Anotação estrutural e morfossintática de um corpus em Português brasileiro utilizando a codificação da TEI P5 e o AeliusHunpos.

Cid Ivan da Costa Carvalho
Doutorando
Programa de Pós Graduação em Linguística da Universidade Federal do Ceará
Linha de Pesquisa: Descrição e Análise Linguística
Davis Macedo Vasconcelos
Doutorando
Programa de Pós Graduação em Linguística da Universidade Federal do Ceará
Linha de Pesquisa: Descrição e Análise Linguística
Ednardo Luiz da Costa
Doutorando
Programa de Pós Graduação em Linguística da Universidade Federal do Ceará
Linha de Pesquisa: Descrição e Análise Linguística
Gezenira Rodrigues
Doutorando
Programa de Pós Graduação em Linguística da Universidade Federal do Ceará
Linha de Pesquisa: Descrição e Análise Linguística
Katiuscia de Moraes Andrade
Mestranda
Programa de Pós Graduação em Linguística da Universidade Federal do Ceará
Linha de Pesquisa: Descrição e Análise Linguística

Leonel Figueiredo de Alencar Araripe (UFC)
Orientador

Início da pesquisa: Novembro de 2011
Data prevista para a conclusão: Novembro de 2012.

O presente trabalho tem por finalidade a anotação estrutural e morfossintática de um *corpus*, em Português brasileiro, com aproximadamente vinte mil *tokens*, utilizando a codificação TEI P5 e o AeliusHunPos, treinado no Corpus Tycho Brahe. Pelo que sabemos, trata-se do primeiro *corpus* de língua portuguesa a implementar esse tipo de codificação. Também serão anotados fenômenos de variação ortográfica por meio da *tag* <choice> e outras similares.

A anotação estrutural permite a marcação dos dados internos e externos do texto. Os dados externos referem-se aos metadados textuais. Os dados internos, por sua vez, compreendem a marcação da estrutura geral e da estrutura de subparágrafos (ALUÍSIO; ALMEIDA, 2006). A anotação estrutural será realizada de forma semiautomática, a partir da estrutura base da TEI dentro do etiquetador Aelius.

A *Text Enconding Iniative* (TEI) é uma codificação cujas diretrizes são dirigidas a todos que desejam trocar informações armazenadas em formato eletrônico, permitindo a manipulação do texto, o que não seria possível através de um texto em formato de imagem. As diretrizes da TEI definem meios para tornar explícitas certas características de um texto, a fim de permitir o seu processamento em diferentes máquinas (BURNARD;SPERBERG-MCQUEEN, 2010).

As diretrizes prescrevem o uso da linguagem XML (*Extensible Markup Language*). Ainda segundo as diretrizes, o design dos documentos precisa obedecer a alguns critérios, como: apresentar as características textuais necessárias para a pesquisa; ser simples, claro e concreto; ser usado facilmente por pesquisadores, sem a necessidade de uso de softwares especiais; permitir definições rigorosas e eficientes para o processamento de textos; permitir extensões definidas pelo

usuário e estar em conformidade com as normas existentes (BURNARD;SPERBERG-MCQUEEN, 2010).

A anotação linguística será a nível morfossintático, utilizando o etiquetador automático Aelius e o modelo HunPos, treinado no Corpus Tycho Brahe, dentro do próprio etiquetador. O Aelius é um software livre em Python que utiliza a biblioteca *Natural Language Toolkit* – NLTK (BIRD; KLEIN; LOPER, 2010). Essa ferramenta destina-se ao pré-processamento de textos, à construção de etiquetador morfossintático e à anotação automática de *corpora* com auxílio de revisão humana (ALENCAR, 2010).

O *corpus* a ser anotado consiste na produção textual (manuscritos) de alunos de escolas públicas das cidades cearenses de Barroquinha, Jijoca de Jericoacoara e Camocim, que participaram das oficinas da 2ª edição do projeto Rota das Especiarias – Temperos Literários.

Os textos foram digitados, preservando-se os desvios da norma culta e a estrutura original e, em seguida, foram nomeados e salvos no formato "txt". Em seguida, servirão de *input* para o etiquetador Aelius. Os erros decorrentes da etiquetagem serão corrigidos manualmente.

A identidade dos autores será resguardada, porém, serão explicitados os dados que permitam identificar as variações diatópicas, diafásicas e diastráticas.

O *output* será um corpus com etiquetas XML, segundo a codificação TEI P5 e etiquetagem morfossintática. Permitirá realizar estudos de variação linguística e embasar o desenvolvimento de ferramentas de tecnologia da linguagem natural, como, por exemplo, etiquetadores morfossintáticos mais robustos, capazes de lidar com a linguagem não padrão, e corretores ortográficos.

REFERÊNCIAS

ALENCAR, L. F. de. Aelius: uma ferramenta para anotação automática de corpora usando o NLTK. ELC 2010 – IX Encontro de Linguística de Corpus, PUCRS, Porto Alegre, 8 e 9 de outubro de 2010. Disponível em: <http://corpuslg.org/gelc/elc2010.php>

ALUÍSIO, Sandra Maria; ALMEIDA, Gladis Maria de Barcellos. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. **Caleidoscópio**. São Leopoldo. Vol. 4, n. 3, p. 155-177, set/dez, 2006.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Toolkit**. [s.l]: [s.n.], 2010. Disponível em: <http://www.nltk.org> Acesso em: 30 sep. 2010.

BURNARD, L.; SPERBERG-MCQUEEN, C. M. **TEI P5: Guidelines for Electronic Text**. [Text Encoding Initiative Consortium]: [Charlottesville, Virginia], 2006. Disponível em: <http://www.tei-c.org/release/doc/tei-p5-exemplars/html/teilite.doc.html> Acesso em: 16. set. 2011.