Evento	Salão UFRGS 2014: SIC - XXVI SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2014
Local	Porto Alegre
Título	Combinatórias Léxicas Especializadas: extração com o uso do MWE Toolkit
Autor	PAULO GUILHERME PILOTTI DUARTE
Orientador	CLECI REGINA BEVILACQUA

Combinatórias Léxicas Especializadas: extração com o uso do MWE Toolkit

O Projeto Combinatórias Léxicas Especializadas (CLEs) da Linguagem legal, normativa e científica (ProjeCom) do Grupo Termisul tem como objetivo desenvolver uma base de dados *on-line* para registrá-las. Os usuários previstos são tradutores e redatores, além de profissionais de outras áreas da Linguística e da Informática, legisladores, advogados, empresários e profissionais da comunicação. A proposta é coletar as expressões multivocabulares prototípicas da legislação ambiental brasileira e seus equivalentes na legislação do meio ambiente dos países do Mercosul (Argentina, Uruguai e Paraguai), Alemanha, Estados Unidos, França e Itália. Nesta apresentação, relato as diferentes etapas realizadas para a aplicação do *Multiword Expression Toolkit* (MWEToolkit) – ferramenta que auxilia na extração de candidatos a CLEs – no *corpus* da legislação ambiental brasileira. A extração das expressões multivocabulares envolveu várias etapas: indexação, extração, filtragem e contagem dos candidatos. Na etapa de indexação, o MWEToolkit cria uma matriz que permite o acesso rápido a qualquer n-grama. Os n-gramas são grupos de palavras repetidas ao longo do *corpus*, podendo ter várias extensões – unigrama, bigrama, trigrama etc, e para os quais é indicada a frequência. Após essa etapa, para a extração, foram criados dois arquivos de padrões morfossintáticos das CLEs, em linguagem XML, utilizados para a identificação dos candidatos. Esses arquivos contêm padrões nominais (Ex.: <pat><w pos="NOM"/><w pos="PRP+DET"/><w pos="NOM"/></pat>) e verbais (Ex.: <pat><w pos="V"/><w pos="NOM"/></pat>) que compõem os filtros linguísticos utilizados na extração. Para o primeiro padrão, espera-se extrair combinatórias como *acondicionamento do produto* e *coleta de resíduos*. Para o segundo, *acondicionar resíduos* e *adquirir energia.* Ao aplicar apenas os filtros linguísticos obtivemos 7638 candidatos nominais e 3244 candidatos verbais. Após uma verificação dos resultados, percebemos que esse elevado número de candidatos trazia consigo uma grande quantidade de ruído, ou seja, expressões multivocabulares que não se caracterizam como CLEs. Para diminuir este ruído, aplicamos filtros de frequência. Passamos, assim, para a etapa de filtragem em que foram aplicados simultaneamente dois filtros de frequência. Para o primeiro filtro, utilizamos os termos mais frequentes do *corpus* na posição de "NOM" nos padrões mencionados acima e aplicamos, ao mesmo tempo, um corte de frequência igual ou superior a dois. Desta forma, para o grupo dos dez termos mais frequentes no *corpus* – *natureza, substâncias, material/materiais, animal/animais, solo, poluição, embalagem/ embalagens, pesca, óleo, vegetação* – aplicados aos padrões verbais, obtivemos 29 candidatos a CLEs e para os padrões nominais, 240 candidatos. Este conjunto de candidatos é contrastado com um *corpus* de língua geral (*corpus* de referência) para que seja possível identificar as combinatórias prototípicas do *corpus* de legislação ambiental. A partir da aplicação desse contraste, baseado em critérios estatísticos, restaram 22 candidatos para os padrões formados por verbos e 225 para os padrões nominais. Após a filtragem dos candidatos, estes são contados e ordenados de acordo com a sua frequência no corpus legislativo, dentro de cada um dos grupos de padrões. Estes resultados são, então, analisados de forma mais detalhada pelos pesquisadores, considerando seu contexto de ocorrência, de forma a selecionar as expressões que se caracterizam como CLEs, seguindo os critérios estabelecidos pela equipe. Os dados extraídos servem para confirmar como CLEs as unidades já identificadas pelo grupo através do uso de outras ferramentas, além de oferecer novas combinatórias que serão incluídas na base de dados. Destacamos ainda a importância da aplicação do MWE, pois ele utiliza critérios linguísticos e estatísticos que otimizam a extração e asseguram a confiabilidade dos dados obtidos.