TÉCNICAS EM SOFTWARES LIVRES PARA LINGUÍSTICA DE CORPUS (10 ETAPA)

XXXVIII Encontro de Iniciação Científica

Fernando Antonio Saraiva Maia, Juliana Lopes Gurgel, Leidiana Iza Adrande Freitas, Leonel Figueiredo de Alencar Araripe

No presente projeto, codificamos a valência verbal dos 500 verbos mais frequentes do português do Brasil, segundo o paradigma da Gramática Léxico-Funcional (LFG, do inglês Lexical-Functional Grammar). A BrGram, a mais extensa gramática computacional do português brasileiro no formalismo LFG/XLE, iniciada por Alencar (2013), ainda carece de um léxico suficiente para a análise de textos reais. Nesse sentido, a codificação dessas valências verbais em moldes no formato LFG/XLE, através da criação do módulo BrVal 1.0, contribuirá para a ampliação da cobertura da gramática. O módulo foi construído a partir da extração dos 500 verbos mais frequentes do corpus NILC/São Carlos, os quais tiveram suas valências codificadas segundo os usos presentes em Borba et al. (1991). A análise manual de um conjunto de 100 sentenças gramaticais (teste positivo) e de um conjunto de 100 sentenças agramaticais (teste negativo) resultou em 87% de sentenças analisadas para o teste positivo, contradizendo nossa hipótese de que a acurácia do módulo BrVal 1.0 seria de pelo menos 95%. Por outro lado, 18% das sentenças do teste negativo foram analisadas, divergindo do valor proposto por Butt et al. (1999), que deve ser 0%, a fim de garantir que a gramática não hipergere. Assim, um estudo mais aprofundado se faz necessário para aumentar a acurácia da gramática. Agradecemos ao apoio da UFC, através das bolsas concedidas, para a realização deste projeto.

Palavras-chave: LINGUÍSTICA COMPUTACIONAL. LÉXICO. VALÊNCIA VERBAL. PROCESSAMENTO DA LINGUAGEM.