Reflexões sobre anotação sintática e ferramentas de busca - Uso da linguagem XML para anotação sintática no *corpus* digital DOViC

Reflections on syntactic annotation and search tools - Using the XML for syntactic annotation in digital corpus DOViC

Cristiane Namiuti Temponi*
Aline Silva Costa**

RESUMO: *Penn TreeBank* para anotação sintática no *corpus* digital DOViC, uma vez que esta linguagem já é utilizada para a anotação de edições e de informações morfológicas neste *corpus*. Assim, uma única tecnologia pode ser usada para os diversos tipos de buscas automáticas. Para uma experimentação da anotação sintática com XML, implementamos um programa que faz a conversão do formato *Penn TreeBank* para a linguagem alvo, e foram realizadas algumas pesquisas sintáticas com a linguagem XPath, uma linguagem de consulta para a tecnologia XML. As buscas realizadas foram comparadas com as mesmas buscas feitas na ferramenta *corpus* Search, uma ferramenta específica para o formato *Penn TreeBank*. O uso de XML para todas as representações favorece a criação de recursos padronizados, que podem ser reutilizados, facilitando a extração de dados de *corpora*. A disponibilidade de anotação usando um padrão como XML também oferece independência tecnológica a outros grupos pesquisadores interessados no *corpus*.

PALAVRAS-CHAVE: *Corpus*. XML. XPath. *Penn TreeBank*.

ABSTRACT: This paper makes reflections on the use of XML as an alternative format for *Penn TreeBank* syntactic annotation in digital corpus DOViC, since this language is already used for the annotation of editions and morphological information in this corpus. Thus, a single technology can be used for various types of automatic searches. For a trial of syntactic annotation with XML, we implemented a program that does the conversion of the *Penn TreeBank* format for the target language, and some syntactic research with the XPath language, a query language for XML technology were performed. The queries were compared with the same search queries made in tool corpus Search, a tool for the specific format *TreeBank Penn*. The use of XML for any representations favors the creation of standard features, which can be re-used, facilitating the extraction of data from corpora. The availability of annotation using XML as a standard also offers technological independence to other researchers interested groups in the corpus.

KEYWORDS: Corpus. XML. XPath. *Penn TreeBank*.

1. Introdução

Nos últimos anos, *corpora* cada vez maiores de recursos linguísticos foram desenvolvidos e anotados pelos estudiosos da linguagem. Certos princípios de representação

* Doutora em Linguística pela Universidade Estadual de Campinas (UNICAMP) e professora do Departamento de Estudos Lingüísticos e Literários e do Programa de Pós-Graduação em Linguística da Universidade Estadual do Sudoeste da Bahia (UESB).
** Bacharel em Ciências da Computação pela Universidade Estadual do Sudoeste da Bahia (UESB) e Mestranda do Programa de Pós-Graduação em Linguística da UESB.

têm sido amplamente adotados, como o uso de anotação *stand-off*[1] ou o uso da linguagem XML, e foram feitas várias tentativas para proporcionar mecanismos e formatos de anotação genéricos. Apesar de tais esforços, os formatos de anotação variam consideravelmente para cada recurso linguístico, projeto ou *corpus*, muitas vezes para satisfazer as restrições impostas por determinado software de processamento. Tratando-se especificamente dos recursos sintáticos, existem diversos formatos para representar a estrutura de sentenças em *corpora* linguísticos digitais. Essa variedade de formatos, no entanto, dificulta o acesso aos dados sintáticos, uma vez que cada formato exige tecnologias de processamento específicas. Na análise de línguas naturais, os dados linguísticos podem ser reutilizados, servindo a diversas pesquisas. Mas pela variedade de representações existentes, as ferramentas e aplicações computacionais desenvolvidas são raramente reutilizadas. A comunidade de processamento de linguagem reconhece que a uniformização e interoperabilidade são cada vez mais prementes para permitir o compartilhamento, fusão e comparação de recursos linguísticos (IDE; ROMARY; DE CLERGERIE, 2004).

O *corpus* de Documentos Oitocentistas de Vitória da Conquista – DOViC – utiliza a linguagem XML para anotação de edições e representação da morfologia dos textos que o compõem. O esquema de anotação e ferramenta utilizados são os mesmos utilizados pelo *Corpus Histórico do Português Tycho Brahe*. No entanto, a representação da estrutura sintática no *Tycho Brahe* não é feita utilizando essa mesma linguagem, mas sim um outro formato, o *Penn TreeBank*. Dessa maneira, a extração dos dados sintáticos demanda o uso de uma tecnologia diversa, uma ferramenta que faça buscas em arquivos nesse formato específico. A ferramenta utilizada para este propósito no *Tycho Brahe* é o programa *Corpus Search*.

