seu</e>
              <m v="PRO$"/>
            </w>
            <w id="87">
              <o>titulo<bk t="l"
id="bk_10"/></o>
              <e t="mod">tíitulo</e>
              <m v="N"/>
            </w>
            <w id="88">
              <o>lhe</o>
              <m v="CL"/>
            </w>
            <w id="89">
              <o>passo</o>
              <m v="VB-P"/>
            </w>
            <w id="90">
              <o>a</o>
              <m v="D-F"/>
            </w>
            <w id="91">
              <o>prezente</o>
              <e
t="mod">presente</e>
              <m v="ADJ-G"/>
            </w>
            <w id="92">
              <o>carta</o>
              <m v="N"/>
            </w>
            <w id="93">
              <o>por</o>
              <m v="P"/>
            </w>
            <w id="94">
              <o>mim</o>
              <m v="PRO"/>
            </w>
            <w id="95">
              <o>escri<bk t="l"
id="bk_11"/> pta</o>
              <e
t="jun">escripta</e>
              <e t="mod">escrita</e>
              <m v="VB-AN-F"/>
            </w>
            <w id="96">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="97">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="98">
              <o>assignada</o>
              <e
t="mod">assinada</e>
              <m v="VB-AN-F"/>
            </w>
            <w id="99">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="100">
              <o>que</o>
              <m v="CONJ"/>
            </w>
            <w id="101">
              <o>quero</o>
              <m v="VB-P"/>
            </w>
            <w id="102">
              <o>tenha</o>
              <m v="TR-SP"/>
            </w>
            <w id="103">
              <o>va<bk t="l"
id="bk_12"/> lidade</o>
              <e
t="jun">validade</e>
              <m v="N"/>
              <m v="N"/>
            </w>
            <w id="104">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="105">
              <o>como</o>
              <m v="CONJS"/>
            </w>
            <w id="106">
              <o>si</o>
              <e t="mod">se</e>
              <m v="CONJS"/>
            </w>
            <w id="107">
              <o>fosse</o>
```

```
              <m v="SR-SD"/>
            </w>
            <w id="108">
              <o>verba</o>
              <m v="N"/>
            </w>
            <w id="109">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="110">
              <o>titulo</o>
              <e t="mod">título</e>
              <m v="N"/>
            </w>
            <w id="111">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="112">
              <o>pe<bk t="l"
id="bk_13"/> dindo</o>
              <e t="jun">pedindo</e>
              <m v="N"/>
              <m v="CONJ"/>
            </w>
            <w id="113">
              <o>as</o>
              <m v="D-F-P"/>
            </w>
            <w id="114">
              <o>Justiças</o>
              <m v="NPR-P"/>
            </w>
            <w id="115">
              <o>do</o>
              <m v="P+D"/>
            </w>
            <w id="116">
              <o>Imperio</o>
              <e t="mod">Império</e>
              <m v="NPR"/>
            </w>
            <w id="117">
              <o>lhe</o>
              <m v="CL"/>
            </w>
            <w id="118">
              <o>deem<bk t="l"
id="bk_14"/></o>
              <m v="VB-SP"/>
            </w>
            <w id="119">
              <o>toda</o>
              <m v="Q-F"/>
            </w>
            <w id="120">
              <o>a</o>
              <m v="D-F"/>
            </w>
            <w id="121">
              <o>validade</o>
              <m v="N"/>
            </w>
            <w id="122">
              <o>que</o>
              <m v="WPRO"/>
            </w>
```

```
            <w id="123">
              <o>o</o>
              <m v="D"/>
            </w>
            <w id="124">
              <o>Direito</o>
              <m v="NPR"/>
            </w>
            <w id="125">
              <o>outorga</o>
              <m v="NPR"/>
            </w>
            <w id="126">
              <o>.</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_3">
            <w id="127">
              <o>São<bk t="l"
id="bk_15"/></o>
              <m v="NPR"/>
            </w>
            <w id="128">
              <o>Felipo</o>
              <m v="NPR"/>
            </w>
          </s>
          <s id="s_4">
            <w id="129">
              <o>cinco</o>
              <m v="NUM"/>
            </w>
            <w id="130">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="131">
              <o>abril</o>
              <m v="NPR"/>
            </w>
            <w id="132">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="133">
              <o>mil</o>
              <m v="NUM"/>
            </w>
            <w id="134">
              <o>oito</o>
              <m v="NUM"/>
            </w>
            <w id="135">
              <o>centos</o>
              <m v="NUM"/>
            </w>
            <w id="136">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="137">
              <o>quatro<bk t="l"
id="bk_16"/></o>
              <m v="NUM"/>
            </w>
            <w id="138">
              <o>digo</o>
```

