Combinatórias Léxicas Especializadas: etapas prévias para sua extração utilizando recursos do Processamento da Linguagem Natural

No âmbito do Projeto Combinatórias Léxicas Especializadas (CLEs) da Linguagem legal, normativa e técnico-científica (ProjeCom) do Grupo Termisul, sou responsável pela informatização do Projeto. No projeto, as CLEs são entendidas como unidades sintagmáticas, recorrentes e condicionadas pela língua, área ou gênero textual. Nesta apresentação, relato a preparação de *corpora* multilíngues de textos legislativos de Direito Ambiental para uso em um extrator de unidades léxicas multivocabulares, *Multiword Expression Toolkit* (MWEToolkit). Sua utilização visa complementar a pesquisa manual assistida pelo computador até então realizada pelos pesquisadores do Grupo. A ferramenta exige a inserção, depois de cada palavra do *corpus,* de uma etiqueta, indicando sua categoria gramatical. Tal anotação morfossintática é realizada por um etiquetador automático, o *TreeTagger*. Nesse contexto, enfrentei problemas relativos à atualização do meu conhecimento em relação a essas ferramentas e à adequação das condições logísticas vigentes no projeto. A metodologia adotada procurou conciliar os aspectos práticos e aplicados da tarefa. Em primeiro lugar, foi necessária a instalação do ambiente operacional Linux para que as ferramentas pudessem ser colocadas em funcionamento. Em seguida, a partir da página (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/) foi feito o *download* do *TreeTagger* e seus respectivos arquivos de parâmetros para as diferentes línguas contempladas pelo Termisul: português, alemão, espanhol, francês, inglês e italiano. Posteriormente, o MWEToolkit foi instalado com a orientação do seu desenvolvedor e os dois programas foram rodados. Iniciamos o processamento com o *corpus* de língua portuguesa. Os arquivos de saída desse *corpus* etiquetados pelo TreeTagger foram submetidos ao MWEToolkit. Nesse momento, surgiram problemas causados pela presença de marcações XML indicando as partes macroestruturais do texto legislativo, como título, ementa, data, e outras. Informado das dificuldades, seu autor realizou correções no programa embutido no interior do programa principal que permite a automatização de tarefas secundárias - denominado *script* - e sugeriu que estes elementos problemáticos fossem retirados do *corpus* original. Neste momento, está sendo realizado, de maneira experimental, o processamento passo a passo dos sete *scripts* que precedem a análise do *corpus*, partindo do *script* de identificação, passando pelo de extração e contagem até chegar ao de avaliação das CLEs. Esta tarefa atualmente encontra-se em andamento, pois exige muito tempo dada a complexidade dos *scripts* e a extensão do *corpus.* Este trabalho enfatiza, de um lado, a validade do uso da ferramenta informatizada MWEToolkit na identificação e extração das CLEs, como recurso complementar para ampliação e aprofundamento da pesquisa linguística empreendida pelo Termisul. De outro lado, testemunha a viabilidade da colaboração de linguistas e informatas em uma tarefa conjunta.