Este trabalho discute o uso da mesma tecnologia já utilizada na representação da morfologia, a linguagem XML, como uma alternativa ao formato *Penn TreeBank* para anotação sintática. XML é uma linguagem que permite descrever qualquer tipo de dado e é um padrão aberto para interoperabilidade e intercâmbio de informações. A existência de uma ampla variedade de tecnologias para esse padrão permite a criação de recursos padronizados, favorecendo a reutilização tecnológica e facilitando a extração de dados de *corpora*.

[1] Anotação *stand- off* é uma estratégia de anotação em que se mantém os dados anotados em documentos separados dos documentos com os dados originais (IDE; ROMARY; DE CLERGERIE, 2004).

Para experimentação de uma representação de estrutura sintática usand XML foi implementado um programa que realiza a conversão do formato *Penn TreeBank* para a linguagem alvo. Buscas automáticas nesse formato foram realizadas com a linguagem de consulta XPath. As buscas realizadas foram então comparadas com as mesmas buscas feitas na ferramenta *Corpus Search*. Finalmente, foi feita uma análise qualitativa de custo/benefício para uso da linguagem em questão no *corpus* DOViC.

Na seção dois, a linguagem XML será abordada brevemente. A seção três apresenta o *corpus* digital DOViC e sua metodologia de anotação. As seções quatro a sete tratam de padrões de anotação para *corpora*, do formato Penn TreeBank e do uso de XML em anotações sintáticas. A seção oito apresenta sucintamente uma linguagem para consultas em XML, a Xpath. As seções seguintes apresentam a proposta do trabalho, mostrando o resultado do programa implementado para a conversão de formatos e as buscas realizadas em XPath. Por fim, a seção onze faz uma análise qualitativa do custo/benefício para uso da proposta, seguida das considerações finais.

2. 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[2] (*World Wide Web Consortium*) como uma nova alternativa à linguagem HTML (*Hiper Text Markup Language*), linguagem dominante na Web, a XML combina extensibilidade, poder e flexibilidade com a simplicidade exigida pela Web (SILVA FILHO, 2004; DEITEL; DEITEL; NIETO; LIN; SADU, 2005).

Documentos XML são documentos de texto que representam dados de maneira estruturada utilizando um conjunto de *tags*[3] 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; DEITEL; NIETO; LIN; SADU, 2005).

[2] O W3C é uma organização, fundada em 1994, destinada a desenvolver tecnologias interoperantes e de domínio público para a *World Wide Web* (DEITEL et al, 2005).
[3] Os termos *marca*, *elemento* ou *etiqueta* podem ser usados como sinônimo de *tag*.

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; DEITEL; NIETO; LIN; SADU, 2005).

Dados representados por XML são estruturados 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; DEITEL; NIETO; LIN; SADU, 2005).

O quadro 1 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 1: 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; DEITEL; NIETO; LIN; SADU, 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; DEITEL; NIETO; LIN; SADU, 2005)

3. O *corpus* DOViC

O *corpus* DOViC (*corpus* de Documentos Oitocentistas de Vitória da Conquista) é um *corpus* digital de documentos manuscritos do século XIX, desenvolvido no âmbito do projeto "Memória conquistense: implementação de um *corpus* digital"[4] (NAMIUTI, 2013) em parceria com o projeto de pesquisa "Sintaxe diacrônica em *corpus* eletrônico: do português pré-clássico às variantes modernas" (NAMIUTI; SANTOS, 2010). Os documentos manuscritos que compõem o *corpus* estão guardados nos arquivos do Fórum de Vitória da Conquista-Bahia.

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 mesmo esquema de anotação. O *Corpus Tycho Brahe* é um *corpus* digital composto de textos em português escritos por autores nascidos entre 1380 e 1845, desenvolvido na Universidade Estadual de Campinas (UNICAMP). O desenvolvimento deste *corpus* se deu a partir de 1998, no âmbito do Projeto "Padrões Rítmicos, Fixação de Parâmetros e Mudança Linguística" (UNICAMP, 1998).