```
              <m v="VB-P"/>
            </w>
            <w id="139">
              <o>mil</o>
              <m v="NUM"/>
            </w>
            <w id="140">
              <o>oito</o>
              <m v="NUM"/>
            </w>
            <w id="141">
              <o>centos</o>
              <m v="NUM"/>
            </w>
            <w id="142">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="143">
              <o>trinta<bk t="l"
id="bk_17"/></o>
              <m v="NUM"/>
            </w>
            <w id="144">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="145">
              <o>quatro</o>
              <m v="NUM"/>
            </w>
            <w id="146">
              <o>=</o>
              <m v="N"/>
            </w>
          </s>
          <s id="s_5" f="i">
            <w id="147">
              <o>Antonio</o>
              <m v="NPR"/>
            </w>
            <w id="148">
              <o>Jose</o>
              <e t="mod">José</e>
              <m v="NPR"/>
            </w>
            <w id="149">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="150">
              <o>Souza</o>
              <m v="NPR"/>
            </w>
            <w id="151">
              <o>Paes<bk t="l"
id="bk_18"/></o>
              <m v="NPR"/>
            </w>
            <w id="152">
              <o>=</o>
              <m v="N"/>
            </w>
          </s>
          <s id="s_6">
            <w id="153">
              <o>Reconheço</o>
              <m v="VB-P"/>
```

```
            </w>
            <w id="154">
              <o>verdadeiras</o>
              <m v="ADJ-F-P"/>
            </w>
            <w id="155">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="156">
              <o>dou</o>
              <m v="VB-P"/>
            </w>
            <w id="157">
              <o>fé</o>
              <m v="N"/>
            </w>
            <w id="158">
              <o>.</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_7">
            <w id="159">
              <o>Caetite<bk t="p"
id="bk_19"><sce t="header"
id="sce_2"><te t="pgn" id="te_2"><w
id="160"><o>Livro1 folha 41
frente</o></w></te></sce></bk></o>
              <e t="mod">Caetité</e>
              <m v="NPR"/>
            </w>
            <w id="161">
              <o>Caetite</o>
              <e t="mod">Caetité</e>
              <m v="NPR"/>
            </w>
            <w id="162">
              <o>vinte</o>
              <m v="NUM"/>
            </w>
            <w id="163">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="164">
              <o>hum</o>
              <e t="mod">um</e>
              <m v="NUM"/>
            </w>
            <w id="165">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="166">
              <o>Fevereiro<bk t="l"
id="bk_20"/></o>
              <m v="NPR"/>
            </w>
            <w id="167">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="168">
              <o>mil</o>
              <m v="NUM"/>
            </w>
            <w id="169">
```

```
              <o>oito</o>
              <m v="NUM"/>
            </w>
            <w id="170">
              <o>centos</o>
              <m v="NUM"/>
            </w>
            <w id="171">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="172">
              <o>trinta</o>
              <m v="NUM"/>
            </w>
            <w id="173">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="174">
              <o>nove<bk t="l"
id="bk_21"/></o>
              <m v="NUM"/>
            </w>
          </s>
          <s id="s_8">
            <w id="175">
              <o>Bras</o>
              <e t="mod">Brás</e>
              <m v="NPR"/>
            </w>
            <w id="176">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="177">
              <o>Souza</o>
              <m v="NPR"/>
            </w>
            <w id="178">
              <o>Barrem</o>
              <m v="NPR"/>
            </w>
            <w id="179">
              <o>Tabelliam</o>
              <e
t="mod">Tabelião</e>
              <m v="NPR"/>
            </w>
            <w id="180">
              <o>a</o>
              <m v="CL"/>
            </w>
            <w id="181">
              <o>escrevi</o>
              <m v="VB-D"/>
            </w>
            <w id="182">
              <o>e<bk t="l"
id="bk_22"/></o>
              <m v="CONJ"/>
            </w>
            <w id="183">
              <o>assignei</o>
              <e t="mod">assinei</e>
              <m v="VB-D"/>
            </w>
            <w id="184">
```

