Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006) - Poster Section, Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

Proposta de um Esquema de Anotação de Corpora Extensível

José Guilherme C. de Souza, Patrícia N. Gonçalves e Renata Vieira

Universidade do Vale do Rio dos Sinos (UNISINOS) *
Av. Unisinos 950 – 93022-000 – São Leopoldo – RS – Brasil
{joseguilhermecs, patricia.unisinos, renata.vieira}@gmail.com

Resumo Este artigo apresenta uma proposta de esquema de anotação lingüística em XML que possa vir a ser adotado na construção de *corpora* anotados da Língua Portuguesa, para uso da comunidade da área.

Para que a web semântica estabeleça-se em larga escala, faz-se necessário que um grande número de documentos seja anotado. Para esse propósito, ferramentas que possibilitam a anotação semi-automática de documentos têm sido desenvolvidas. No entanto, dado o grande número de documentos na web que consistem total ou parcialmente de texto não estruturado, mostra-se necessário o desenvolvimento de ferramentas que possibilitem a análise automática da estrutura semântica de documentos textuais [1].

Existem diversas ferramentas para anotação manual e automática de corpus com informações lingüísticas de vários níveis. Essas informações devem ser armazenadas de uma forma eficiente. Por eficiente, entende-se que os repositórios de dados com anotações lingüísticas devem permitir a expansão e a facilidade de uso e reuso dessas informações. Por se tratar de uma área recente, modelos de anotação que atendam às exigências citadas ainda estão sendo estudados.

Neste trabalho, é proposto um padrão de codificação para anotação lingüística para a Língua Portuguesa baseado na linguagem de marcação XML que atenda às características acima. Para isso, foram estudados diversos formatos que vêm sendo adotados pela comunidade de Processamento de Linguagem Natural. Entre eles, os padrões de anotação para recursos lingüísticos que têm sido discutidos e desenvolvidos pelo grupo ISO TC37 SC 4 (International Standards Organization-Language Resources Standards)[2], padrões utilizados por anotadores lingüísticos como o PALAVRAS[1][3] e por projetos, como o MuchMore[1] e o TIGER[4]. Através da análise e crítica dos modelos existentes, queremos chegar a uma proposta para atender necessidades de trabalhos com *corpora* anotados da Língua Portuguesa.

Dividimos as informações em três arquivos, de acordo com o esquema inicialmente proposto pelo PALAVRAS XTRACTOR[5]. São eles: arquivo de codificação do texto (figura 1), arquivo de informações morfossintáticas (figura 2) e arquivo com as informações estruturais sintagmáticas do texto (figura 3).

* Agradecimentos ao CNPq (Conselho Nacional de Desenvolvimento Tecnológico e Científico) pelo apoio financeiro - Projeto PLN-BR Proc. número 550.388/2005-2

[1] http://visl.hum.sdu.dk/visl/pt

Optamos por essa divisão pois, através dela, é possível desvincular o armazenamento do texto propriamente dito, das informações obtidas na anotação lingüística. Assim, diferentes versões do mesmo tipo de anotação, e, até mesmo, diferentes anotações (de *part-of-speech*, por exemplo), podem ser associadas ao texto. Essa opção vai de encontro aos princípios estudados pela ISO TC37 SC 4 em [2].

```
<text id="t1">
 <paragraph id="p1">
  <sentence id="s1">
   <word id="w1">A</word>
   <word id="w2">menina</word>
   <word id="w3">gosta</word>
   <word id="w4">de</word>
   <word id="w5">maçãs</word>
   <word id="w6">verdes</word>
   <word id="w7">.</word>
  </sentence>
 </paragraph>
</text>
```

Figura 1. Codificação do texto "A menina gosta de maçãs verdes".

O arquivo mostrado na figura 1 propõe um esquema de codificação para o texto. Ele consiste de quatro elementos: <text>, <paragraph>, <sentence> e <word>. O primeiro referencia um texto. Um texto pode ter vários elementos <paragraph>, que marcam os parágrafos do texto. Esses, por sua vez, podem ter sentenças definidas pelo elemento <sentence>. Cada sentença é formada por palavras. As palavras são representadas pelo elemento <word>. Todos elementos citados possuem um atributo de identificação *id*.

Através dessa estruturação, acreditamos que o texto fica melhor organizado e caracterizado, mais estruturado do que com as soluções propostas utilizadas pelos projetos Muchmore[1] e Tiger-XML[6], que codificam o texto no mesmo arquivo que as informações linguísticas. Em relação ao PALAVRAS XTRACTOR, a diferença é a inclusão de novas marcações relacionadas a estrutura do texto (parágrafos e frases).

Para o desenvolvimento do arquivo que representa as informações morfossintáticas (figura 2), utilizamos algumas das idéias usadas no projeto Muchmore. O arquivo é formado por elementos <token>. Esses elementos representam as palavras do texto anotado. Cada <token> possui um atributo identificador (*id*) seguido de sua classe gramatical (*class*), gênero (*gender*), número (*number*), forma canônica (*canon*) e, ainda, uma referência para a palavra no texto (*wordref*).

