Etiquetagem morfossintática de 1 milhão de palavras: relatos da experiência lingüística num processo semi-automático

O objetivo deste trabalho é demonstrar o processo utilizado pela equipe do NILC na etiquetagem morfossintática de um corpus com 1 (um) milhão de palavras - recurso básico para o treinamento de etiquetadores estatísticos e híbridos.
A etiquetagem morfossintática é uma tarefa bastante conhecida no domínio do Processamento de Língua Natural (PLN). Esse processo consiste em atribuir uma categoria gramatical (e talvez outros atributos referentes à categoria, por exemplo, gênero, número, etc) a cada palavra de um texto. A anotação de um texto com etiquetas morfossintáticas é útil para várias manipulações subseqüentes de um texto, por exemplo: análises lingüísticas, alinhamento de textos, sumarização, etc. No entanto, essa anotação é tradicionalmente feita de modo manual, o que requer tempo e equipe interdisciplinar especializada – fatores complicadores do processo. Diante disso, elaboramos uma nova proposta metodológica, que consistiu de várias etapas de etiquetagem semi-automáticas. Para tanto, partimos de um corpus com aproximadamente 100 mil palavras etiquetadas manualmente para, por meio de um processo incremental, chegar ao corpus de 1 (um) milhão de palavras. Essa proposta metodológica semi-automática foi dividida em cinco fases, sendo que a cada fase o corpus aumentaria o equivalente ao tamanho anterior mais 100 mil palavras. Em todas as fases, o corpus seria etiquetado pelo NILC Tagset e, em seguida, 10% desse corpus seriam corrigidos manualmente. Dessa correção manual, elaboraríamos regras que seriam utilizadas como um filtro para eliminar parte dos erros da etiquetagem daquela fase. Entretanto, no processo de criação dessas regras, os lingüistas se depararam com problemas extremamente complexos, como: ambigüidade categorial, ambigüidade de cunho semântico, pertinência e abrangência das regras, etc. E são exatamente essas experiências lingüísticas que serão enfatizados neste trabalho.