```
              <o>em</o>
              <m v="P"/>
            </w>
            <w id="185">
              <o>publico</o>
              <e t="mod">público</e>
              <m v="ADJ"/>
            </w>
            <w id="186">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="187">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="188">
              <o>rogo</o>
              <m v="VB-P"/>
            </w>
            <w id="189">
              <o>seguintes</o>
              <m v="ADJ-G-P"/>
            </w>
            <w id="190">
              <o>de<bk t="l"
id="bk_23"/></o>
              <m v="P"/>
            </w>
            <w id="191">
              <o>que</o>
              <m v="WPRO"/>
            </w>
            <w id="192">
              <o>uzo</o>
              <e t="mod">uso</e>
              <m v="VB-P"/>
            </w>
            <w id="193">
              <o>.</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_9">
            <w id="194">
              <o>Em</o>
              <m v="P"/>
            </w>
            <w id="195">
              <o>testemunho</o>
              <m v="N"/>
            </w>
            <w id="196">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="197">
              <o>verdade</o>
              <m v="N"/>
            </w>
            <w id="198">
              <o>=</o>
              <m v="."/>
            </w>
            <w id="199">
              <o>es<bk t="l"
id="bk_24"/> tava</o>
              <e t="jun">estava</e>
```

```
              <m v="FW"/>
              <m v="VB-D"/>
            </w>
            <w id="200">
              <o>o</o>
              <m v="D"/>
            </w>
            <w id="201">
              <o>signal</o>
              <m v="N"/>
            </w>
            <w id="202">
              <o>publico</o>
              <e t="mod">público</e>
              <m v="ADJ"/>
            </w>
            <w id="203">
              <o>=</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_10">
            <w id="204">
              <o>Bras</o>
              <e t="mod">Brás</e>
              <m v="NPR"/>
            </w>
            <w id="205">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="206">
              <o>Sou<bk t="l"
id="bk_25"/> za</o>
              <e t="jun">Souza</e>
              <m v="P"/>
              <m v="NPR"/>
            </w>
            <w id="207">
              <o>Barrem</o>
              <m v="NPR"/>
            </w>
            <w id="208">
              <o>=</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_11">
            <w id="209">
              <o>Numero</o>
              <e t="mod">Número</e>
              <m v="NPR"/>
            </w>
            <w id="210">
              <o>noventa</o>
              <m v="NUM"/>
            </w>
            <w id="211">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="212">
              <o>seis</o>
              <m v="NUM"/>
            </w>
            <w id="213">
              <o>=<bk t="l"
id="bk_26"/></o>
```

```
              <m v="."/>
            </w>
          </s>
          <s id="s_12">
            <w id="214">
              <o>Pagou</o>
              <m v="VB-D"/>
            </w>
            <w id="215">
              <o>do</o>
              <m v="P+D"/>
            </w>
            <w id="216">
              <o>sello</o>
              <e t="mod">selo</e>
              <m v="N"/>
            </w>
            <w id="217">
              <o>oitenta</o>
              <m v="NUM"/>
            </w>
            <w id="218">
              <o>reis</o>
              <e t="mod">réis</e>
              <m v="N-P"/>
            </w>
          </s>
          <s id="s_13">
            <w id="219">
              <o>Caetite</o>
              <e t="mod">Caetité</e>
              <m v="NPR"/>
            </w>
            <w id="220">
              <o>vin<bk t="l"
id="bk_27"/> te</o>
              <e t="jun">vinte</e>
              <m v="NUM"/>
              <m v="CL"/>
            </w>
            <w id="221">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="222">
              <o>hum</o>
              <e t="mod">um</e>
              <m v="D-UM"/>
            </w>
            <w id="223">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="224">
              <o>Fevereiro</o>
              <m v="NPR"/>
            </w>
            <w id="225">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="226">
              <o>mil</o>
              <m v="NUM"/>
            </w>
            <w id="227">
              <o>oito</o>
              <m v="NUM"/>
```