A transcrição e edição dos textos do *corpus* DOViC são feitos com o auxílio da ferramenta E-Dictor (KEPLER; PAIXÃO DE SOUSA; FARIA, 2010). O texto transcrito é salvo em um arquivo no formato texto simples. Edições como modernização, junção, segmentação e modernização de grafia são feitas por meio da interface gráfica da ferramenta, produzindo como resultado um arquivo anotado na linguagem XML. O software realiza anotação das informações morfológicas dos textos, também no formato XML e ambas as anotações são feitas num único arquivo. Esse esquema de anotação suportado pelo E-Dictor, utilizado tanto no *corpus Tycho Brahe* quanto no DOViC, foi concebido dentro do projeto "Memória dos Texto" (PAIXÃO DE SOUSA, 2006). Como esse processo é feito por meio da interface gráfica, o uso da linguagem XML é transparente para o usuário, ou seja, ele não lida diretamente com essa estrutura.

A anotação de edição é realizada identificando todos os itens acrescentados ao texto pelo editor como elementos <e>, e os itens originais correspondentes como elementos <o>. O tipo de edição é identificado através da propriedade "t" dos elementos <e> e os tipos possíveis são listados na tabela 1. A numeração dos elementos, identificados pela propriedade "id" são atribuídas automaticamente pela ferramenta (UNICAMP, 2007).

[4] Projeto financiado pelo CNPq (CNPq 485098/2013-0).

As anotações de informações morfológicas dos textos também são feitas em XML e mantidas no mesmo arquivo com as edições. A identificação de informação morfológica dá-se pela marcação do item lexical com o elemento <m>. A propriedade "v" marca o valor da categoria lexical. A figura 1 mostra um trecho da anotação gerada pelo E-Dictor para um texto do *corpus* DoVic.

Tabela 1: Tipos de edição possíveis para o *corpus Tycho Brahe* e representação na anotação XML.

Tipo de edição	**Atributo de <e>**	**Exemplo**
uniformização grafemática e de módulo	t="gra"	<e t="gra">serviço</e><o>feruiço</o>
separação ou junção de vocábulos	t="seg"	<e t="seg">que fe</e><o>quefe</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><or> " </or>
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: Unicamp (2007).

```
<text t="full" words="130" id="text_1">
  <sc id="sc_1">
    <p id="p_1">
      <s id="s_1">
        <w id="s_1#0">
          <o>eordeno</o>
          <e t="seg">e ordeno</e>
          <m v="CONJ"/>
          <m v="VB-P"/>
        </w>
        <w id="s_1#1">
          <o>atodos</o>
          <e t="seg">a todos</e>
          <m v="P"/>
          <m v="Q-P"/>
        </w>
        <w id="s_1#2">
          <o>osOfficiaes</o>
          <e t="mod">os oficiais</e>
          <e t="seg">os Officiaes</e>
          <m v="D-P"/>
          <m v="N-P"/>
        </w>
        <w id="s_1#3">
          <o>de</o>
          <m v="P"/>
        </w>
        <w id="s_1#4">
          <o>Justiça</o>
          <m v="NPR"/>
        </w>
        <w id="s_1#5">
          <o>desta</o>
          <m v="P+D-F"/>
        </w>
        <w id="s_1#6">
          <o>sobre</o>
```

Figura 1 - Arquivo XML gerado pelo E-Dictor para um documento do *corpus* DOViC.

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 anotado no formato POS (*Part of Speech*), e gera como saída um arquivo texto no formato *Penn TreeBank*, que será detalhado na seção 5. O treinamento do *parser* foi feito para o português clássico na Universidade da Pensilvânia. Para obtenção da

representação sintática nos textos do *corpus* DOViC, os textos deverão ainda passar pelo mesmo processo de etiquetagem.

4. Padrões para anotações de *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[5], Text Encoding Initiative (TEI)[6], Corpus Encoding Standard (CES), Corpus Encoding Standard for XML (XCES) e padrão ISO TC37/SC4.

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

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 p 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

[5] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html
[6] http://www.tei-c.org/index.xml

flexibilidade para codificadores e anotadores, e ao mesmo tempo permitir e estimular o intercâmbio e reutilização de recursos linguísticos anotados (IDE; ROMARY; DE LA CLERGERIE, 2004).

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).

5. O formato *Penn TreeBank*

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 (LEZIUS; MENGEL, 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 incuídas nessa representação. Outras categorias também são omitidas nesse esquema de anotação, como por exemplo, VP e DP (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. A categoria DP, por exemplo, é omitida porque o custo de incluí-la supera sua utilidade. 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 2. A figura 2 mostra a representação gráfica correspondente a esta mesma estrutura de análise.

Quadro 2 - 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).

