NHEENTIQUETADOR: UM ETIQUETADOR MORFOSSINTÁTICO PARA O SINTAGMA NOMINAL DO NHEENGATU

Leidiana Iza Andrade Freitas, Dominick Maia Alexandre, Juliana Lopes Gurgel, Leonel Figueiredo de Alencar Araripe

O presente trabalho tem como objetivo descrever as etapas de construção e testagem do Nheentiquetador 1.0, o primeiro etiquetador morfossintático para as classes de palavras que ocorrem no sintagma nominal da Língua Geral Amazônica, ou nheengatu. Embora a existência de etiquetadores morfossintáticos seja comum para as línguas majoritárias (ALENCAR, 2013, 2015), às línguas minoritárias é reservado um lugar à margem no que se refere ao tratamento computacional. O desenvolvimento de um recurso de PLN básico como um etiquetador, bem como a disponibilização de um corpus anotado, estabelecem um novo lugar para o nheengatu no atual contexto científico e tecnológico. A abordagem para o desenvolvimento do etiquetador foi baseada no conhecimento, devido à escassez de corpora anotados para o referido idioma. Desse modo, as regras implementadas foram baseadas em descrições gramaticais disponíveis, mais especificamente nos trabalhos de Navarro (2011) e Cruz (2011). A metodologia da pesquisa foi dividida em duas partes: (i) a compilação do corpus e (ii) a implementação do etiquetador. Na parte (i), foram feitas a tokenização, a extração e a normalização de parte dos textos de Navarro (2011). Na parte (ii), foram feitas a implementação do algoritmo do etiquetador e a testagem da ferramenta (VOUTILAINEN, 2004; JURAFSKY; MARTIN, 2019). O Nheentiquetador 1.0 atingiu um F-score de 0.83 na etiquetagem morfossintática automática de uma amostra de 10% das sentenças do corpus compilado. A acurácia obtida nos primeiros testes distancia-se do estado da arte, que é 95%. Deste modo, um aperfeiçoamento da ferramenta se faz necessário para aumentar sua acurácia. Ainda assim, os resultados e produtos derivados desta pesquisa fornecem, para a comunidade acadêmica, um conjunto de sentenças em nheengatu úteis à aplicação de diferentes ferramentas destinadas ao processamento computacional desta língua.

Palavras-chave: linguística computacional. nheengatu. etiquetador morfossintático. processamento de linguagem.