```
            </w>
            <w id="228">
              <o>centos<bk t="l"
id="bk_28"/></o>
              <m v="NUM"/>
            </w>
            <w id="229">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="230">
              <o>trinta</o>
              <m v="NUM"/>
            </w>
            <w id="231">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="232">
              <o>nove</o>
              <m v="NUM"/>
            </w>
            <w id="233">
              <o>=</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_14">
            <w id="234">
              <o>Neves</o>
              <m v="NPR"/>
            </w>
            <w id="235">
              <o>=</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_15">
            <w id="236">
              <o>Irlanda</o>
              <m v="NPR"/>
            </w>
            <w id="237">
              <o>Carva<bk t="l"
id="bk_29"/> lho</o>
              <e
t="jun">Carvalho</e>
              <m v="VB-D"/>
              <m v="D"/>
            </w>
            <w id="238">
              <o>=</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_16">
            <w id="239">
              <o>Lançada</o>
              <m v="NPR"/>
            </w>
            <w id="240">
              <o>no</o>
              <m v="P+D"/>
            </w>
            <w id="241">
              <o>livro</o>
              <m v="N"/>
            </w>
```

```
            <w id="242">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="243">
              <o>notas</o>
              <m v="N-P"/>
            </w>
            <w id="244">
              <o>de<bk t="l"
id="bk_30"/> cimo</o>
              <e t="jun">decimo</e>
              <e t="mod">décimo</e>
              <m v="ADJ"/>
            </w>
            <w id="245">
              <o>quarto</o>
              <m v="N"/>
            </w>
            <w id="246">
              <o>a</o>
              <m v="P"/>
            </w>
            <w id="247">
              <o>folhas</o>
              <m v="N-P"/>
            </w>
            <w id="248">
              <o>noventa</o>
              <m v="NUM"/>
            </w>
            <w id="249">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="250">
              <o>duas</o>
              <m v="NUM-F"/>
            </w>
          </s>
          <s id="s_17">
            <w id="251">
              <o>Cae<bk t="l"
id="bk_31"/> tite</o>
              <e t="jun">Caetite</e>
              <e t="mod">Caetité</e>
              <m v="NPR"/>
            </w>
            <w id="252">
              <o>quatro</o>
              <m v="NUM"/>
            </w>
            <w id="253">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="254">
              <o>Abril</o>
              <m v="NPR"/>
            </w>
            <w id="255">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="256">
              <o>mil</o>
              <m v="NUM"/>
            </w>
```

```
            <w id="257">
              <o>oito</o>
              <m v="NUM"/>
            </w>
            <w id="258">
              <o>centos<bk t="l"
id="bk_32"/></o>
              <m v="NUM"/>
            </w>
            <w id="259">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="260">
              <o>trinta</o>
              <m v="NUM"/>
            </w>
            <w id="261">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="262">
              <o>novi</o>
              <e t="mod">nove</e>
              <m v="NUM"/>
            </w>
            <w id="263">
              <o>=</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_18">
            <w id="264">
              <o>Souza</o>
              <m v="NPR"/>
            </w>
            <w id="265">
              <o>Barrem</o>
              <m v="NPR"/>
            </w>
            <w id="266">
              <o>=</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_19">
            <w id="267">
              <o>Não<bk t="l"
id="bk_33"/></o>
              <m v="NEG"/>
            </w>
            <w id="268">
              <o>se</o>
              <m v="SE"/>
            </w>
            <w id="269">
              <o>continha</o>
              <m v="VB-D"/>
            </w>
            <w id="270">
              <o>mais</o>
              <m v="ADV-R"/>
            </w>
            <w id="271">
              <o>outra</o>
              <m v="OUTRO-F"/>
            </w>
            <w id="272">
```

```
              <o>alguma</o>
              <m v="Q-F"/>
            </w>
            <w id="273">
              <o>coi<bk t="l"
id="bk_34"/> za</o>
              <e t="jun">coiza</e>
              <e t="mod">coisa</e>
              <m v="N"/>
            </w>
            <w id="274">
              <o>em</o>
              <m v="P"/>
            </w>
            <w id="275">
              <o>a</o>
              <m v="D-F"/>
            </w>
            <w id="276">
              <o>dita</o>
              <m v="VB-AN-F"/>
            </w>
            <w id="277">
              <o>carta</o>
              <m v="N"/>
            </w>
            <w id="278">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="279">
              <o>Liberdade</o>
              <m v="NPR"/>
            </w>
            <w id="280">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="281">
              <o>a</o>
              <m v="D-F"/>
            </w>
            <w id="282">
              <o>qual</o>
              <m v="WPRO"/>
            </w>
            <w id="283">
              <o>,<bk t="l"
id="bk_35"/></o>
              <m v=","/>
            </w>
            <w id="284">
              <o>sendo</o>
              <m v="SR-G"/>
            </w>
            <w id="285">
              <o>por</o>
              <m v="P"/>
            </w>
            <w id="286">
              <o>mim</o>
              <m v="PRO"/>
            </w>
            <w id="287">
              <o>Tabelliam</o>
              <e
t="mod">Tabelião</e>
              <m v="NPR"/>
```