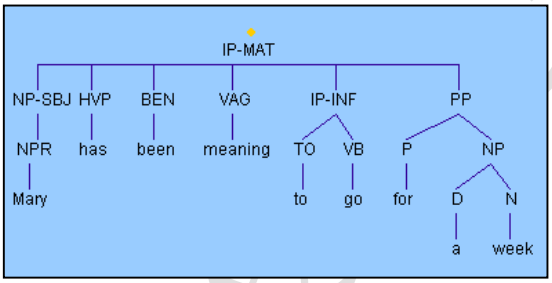

Figura 2 - Representação gráfica de estrutura de análise de uma sentença na anotação *Penn TreeBank*

6. A Ferramenta *Corpus Search*

Assim como há vários formatos para anotação sintática de *corpora*, há também várias ferramentas para extrair informação destes dados anotados, dentre as quais podemos citar: Tgrep2, TIGERsearch, Emu, Corpus Search, NiteQL, Lpath (IMS, 2013).

O *Corpus Search* é um programa que realiza pesquisas sintáticas em *corpora* anotados no formato *Penn TreeBank*. Asssim como o esquema de anotação, o software também foi desenvolvido na Universidade da Pensilvânia (CORPUS SEARCH, 2009).

Corpus Search é implementado na linguagem de programação Java, e portanto, é multiplataforma[7] e requer que o programa JRE[8] (*Java Runtime Environment*) esteja instalado no computador do usuário.

[7] Um programa multiplataforma pode ser executado em qualquer sistema operacional, desde que haja uma máquina virtual apropriada instalada.
[8] JRE (*Java Runtime Environment)* é um software desenvolvido para executar programas feitos na linguagem Java. O JRE possui como componente a máquina virtual Java (JVM- Java Virtual Machine) (ORACLE, 2010).

A execução do *Corpus Search* para realizar buscas sintáticas requer duas entradas: o arquivo do *corpus*, anotado no formato *Penn TreeBank*; e o arquivo com a especificação da consulta a ser realizada, também chamado de *command file*, em formato texto simples.

A especificação das buscas no arquivo de entrada 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 resultados de uma busca realizada pelo *Corpus Search* podem ser vistos no arquivo de saída gerado pelo programa. O arquivo é produzido no formato texto simples e reúne informações sobre as sentenças contendo as restrições especificadas pela busca (CORPUS SEARCH, 2009).

7. Utilização da linguagem XML na anotação sintática

Este trabalho discute a utilização da linguagem XML como alternativa para anotação sintática de textos do *corpus* DOViC. Existem numerosos exemplos da implementação de XML em anotações de *corpora*. Entre eles estão os projetos Alpino Dependency *TreeBank*, Europarl Parallel Corpus, Wikipedia XML *corpora*, PDTB XML, e outros. Há ainda outros estudos que visam converter dados de *corpora* em XML ou desenvolver representações XML para unir os dados de *corpora* de múltiplas fontes. O *Expert Advisory Group on Language Engineering Standards* lançou uma codificação XML padrão para *corpus*, o XCES (*XML Corpus Encoding Standard*) (YAO; BORISOVA; ALAM, 2010).

Lezius e Mengel (2000) propõem um esquema de anotação sintática baseado em XML. Nessa abordagem, é proposta a utilização de basicamente quatro elementos XMLpara descrever a estrutura arbórea: elementos sentença <s> , elementos não-terminais <n>, elementos terminais ou palavras <w> e elementos de aresta <edge> , usado para nós de ligação. As categorias dos nós, como categoria sintática ou rótulo POS são representados como atributos das *tags* XML. Um exemplo da anotação proposta é mostrado na figura 3.

O padrão XCES (seção 4) descreve um padrão de codificação em XML para anotações linguísticas com informações morfossintáticas. Assim, informações sobre estrutura de sentenças e informações morfológicas são mantidas 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 4 mostra um fragmento de um texto contendo anotações de informação morfossintática neste padrão.

```
<s id="s1" href="#id(n1_500)"/>

<n id="n1_500" cat="S">
  <edge id="edge1_1" href="#id(n1_501)"/>
  <edge id="edge1_2" href="#id(n1_502)"/>
</n>
<n id="n1_501" cat="NP">
  <edge id="edge1_3" href="#id(w1_0)"/>
  <edge id="edge1_4" href="#id(w1_1)"/>
</n>
<n id="n1_502" cat="VP">
  <edge id="edge1_5" href="#id(w1_2)"/>
  <edge id="edge1_6" href="#id(n1_503)"/>
</n>
<n id="n1_503" cat="NP">
  <edge id="edge1_7" href="#id(w1_3)"/>
  <edge id="edge1_8" href="#id(w1_4)"/>
</n>

<w id="w1_0" word="The"/>
<w id="w1_1" word="boy"/>
<w id="w1_2" word="likes"/>
<w id="w1_3" word="the"/>
<w id="w1_4" word="girl"/>
```

Figura 3 - Exemplo de anotação sintática usando XML proposta por Lezius e Mengel (2000).
Fonte: Lezius; Mengel (2000).

