014 **UM SISTEMA DE IDENTIFICAÇÃO DE EXPRESSÕES COMPOSTAS PARA AUXÍLIO À CONSTRUÇÃO DE DICIONÁRIOS.** *Carlos Eduardo Ramisch, Marco Aurélio Pires Idiart, Aline Villavicencio (orient.)* (UFRGS).

Na área de Processamento de Linguagem Natural (PLN), a identificação de Expressões Compostas é considerada um problema fundamental. Esse tipo de expressão é bastante heterogênea, incluindo substantivos compostos (traffic light), locuções verbais (take into account), locuções preposicionais (on top of), entre outros. Os sistemas de PLN – sistemas de tradução automática e de geração automática de textos, por exemplo – costumam ser baseados em ferramentas e recursos lingüístico-computacionais, tais como gramáticas e dicionários de grande cobertura. É estimado que, em uma língua, o número de Expressões Compostas utilizadas se equipare ao número de palavras individuais no vocabulário de um falante nativo, portanto, é muito importante que essas expressões sejam inseridas adequadamente nos recursos lingüístico-computacionais. Em sistemas de PLN, é interessante que se possua conhecimento dessas expressões, a fim de evitar que algumas construções, apesar de gramaticais, soem artificiais para um falante nativo (como por exemplo "café poderoso" ao invés de "café forte"). O objetivo desse trabalho é, através da integração e refinamento de diversas ferramentas estatísticas (testes de hipóteses, informação mútua, etc.) e baseadas em conhecimento (etiquetadores, filtros, etc) em desenvolvimento, construir um ambiente integrado de identificação semi-automática de Expressões Compostas para auxiliar na tarefa de criação de recursos lingüístico-computacionais por lexicógrafos e desenvolvedores de sistemas de PLN. O ambiente provê uma interface entre o engenheiro de gramática e os métodos baseados no uso de textos de corpora e de consultas à Web para realizar inferências sobre candidatos a Expressões Compostas. (BIC).