```
            </w>
            <w id="288">
              <o>abaixo</o>
              <m v="ADV"/>
            </w>
            <w id="289">
              <o>assigna<bk t="l"
id="bk_36"/> da</o>
              <e
t="jun">assignada</e>
              <e
t="mod">assinada</e>
              <m v="VB-AN-F"/>
            </w>
            <w id="290">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="291">
              <o>aqui</o>
              <m v="ADV"/>
            </w>
            <w id="292">
              <o>lançada</o>
              <m v="VB-AN-F"/>
            </w>
            <w id="293">
              <o>bem</o>
              <m v="ADV"/>
            </w>
            <w id="294">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="295">
              <o>fielmente</o>
              <m v="ADV"/>
            </w>
            <w id="296">
              <o>neste<bk t="l"
id="bk_37"/></o>
              <m v="P+D"/>
            </w>
            <w id="297">
              <o>livro</o>
              <m v="N"/>
            </w>
            <w id="298">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="299">
              <o>Nottas</o>
              <e t="mod">Notas</e>
              <m v="NPR-P"/>
            </w>
            <w id="300">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="301">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="302">
              <o>a</o>
              <m v="P"/>
            </w>
```

```
            <w id="303">
              <o>ella</o>
              <e t="mod">ela</e>
              <m v="PRO"/>
            </w>
            <w id="304">
              <o>em</o>
              <m v="P"/>
            </w>
            <w id="305">
              <o>tudo</o>
              <m v="Q"/>
            </w>
            <w id="306">
              <o>me</o>
              <m v="CL"/>
            </w>
            <w id="307">
              <o>repor<bk t="l"
id="bk_38"/> tando</o>
              <e
t="jun">reportando</e>
              <m v="VB"/>
              <m v="N"/>
            </w>
            <w id="308">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="309">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="310">
              <o>depois</o>
              <m v="ADV"/>
            </w>
            <w id="311">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="312">
              <o>com</o>
              <m v="P"/>
            </w>
            <w id="313">
              <o>outro</o>
              <m v="OUTRO"/>
            </w>
            <w id="314">
              <o>official</o>
              <e t="mod">oficial</e>
              <m v="N"/>
            </w>
            <w id="315">
              <o>de<bk t="l"
id="bk_39"/></o>
              <m v="P"/>
            </w>
            <w id="316">
              <o>banca</o>
              <m v="N"/>
            </w>
            <w id="317">
              <o>comigo</o>
              <m v="P+PRO"/>
            </w>
            <w id="318">
```

```
              <o>ao</o>
              <m v="P+D"/>
            </w>
            <w id="319">
              <o>concerto</o>
              <m v="N"/>
            </w>
            <w id="320">
              <o>abaxo</o>
              <m v="ADV"/>
            </w>
            <w id="321">
              <o>assi<bk t="l"
id="bk_40"/> gnado</o>
              <e
t="jun">assignado</e>
              <e
t="mod">assinado</e>
              <m v="VB-AN"/>
            </w>
            <w id="322">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="323">
              <o>lel-a</o>
              <e t="mod">lê-la</e>
              <m v="VB+CL"/>
            </w>
            <w id="324">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="325">
              <o>conferil-a</o>
              <e t="mod">conferi-
la</e>
              <m v="VB+CL"/>
            </w>
            <w id="326">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="327">
              <o>concertal-a</o>
              <e t="mod">concertá-
la</e>
              <m v="VB+CL"/>
            </w>
            <w id="328">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="329">
              <o>es<bk t="l"
id="bk_41"/> crevel-a ,</o>
              <e t="jun">escrevel-
a,</e>
              <e t="mod">escrevê-
la</e>
              <m v="VB+CL"/>
            </w>
            <w id="330">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="331">
              <o>assignal-a ,</o>
```