O projeto MuchMore (seção 4) propõe um formato de anotação baseado em XML onde diversos níveis de informação podem ser mantidos separadamente mas integrados através de identificadores. A figura 5 exemplifica um trecho de texto anotado no formato do MuchMore. O texto é representado pelo elemento <text>, que por sua vez, é composto de um ou mais elementos <token>, que identificam as palavras, os quais marcam através de atributos as informações morfossintáticas, além da forma canônica de cada palavra. As estruturas sintáticas são representadas pelos elementos <chunk> , cujos atributos "from" e "to" marcam onde começa e onde termina a estrutura. O atributo "type" classifica a estrutura como NP, PP, etc. Os atributos "id" dos elementos permitem a referência para a integração de múltiplos níveis de informação linguística.

Yao, Alam e Borisova (2010) apresentam o projeto PDTB XML, um projeto que converte os textos do *corpus Penn* Discourse *TreeBank* 2.0 para o formato XML. O PDTB é um grande *corpus* construído na Universidade da Pensilvânia, anotado com informações sintáticas e relações de discurso, argumentos, atribuições e sentido. O esquema de anotação utilizado possui o mesmo nome do *corpus*, *Penn* Discourse *TreeBank*.

```
<?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>Pp3ns</msd>
        <ctag>PPER3</ctag></lex>
      <lex>
        <base>it</base>
        <msd>Pp3ns</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>Vmis3s</msd>
       <ctag>PAST3</ctag></lex>
     <lex>
       <base>be</base>
       <msd>Vais1s</msd>
       <ctag>AUX1</ctag></lex>
     <lex>
   <base>be</base>
   <msd>Vais3s</msd>
   <ctag>AUX3</ctag></lex>
<lex>
   <base>be</base>
   <msd>Vmis1s</msd>
   <ctag>PAST1</ctag></lex>
<lex>
   <base>be</base>
   <msd>Vmis3s</msd>
   <ctag>PAST3</ctag></lex></tok>…
```

Figura 4 - Fragmento de texto com anotação no padrão XCES.
Fonte: Ide; Bonhomme; Romary (2000).

```
Balint syndrom is a combination of symptoms including simultanagnosia, a
disorder of spatial and object-based attention, disturbed spatial percep-
tion and representation, and optic ataxia resulting from bilateral pa-
rieto-occipital lesions.

<text>
  <token id="w1" pos="NN">Balint</token>
  <token id="w2" pos="NN">syndrom</token>
  <token id="w3" pos="VBZ" lemma="be">is</token>
  <token id="w4" pos="DT" lemma="a">a</token>
  <token id="w5" pos="NN" lemma="combination">combination</token>
  ...
  <token id="w20" pos="JJ" lemma="spatial">spatial</token>
  <token id="w21" pos="NN" lemma="perception">perception</token>
  <token id="w22" pos="CC" lemma="and">and</token>
  <token id="w23" pos="NN" lemma="representation">representation</token>
  ...
</text>

<chunks>
      <chunk id="c1" from="w1" to="w2" type="NP"/>
      <chunk id="c7" from="w20" to="w23" type="NP"/>
</chunks>
```

Figura 5 – Exemplo de anotações linguísticas no MUCHMORE.
Fonte: Buitelaar et al. (2003).

8. XPath

A linguagem XML descreve dados de forma flexível e eficiente através da marcação dos dados com *tags* 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 3 a 5. Todos os exemplos podem ser aplicados ao documento XML dado como exemplo na figura 1. A expressão na figura 8 localiza todos os nós <titulo>, que sejam filhos de <livro>. A expressão na figura 9 localiza o nó <livro> que possua um atributo ISBN cujo valor seja "978-85-7244-800-0". E por fim, a expressão na figura 10 localiza o terceiro nó filho <autor> do nó <livro>.

Quadro 3 - Exemplo de expressão XPath para localizar nós <titulo> filhos de <livro>

```
/livro/titulo
```

Quadro 4 - 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 5 - Exemplo de expressão XPath para localizar o terceiro nó <autor> filho de nós <titulo>

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

9. Conversor do formato *Penn TreeBank* para XML

Para a transformação do formato *Penn TreeBank* para XML, foi desenvolvido neste trabalho um programa na linguagem Java, que recebe como entrada um arquivo no primeiro formato e gera um arquivo de saída XML correspondente. O programa não implementa ainda a função de *parser*, e portanto, o arquivo de entrada deve ser um documento *Penn TreeBank* bem formado. Para uso futuro, o programa deve implementar a função de *parser* a fim de evitar entradas errôneas. O programa não possui interface gráfica, exibindo apenas uma janela de diálogo para fornecimento do arquivo de entrada pelo usuário.

