Corpora e Neologismos

Ieda Maria Alves (FFLCH-USP)
Sandra Maria Aluísio (ICMC-USP)
Thiago Alexandre Salgueiro Pardo (ICMC-USP)
Bruno Maroneze (FFLCH-USP)

O Projeto TermNeo (Observatório de Neologismos Científicos e Técnicos do Português Contemporâneo – Projeto CNPq n. 550520/2002-3), tem se dedicado, desde 1993, a coletar e analisar a neologia do português contemporâneo do Brasil com base em um corpus jornalístico constituído por jornais (Folha de S. Paulo e O Globo) e revistas de divulgação (IstoÉ e Veja). Os primeiros materiais analisados, sob forma impressa, foram aos poucos sendo substituídos por materiais informatizados, à medida que os jornais e revistas foram sendo disponibilizados na Internet. São, atualmente, analisados com o auxílio de programas como Folio Builder e WordSmith tools. Esses recursos, no entanto, não se mostram eficazes para agilizar a coleta neológica, uma vez que não dispensam a leitura por parte de pesquisadores do Projeto. Desse modo, com a finalidade de se conseguir uma extração automática de neologismos, formou-se uma parceria junto a pesquisadores do NILC (Núcleo Interinstitucional de Lingüística Computacional), que disponibilizou o léxico do ReGra - Revisor Gramatical do português brasileiro (www.nilc.icmc.usp.br/nilc/projects/regra.htm), os recursos desenvolvidos no Projeto Unitex-PB (http://nilc.icmc.usp.br:8180/unitex-pb/index.html), o módulo de criação de palavras-compostas do próprio Regra, além de interfaces computacionais com o VOLP – Vocabulário Ortográfico da Língua Portuguesa, 3ª edição, 1999 (www.academia.org.br) – para a criação de um ambiente informatizado que, a partir de um corpus de entrada, lista as palavras candidatas a neologismos junto a seu contexto para avaliação dos pesquisadores do Projeto TermNeo. O léxico atual do ReGra possui 68.530 canônicas e 1.563.136 entradas e, dentre essas, 5.763 são locuções (de nomes próprios, prepositivas, adjetivas, adverbiais, etc.). Os recursos do Unitex-PB são compostos de um Dicionário de Palavras Simples para o Português Brasileiro que possui aproximadamente 67.500 canônicas associadas a suas regras de flexão; um Dicionário de Palavras Simples Flexionadas para o Português Brasileiro que possui aproximadamente 880.000 palavras flexionadas; e um Dicionário de Palavras Compostas Flexionadas para o Português Brasileiro que possui aproximadamente 4.000 entradas. Nessa apresentação, mostraremos os recursos do ambiente e uma avaliação de seu desempenho com vistas a agilizar a coleta neológica. O ambiente informatizado para a extração de neologismos será disponibilizado para uso por outros pesquisadores que enfrentam o mesmo problema na coleta neológica.