```
              <e t="jun">assignal-
a,</e>
              <e t="mod">assigná-
la,</e>
              <m v="VB+CL"/>
            </w>
            <w id="332">
              <o>foi</o>
              <m v="SR-D"/>
            </w>
            <w id="333">
              <o>entregue</o>
              <m v="VB-AN"/>
            </w>
            <w id="334">
              <o>a</o>
              <m v="P"/>
            </w>
            <w id="335">
              <o>pro<bk t="l"
id="bk_42"/> pria</o>
              <e t="jun">propria</e>
              <e t="mod">própria</e>
              <m v="ADJ-F"/>
            </w>
            <w id="336">
              <o>sorte</o>
              <m v="N"/>
            </w>
            <w id="337">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="338">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="339">
              <o>dou</o>
              <m v="VB-P"/>
            </w>
            <w id="340">
              <o>fé</o>
              <m v="N"/>
            </w>
            <w id="341">
              <o>.</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_20">
            <w id="342">
              <o>Imperial</o>
              <m v="ADJ-G"/>
            </w>
            <w id="343">
              <o>Villa</o>
              <m v="NPR"/>
            </w>
            <w id="344">
              <o>da<bk t="l"
id="bk_43"/></o>
              <m v="P+D-F"/>
            </w>
            <w id="345">
              <o>Victoria</o>
              <m v="NPR"/>
            </w>
```

```
            <w id="346">
              <o>aos</o>
              <m v="P+D-P"/>
            </w>
            <w id="347">
              <o>vinte</o>
              <m v="NUM"/>
            </w>
            <w id="348">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="349">
              <o>hum</o>
              <e t="mod">um</e>
              <m v="NUM"/>
            </w>
            <w id="350">
              <o>dias</o>
              <m v="N-P"/>
            </w>
            <w id="351">
              <o>do</o>
              <m v="P+D"/>
            </w>
            <w id="352">
              <o>mez</o>
              <e t="mod">mês</e>
              <m v="N"/>
            </w>
            <w id="353">
              <o>de<bk t="l"
id="bk_44"/></o>
              <m v="P"/>
            </w>
            <w id="354">
              <o>Oitubro</o>
              <e t="mod">Outubro</e>
              <m v="NPR"/>
            </w>
            <w id="355">
              <o>do</o>
              <m v="P+D"/>
            </w>
            <w id="356">
              <o>anno</o>
              <e t="mod">ano</e>
              <m v="N"/>
            </w>
            <w id="357">
              <o>do</o>
              <m v="P+D"/>
            </w>
            <w id="358">
              <o>Nascimento</o>
              <m v="NPR"/>
            </w>
            <w id="359">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="360">
              <o>Nos<bk t="l"
id="bk_45"/> so</o>
              <e t="jun">Nosso</e>
              <m v="PRO$"/>
              <m v="N-P"/>
            </w>
```

```
            <w id="361">
              <o>Senhor</o>
              <m v="NPR"/>
            </w>
            <w id="362">
              <o>Jesus</o>
              <m v="NPR"/>
            </w>
            <w id="363">
              <o>Christo</o>
              <e t="mod">Cristo</e>
              <m v="NPR"/>
            </w>
            <w id="364">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="365">
              <o>mil</o>
              <m v="NUM"/>
            </w>
            <w id="366">
              <o>oito</o>
              <m v="NUM"/>
            </w>
            <w id="367">
              <o>cen<bk t="l"
id="bk_46"/> tos</o>
              <e t="jun">centos</e>
              <m v="P+D"/>
              <m v="NUM"/>
            </w>
            <w id="368">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="369">
              <o>quarenta</o>
              <m v="NUM"/>
            </w>
            <w id="370">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="371">
              <o>hum</o>
              <e t="mod">um</e>
              <m v="D-UM"/>
            </w>
            <w id="372">
              <o>vigessimo</o>
              <e
t="mod">vigésimo</e>
              <m v="NPR"/>
            </w>
            <w id="373">
              <o>da</o>
              <m v="P+D-F"/>
            </w>
            <w id="374">
              <o>Inde<bk t="l"
id="bk_47"/> pendencia</o>
              <e
t="jun">Independencia</e>
              <e
t="mod">Independência</e>
              <m v="N"/>
            </w>
```

