EXTRAÇÃO (SEMI)AUTOMÁTICA DE TERMOS EM CORPORA DE LÍNGUA PORTUGUESA: MÉTODOS, TÉCNICAS E FERRAMENTAS COMPUTACIONAIS

Resumo: A pesquisa terminológica baseada em corpus (*corpus-based*) ou direcionada pelo corpus (*corpus-driven*) é hoje a tônica nos grupos que desenvolvem projetos em Terminologia, tanto no Brasil como no exterior. Temos fácil acesso a determinadas ferramentas de processamento automático de língua natural (PLN) que auxiliam a tratar os dados que nos chegam em formato eletrônico, tais como compiladores automáticos de corpora, contadores de frequência, concordanciadores, extratores de palavras-chave, anotadores (estrutural e linguístico), etc. Entretanto, há ainda para o cenário da língua portuguesa algumas lacunas no que se refere à extração (semi)automática de termos, o que acaba gerando grande trabalho humano para limpar enormes listas e filtrar o que realmente é termo do domínio especializado, sem contar os baixos índices obtidos quando aplicamos métricas clássicas da área de processamento, como a precisão e a revocação (*recall*). Pretendemos, pois, em nossa apresentação, mostrar diferentes métodos, técnicas e ferramentas, todos testados e avaliados em projetos finalizados e em andamento, de forma a auxiliar pesquisadores que, de alguma forma, têm de lidar com recuperação de informação (RI) em corpora.

Palavras-chave: corpus; Terminologia; ferramenta computacional; extração (semi)automática de termos.