A principal mudança em relação ao formato utilizado pelo Muchmore é a adoção do atributo *wordref*. O formato proposto não repete a palavra no arquivo de informações morfossintáticas, apenas a referencia utilizando seu identificador, mas contém sua forma canônica. Além disso, assim como o formato do PALAVRAS XTRACTOR, utilizamos um elemento para especificar informações

```
<token id="t1" class="art" gender="F" number="S" canon="o" wordref="w1">
 <complement tag="artd" />
</token>
<token id="t2" class="noun" gender="F" number="S" canon="menina" wordref="w2">
 <complement tag="Hfam" />
</token>
<token id="t3" class="verb" n_form="fin" tense="PR" person="3S" mode="IND"
     canon="gostar" wordref="w3"/>
<token id="t4" class="prp" canon="de" wordref="w4"/>
<token id="t5" class="noun" gender="F" number="P" canon="maçã" wordref="w5">
 <complement tag="food-c" />
</token>
<token id="t6" class="adj" gender="F" number="P" canon="verde" wordref="w6"/>
```

Figura 2. Codificação dos dados morfossintáticos do texto da figura 1.

tais como informações semânticas (Hfam, que indica categoria humano e família) e refinamentos morfossintáticos (artd, indicando artigo definido) a respeito da palavra neste arquivo. Esse elemento é o elemento <complement>.

```
<struct id="st1" type="S" head="st5" paragraph_ref="p1" sentence_ref="s1">
  <struct id="st2" type="NP" function="subj" head="st4" from="w1" to="w2"/>
    <struct id="st3" type="art" from="w1" to="w1"/>
    <struct id="st4" type="n" from="w2" to="w2"/>
  </struct>
  <struct id="st5" type="VP" function="p" head="st6" from="w3" to="w3">
    <struct id="st6" type="v_fin" from="w3" to="w3"/>
  </struct>
  <struct id="st7" type="PP" function="piv" head="st8" from="w4" to="w6">
    <struct id="st8" type="prp" from="w4" to="w4"/>
    <struct id="st9" type="NP" function="p" head="st10" from="w5" to="w6">
      <struct id="st10" type="n" from="w5" to="w5"/>
      <struct id="st11" type="adj" from="w6" to="w6"/>
    </struct>
  </struct>
</struct>
```

Figura 3. Codificação dos dados estruturais sintagmáticos do texto da figura 1.

Para a concepção do arquivo com informações estruturais sintagmáticas do texto, utilizamos os elementos e atributos presentes na proposta do XCES[2] /ISO TC37 SC 4[7][2]. Esse arquivo consiste de elementos <struct> que podem ser aninhados como numa estrutura arbórea. Cada <struct> contém um identicador (*id*) e um tipo (*type*) como atributos obrigatórios. Conforme o tipo do elemento, diferentes atributos serão necessários.

Caso o elemento represente um sintagma, ou um elemento não-terminal da árvore sintática (contenha valores como NP, VP, PP, entre outros), ele deve apresentar atributos referentes àquele sintagma. São eles: a função sintática do sintagma (*function*), o núcleo do sintagma (*head*) e as palavras que compõem o sintagma, representadas por um intervalo (o atributo *from* denota o início e o

[2] http://www.cs.vassar.edu/XCES/

atributo *to*, o fim do intervalo). Caso o elemento represente um elemento terminal da árvore sintática, ou seja, represente uma palavra, os únicos parâmetros obrigatórios, além de *id* e *type*, são *from* e *to*, que indicam a qual palavra o elemento se refere.

A linguagem XML tem diversas vantagens que decorrem do fato de ser uma linguagem livre e flexível. Isso, no entanto, requer que um esforço maior seja necessário na definição de esquemas com propósitos específicos.

A proposta aqui apresentada levou em consideração diversas características identificadas como positivas em outros trabalhos e evitou reproduzir as características negativas. Entre elas, a utilização de identificadores em todos os elementos, a separação das informações linguísticas distintas em arquivos diferentes e a estruturação da informação utilizando os atributos. Além disso, uma vez que uma das ferramentas bastante difundidas e com uma anotação bem completa para a Língua Portuguesa tem sido o PALAVRAS[3], nossa proposta buscou acomodar as informações produzidas por essa ferramenta.

Esta proposta se insere no contexto do projeto PLN-BR[3] e tem como propósito o tratamento da informação mobilizada em um mesmo corpus do português do Brasil. É importante que essa proposta seja apresentada e discutida pela comunidade para que o esforço aqui realizado possa trazer benefícios para diferentes grupos que trabalham com PLN da Língua Portuguesa.

Referências

1. Buitelaar, P., Declerck, T.: Linguistic Annotation for the Semantic Web. In: Annotation for the Semantic Web. Volume 96 of Frontiers in Artificial Intelligence and Applications Series. IOS Press (2003)
2. Ide, N., Romary, L.: International standard for a linguistic annotation framework. Journal of Natural Language Engineering **10:3-4** (2004) 211–225
3. Bick, E.: The Parsing System "PALAVRAS- Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Department of Linguistics, University of Århus, DK. (2000)
4. König, E., Lezius, W.: The TIGER language - a description language for syntax graphs, Formal definition. Technical report (2003)
5. Gasperin, C., Vieira, R., Goulart, R., Quaresma, P.: Extracting xml syntactic chunks from portuguese corpora. In: Proceedings of the Workshop TALN 2003 Natural Language Processing of Minority Languages and Small Languages. (2003)
6. Vilela, R., Simoes, A., Bick, E., Almeida, J.J.: Representacao em xml da floresta sintactica. In Ramalho, J.C., Simões, A., Lopes, J.C., eds.: XATA2005, XML: Aplicações e Tecnologias Associadas (Vila Verde, Braga, 10 e 11 de Fevereiro de 2005), Universidade do Minho (2005) 351–361 http://hdl.handle.net/1822/865.
7. Ide, N., Bonhomme, P., Romary, L.: Xces: An xml-based encoding standard for linguistic corpora. In: Proceedings of the Second International Language Resources and Evaluation Conference. (2000)

[3] http://safe.icmc.usp.br/plnbr/