```
            <w id="375">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="376">
              <o>do</o>
              <m v="P+D"/>
            </w>
            <w id="377">
              <o>Imperio</o>
              <e t="mod">Império</e>
              <m v="NPR"/>
            </w>
            <w id="378">
              <o>.</o>
              <m v="."/>
            </w>
          </s>
          <s id="s_21">
            <w id="379">
              <o>Cesario</o>
              <m v="NPR"/>
            </w>
            <w id="380">
              <o>da</o>
              <m v="P+D-F"/>
            </w>
            <w id="381">
              <o>Sil<bk t="l"
id="bk_48"/> va</o>
              <e t="jun">Silva</e>
              <m v="NPR"/>
              <m v="VB-D"/>
            </w>
            <w id="382">
              <o>Mello</o>
              <m v="NPR"/>
            </w>
            <w id="383">
              <o>Tabelliam</o>
              <e
t="mod">Tabelião</e>
              <m v="NPR"/>
            </w>
            <w id="384">
              <o>que</o>
              <m v="C"/>
            </w>
            <w id="385">
              <o>a</o>
              <m v="CL"/>
            </w>
            <w id="386">
              <o>escrevi</o>
              <m v="VB-D"/>
            </w>
            <w id="387">
              <o>,</o>
              <m v=","/>
            </w>
            <w id="388">
              <o>e</o>
              <m v="CONJ"/>
            </w>
            <w id="389">
              <o>as<bk t="l"
id="bk_49"/> signei</o>
```

```
              <e
t="jun">assignei</e>
              <e t="mod">assinei</e>
              <m v="VB-D"/>
            </w>
          </s>
        </p>
        <p id="p_3" t="signature">
          <s id="s_22">
            <w id="390">
              <o>[</o>
            </w>
            <w id="391">
              <o>........</o>
            </w>
            <w id="392">
              <o>]<bk t="c"
id="bk_50"/></o>
            </w>
          </s>
          <s id="s_23"
t="signature">
            <w id="393">
              <o>[</o>
            </w>
            <w id="394">
              <o>.....</o>
            </w>
            <w id="395">
              <o>]<bk t="l"
id="bk_51"/></o>
            </w>
          </s>
        </p>
        <p id="p_4" t="signature">
          <s id="s_24">
            <w id="396">
              <o>[</o>
            </w>
            <w id="397">
              <o>........</o>
            </w>
            <w id="398">
              <o>]<bk t="c"
id="bk_52"/></o>
            </w>
          </s>
          <s id="s_25"
t="signature">
            <w id="399">
              <o>Cesario</o>
            </w>
            <w id="400">
              <o>da</o>
            </w>
            <w id="401">
              <o>Silva</o>
            </w>
            <w id="402">
              <o>Mello<bk t="l"
id="bk_53"/></o>
            </w>
          </s>
          <s id="s_26"
t="signature">
            <w id="403">
              <o>[</o>
            </w>
```

```
            <w id="404">
              <o>...</o>
            </w>
            <w id="405">
              <o>...</o>
            </w>
            <w id="406">
              <o>...</o>
            </w>
            <w id="407">
              <o>]</o>
            </w>
          </s>
        </p>
        <p id="p_5" t="closing">
          <s id="s_27">
            <w id="408">
              <o>Conta</o>
              <m v="VB-I"/>
            </w>
          </s>
        </p>
        <p id="p_6" t="closing">
          <s id="s_28">
            <w id="409">
              <o>Imp</o>
              <m v="N"/>
            </w>
            <w id="410">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="411">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="412">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="413">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="414">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="415">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="416">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="417">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="418">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="419">
              <o>-</o>
              <m v="("/>
            </w>
```

```
            <w id="420">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="421">
              <o>&lt;Símbolo</o>
              <m v="N"/>
            </w>
            <w id="422">
              <o>de</o>
              <m v="P"/>
            </w>
            <w id="423">
              <o>Real/réis&gt;</o>
              <m v="NPR"/>
            </w>
            <w id="424">
              <o>#</o>
              <m v="NPR"/>
            </w>
            <w id="425">
              <o>552<bk t="l"
id="bk_54"/></o>
              <m v="NUM"/>
            </w>
            <w id="426">
              <o>Cont</o>
              <m v="NPR"/>
            </w>
            <w id="427">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="428">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="429">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="430">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="431">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="432">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="433">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="434">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="435">
              <o>-</o>
              <m v="("/>
            </w>
            <w id="436">
              <o>-</o>
              <m v="("/>
```

```
            </w>
            <w id="437">
              <o>-#</o>
              <m v="N"/>
            </w>
            <w id="438">
              <o>150<bk t="l"
id="bk_55"/></o>
              <m v="NUM"/>
            </w>
            <w id="439">
              <o>R#</o>
              <m v="NPR"/>
            </w>
            <w id="440">
              <o>602</o>
              <m v="NUM"/>
            </w>
          </s>
        </p>
      </sc>
    </text>
  </body>
</document>
```