Autores: Paulo Guilherme Pilloti, Anna Maria Becker Maciel, Cleci Regina Bevilacqua
Linha de pesquisa: Teorias lingüísticas do léxico.
Data de início da pesquisa: janeiro de 2012
Previsão de data de conclusão: dezembro de 2012

Pré-processamento de um *corpus* especializado para extração de expressões multivocabulares

Os avanços dos estudos da linguagem favorecem o intercâmbio de pesquisas que permitem, de um lado, a complementação dos procedimentos manuais do linguista por dados empíricos oferecidos pela máquina, de outro, oportunizam a avaliação do desempenho de novos programas de Processamento da Linguagem Natural (PLN). O presente trabalho relaciona-se a um projeto que desenvolve a pesquisa plurilíngue de expressões multivocabulares que ocorrem em vários *corpora* especializados. Um desses *corpora* é formado por textos legais em português e em línguas estrangeiras (alemão, espanhol, francês, inglês e italiano). Enfocamos aqui a coleta e extração de candidatos a expressões multivocabulares recorrentes nos textos legais em língua portuguesa, mais especificamente, os referentes à Legislação Ambiental Brasileira. No âmbito de tal empreendimento, realizamos o pré-processamento de um *corpus* de linguagem especializada para aplicação de uma ferramenta de PLN, o extrator de expressões multivocabulares, MWEToolkit. Nosso objetivo é descrever o caminho percorrido em busca da adequação das condições humanas e logísticas de nossa pesquisa à utilização desse recurso. As diferentes etapas do processo compreenderam aspectos teóricos e práticos que se, para uns parecem triviais, para outros são difíceis de dominar. Caracterizamos, em primeiro lugar, nosso *corpus* de estudo, detalhando sua composição, a linguagem especializada contemplada e a tipologia dos textos incluídos. Em seguida, descrevemos as ferramentas escolhidas para seu pré-processamento. Relatamos, então, os primeiros passos dados desde a familiarização dos participantes do projeto com o sistema de anotação morfossintática do etiquetador TreeTagger, pré-requisito para a utilização do MWEToolkit, até a atualização dos recursos operacionais disponíveis. Finalmente, tratamos da aplicação do TreeTagger e da análise do seu desempenho e demonstramos de que maneira os problemas detectados foram dirimidos. Como conclusão, apresentamos uma reflexão sobre a aplicação na linguagem especializada de etiquetadores e extratores treinados para língua geral e tecemos considerações sobre a validade da colaboração de informatas e lingüistas em um empreendimento conjunto.