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 *tags* <POINT>. Nós com o símbolo "," foram mapeados para *tags* <COMMA>. Houve a necessidade de substituir o caracter "$" pelo caracter "S". Assim, rótulos como "PRO$" foram mapeados para etiquetas "PROS". Os demais nomes das tags para o documento XML permaneceram os mesmos utilizados no formato *Penn* Tree Bank. Assim, cada nó do arquivo de entrada é mapeado numa tag XML com mesmo nome. Cada nó folha (nó sem filho) é gerado na saída como texto puro entre as tags.

O quadro 6 mostra um trecho de um arquivo do *corpus Tycho Brahe* com anotação sintática *Penn TreeBank* e o quadro 7 mostra o arquivo saída correspondente em XML gerado pelo programa.

Quadro 6 - Trecho de arquivo do *corpus Tycho Brahe* com anotação *Penn TreeBank*

```
( (IP-MAT (NP-SBJ *pro*)
          (VB-R Darei)
          (NP-ACC (N princípio))
          (PP (P a)
             (NP (D-F-P estas) (PRO$-F-P minhas) (N-P memórias)))
          (RRC (P por)
              (NP (D-F a) (PRO$-F minha) (N genealogia)))
          (. .)) (ID A_003_PSD,03.1))
```

Quadro 7 - Trecho de arquivo com anotação sintática em XML gerado pelo programa conversor

```
<IP-MAT>
<NP-SBJ>
*pro*
</NP-SBJ>
<VB-R>
Darei
</VB-R>
<NP-ACC>
<N>
princípio
</N>
</NP-ACC>
<PP>
<P>
a
```

</P>
<NP>
<D-F-P>
estas
</D-F-P>
<PROS-F-P>
minhas
</PROS-F-P>
<N-P>
memórias
</N-P>
</NP>
</PP>
<RRC>
<P>
por
</P>
<NP>
<D-F>
a
</D-F>
<PROS-F>
minha
</PROS-F>
<N>
genealogia
</N>
</NP>
</RRC>
<POINT>
.
</POINT>
</IP-MAT>
<ID>
A_003_PSD,03.1
</ID>

A hierarquia na estrutura gerada pode ser melhor visualizada usando qualquer ferramenta que represente as relações hierárquicas inserindo tabulações, como navegadores e outros. A figura 6 mostra a visualização do documento no navegador Firefox.

10. Buscas sintáticas utilizando anotação XML

As buscas nos arquivos de anotação sintática com XML podem ser feitas utilizando uma linguagem de consulta para esta linguagem. Neste trabalho, a linguagem XPath foi utilizada. Como exemplo, foram realizadas duas buscas em um arquivo com anotação sintática *Penn TreeBank* do *corpus Tycho Brahe*, envolvendo relações de dominância ou maternidade e irmandade. Depois de convertido o arquivo para a anotação XML proposta, buscas equivalentes foram realizadas dentro deste arquivo XML com a linguagem XPath.

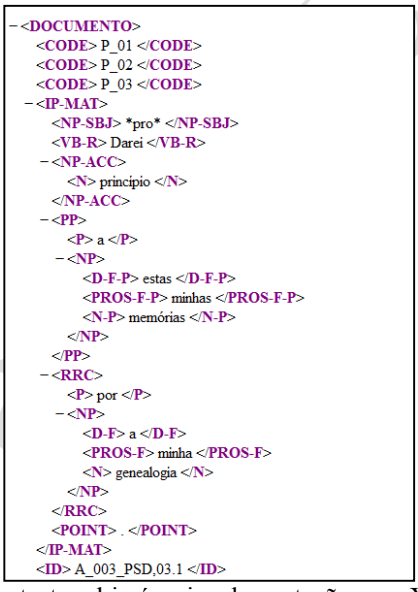

Figura 6 - Visualização da estrutura hierárquica de anotação com XML no navegador Firefox

Para as buscas com XPath, foi implementado um segundo programa na linguagem Java, utilizando a API[9] (*Application Programming Interface*) para XPath. Além da implementação e utilização deste programa, as mesmas buscas também foram feitas no navegador FireFox, através da instalação do *plugin XPath Checker*, disponível gratuitamente

[9] API (Application Programming Interface) é um conjunto de funções e padrões estabelecidos por um software para a utilização das suas funcionalidades por aplicativos que não pretendem conhecer detalhes da implementação do software, mas apenas em usar seus serviços.

na página de complementos para este navegador. Outros programas editores de XML com processadores XML integrados também estão disponíveis na Internet.

