Unitex-PB: desenvolvimento e disponibilização de recursos e ferramentas lingüístico-computacionais para o Português do Brasil

Aplicações em Processamento de Línguas Naturais (PLN) necessitam, em geral, de recursos e ferramentas lingüístico-computacionais para realizarem as tarefas a que se propõem. Tradutores automáticos e sumarizadores de textos, por exemplo, podem utilizar dicionários e anotadores morfossintáticos para reconhecer e processar cada palavra e sua estrutura morfológica; revisores gramaticais utilizam, ainda, critérios de desambiguação lexical e regras gramaticais para construir a estrutura sintática das sentenças de um texto e sugerir possíveis revisões gramaticais; ferramentas de auxílio à escrita utilizam padrões lexicais para o reconhecimento da estrutura textual. A construção individualizada de recursos e ferramentas desta natureza por diferentes grupos de pesquisa para diferentes aplicações causa a replicação dispendiosa desse conhecimento em formalismos próprios e de difícil reutilização em outras aplicações e por outros grupos. Neste cenário, seguindo as tendências internacionais, construiu-se e disponibilizou-se (via Internet), para o português do Brasil, o Unitex-PB, um ambiente computacional que contém um conjunto de recursos e ferramentas lingüístico-computacionais usando um formalismo de codificação de conhecimento chamado DELA (*Dictionnarie Electronique du LADL*), desenvolvido na França e amplamente difundido e utilizado. Em seu estado atual, o Unitex-PB já conta com três grandes dicionários: um dicionário de palavras simples, com 67.500 lemas e as regras de flexão correspondentes associadas aos lemas; um dicionário de palavras simples já flexionadas, com 880.000 entradas; e um dicionário de palavras compostas flexionadas, com 4.000 entradas. Há, também, um conjunto de regras de desambiguação lexical importadas do ReGra, revisor gramatical desenvolvido pelo NILC (Núcleo Interinstitucional de Lingüística Computacional), seguindo-se o formalismo ELAG (*Elimination of Lexical Ambiguities by Grammars*). Cerca de 80 regras já foram incorporadas ao sistema. O Unitex-PB, assim como algumas aplicações desenvolvidas com base nele, encontra-se disponível on-line e tem servido como base para a construção de novas aplicações em PLN, permitindo a uniformização e reutilização do conhecimento lingüístico-computacional.