A busca para relação de dominância no *Corpus Search* utiliza a função *dominates*. Para realizar uma busca de nós NP que dominam nós PP utilizando essa ferramenta, a seguinte expressão foi utilizada no arquivo de consulta:

(NP dominates PP) (10.1)

Para realizar a busca equivalente em XPath, a expressão de consulta utilizada foi:

(//NP[PP]) (10.2)

Para pesquisas com relação de irmandade, o *Corpus Search* utiliza a função *hasSister*. Para realizar uma busca de nós P que possuem nós irmãos NP utilizando esta ferramenta, a seguinte expressão foi utilizada no arquivo de consulta:

(P hasSister NP) (10.3)

A busca equivalente na linguagem XPath foi feita da seguinte maneira:

//P/following-sibling::NP|NP/following-sibling::P (10.4)

11. Análise do uso de XML/XPath para anotação e buscas sintáticas

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.

Para avaliar o uso do *Corpus Search* e da linguagem XPath como ferramentas de busca, a análise pode ser feita sob diversas perspectivas, tanto nos aspectos tecnológicos, quanto na utilização por usuários finais.

Considerando linguistas como usuários finais de tais ferramentas[10], o *Corpus Search* possui uma linguagem mais simples e mais fácil de aprender que XPath. Como o *Corpus Search* é específico para buscas sintáticas, os comandos foram projetados para este fim, trazendo assim mais simplicidade se comparado à XPath. Para pesquisar uma relação de C-comando, por exemplo, o *Corpus Search* possui a função *ccomands*. Na linguagem XPath seria necessária a combinação de várias expressões utilizando-se de operadores para realizar a busca equivalente. A linguagem XPath não é destinada ao uso por usuários leigos em programação de computadores. Ainda assim, o uso do *Corpus Search* também requer aprendizado de sua linguagem específica, além de requerer que o usuário final saiba trabalhar com linhas de comando, instalação e configuração da máquina virtual Java (JVM), não podendo ser um usuário totalmente leigo.

Se considerarmos a existência de uma aplicação intermediária que forneça uma interface para realização das buscas, a tecnologia utilizada para o usuário torna-se transparente, uma vez que não terá conhecimento do que está realmente sendo utilizado na busca, se *Corpus Search*, XPath ou qualquer outra tecnologia. A comparação neste caso, deverá ser feita apenas tratando aspectos tecnológicos sob o ponto de vista do desenvolvedor da aplicação, como facilidade de implementação e esforço exigido de programação.

Em se tratando dos aspectos tecnológicos, o uso do *Corpus Search* traz a vantagem de trazer no arquivo de saída os resultados das buscas, com os trechos encontrados e as estatísticas. Para buscas em XPath, os conjunto dos nós encontrados também é retornado, mas é preciso que o desenvolvedor da aplicação implemente um tratamento deste resultado para exibi-lo para o usuário final. De qualquer sorte, isto pode ser feito sem um exagerado esforço de programação uma vez que as APIs para XML e XPath da linguagem Java fornecem várias funções para isso. Se for desejável não mostrar para os usuários finais o arquivo de saída do *Corpus Search* tal como é exibido, também será exigido um esforço de programação para tratá-lo a fim de apresentar as informações de outra maneira. Se as buscas são feitas em XML, haverá maior flexibilidade ao desenvolvimento da aplicação para exibição dos resultados da consulta. No *Corpus Search*, os resultados são restritos ao que é trazido no arquivo de saída.

[10] Usuários finais usam a ferramenta diretamente, sem a existência de um outro programa que disponibilize uma interface para facilitar o uso

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 morfológicas. Como a anotação morfológica 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 *Corpus Search*.

A disponibilidade de uma vasta gama de implementações para XML torna o acesso a este formalismo mais fácil. Além de XPath, existem outras linguagens de busca, como XQuery, com mais poder e flexibilidade. Muitos SGBDs (Sistemas Gerenciadores de Bancos de Dados) já implementam o suporte a XML e fornecem um mecanismo de processamento das linguagens de consulta, como o banco de dados PostgreSQL.

As linguagens de consulta não são o único mecanismo de recuperação de dados num documento XML. As buscas podem ser realizadas utilizando-se apenas das APIs para XML, que estão disponíveis em diversas linguagens de programação, com variadas funções para navegação na estrutura arbórea do arquivo XML. Há também ferramentas de visualização que proporcionam uma visão geral da estrutura. O suporte a XML implementado pelos navegadores em combinação com folhas de estilo poderão ser utilizados para uma exibição customizada da estrutura sintática.

Outra vantagem importante de XML como formalismo de codificação para anotação sintática é que a marcação XML é altamente expansível. Isto significa que diferentes níveis de anotação podem ser combinados, como por exemplo o discurso e a sintaxe. O uso de XML na anotação sintática através da conversão *Penn TreeBank* para XML pode dispensar o uso do *Corpus Search*, mas ainda mantém a dependência do *parser* que gera o arquivo PTB. Como a XML já tem sido usada por diversos *corpora*, projetos futuros podem considerar o desenvolvimento de um *parser* que já produza a estrutura sintática em XML. De qualquer sorte, o uso de XML para todas as representações favorece a criação de recursos padronizados, que podem ser reutilizados, facilitando assim a extração de dados de *corpora*. A disponibilidade de anotação usando um padrão como XML se faz importante porque também oferece independência tecnológica a outros grupos pesquisadores interessados no *corpus* DOViC.

12. Conclusão

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*.

Trabalhos futuros poderão considerar o desenvolvimento ou emprego de um *parser* que faça anotação da estrutura sintática em XML, sem a necessidade de conversão. Este trabalho 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 desenvolvido, 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

BUITELAAR, P.; DECLERCK, T.; SACALEANU, B.; VINTAR, S.; RAILEANU, D.; CRISPI, C. A multilayered, XML-based approach to the integration of linguistic and semantic annotations. In: **EACL 2003 Workshop on language technology and the semantic web (NLPXML'03)**, 2003, Budapeste. Proceedings of EACL 2003 Workshop on Language Technology and the Semantic Web (NLPXML'03). Cunningham: EACL, 2003. Disponivel em: <http://www.dfki.de/dfkibib/publications/docs/eacl03-xmlnlp.ps>. Acesso em: 23 set 2014.

CORPUS SEARCH**.** *Corpus Search* **Users Guide**. 2009. Disponível em: <http://corpussearch.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.

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.; DE LA CLERGERIE, E.. International standard for a linguistic annotation framework. **Journal of Natural Language Engineering**, Cambridge, v. 10 n. 3-4, pp. 307-326, Sept. 2004.

IMS (Institut für Maschinelle Sprachverarbeitung). **The TIGER-XML treebank encoding format.** Disponível em: <http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/TigerXML.html>. [ca. 2002]. Acesso em: 08 out 2013.

LEZIUS, W.; MENGEL, A. **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: 08 out 2013.

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

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

NAMIUTI, C. (Coord.); SANTOS, J. V. (Co-coordenador) **Memória Conquistense:** implementação de um corpus digital. CNPq 485098/2013-0. UESB, Vitória da Conquista, 2013. (Projeto de Pesquisa).

ORACLE. **JDK 1.1 for Solaris Developer's Guide**. Java Programming Environment and the Java Runtime Environment (JRE). 2010. Disponível em: <http://docs.oracle.com/cd/E19455-01/806-3461/6jck06gqd/index.html>.Acesso em: 14 out. 2013.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. P. E-Dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. In: Tania Shepherd; Tony Berber Sardinha; Marcia Veirano Pinto. (Org.). **Caminhos da Linguística de Corpus**. Campinas: Mercado de Letras, 2010.

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: 01 out 2012.

SANTORINI, B. **Annotation manual for the Penn Historical *Corpora* and the PCEEC**. Disponível em: <http://www.ling.u*Penn*.edu/hist-*corpora*/annotation/index.html>. 2010. Acesso em: 08 out 2013.

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

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

_____. **Corpus do Português Histórico Tycho Brahe.** Manual de Preparação dos Textos. Sistema de Edições Eletrônicas do corpus *Tycho Brahe*. Campinas, 2007. Disponível em: < http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/prep/manual_frameset.html>. Acesso em: 5 ago 2014.

W3C. **XML Technology**. 2010. Disponível em: < http://www.w3.org/standards/xml/> Acesso em 08 de outubro de 2013.

W3C. **XML Path Language (**XPath**)**. 1999. Disponível em: <http://www.w3.org/TR/XPath/>. Acesso em 08 de outubro de 2013.

YAO, X.; BORISOVA, I.; ALAM, M. **PDTB XML: the XMLization of the *Penn* Discourse *TreeBank*** 2.0. 2010. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2010/summaries/336.html>. Acesso em: 12 out 2013.

Artigo recebido em: 30.09.2014
Artigo aprovado em: 23.11.2014