Contributions to the computational processing of diachronic linguistic corpora

Cunha, E. Landulfo Teixeira Paradela

Resumo

A linguística de corpus assistida por computador é um dos principais pontos de convergência entre métodos linguísticos e computacionais. Em particular, o uso de corpora linguísticos diacrônicos oferece oportunidades para a análise quantitativa e computacional de fenômenos relacionados à mudança linguística ao longo do tempo. Esta tese apresenta o resultado de três projectos independentes, embora relacionados, que partilham o interesse na aplicação de métodos computacionais na linguística de corpus diacrônica. Seu principal objetivo é oferecer contribuições para três das etapas da pesquisa envolvendo corpora linguísticos diacrônicos: (a) construção e compilação de corpora; (b) elaboração de ferramentas e algoritmos para exploração de dados; e (c) análise de dados para pesquisa linguística, cultural e histórica. Dois recursos úteis para a linguística de corpus em uma língua diferente do inglês são inicialmente apresentados: um coletor de comentários de portais de notícias, desenvolvido em código aberto e gratuito para uso, modificação e distribuição; e um corpus diacrônico gratuito composto por comentários publicados no UOL, um dos principais portais de notícias brasileiros. Esses recursos são relevantes não apenas para linguistas, mas também para profissionais de outras áreas, incluindo cientistas sociais e jornalistas interessados na percepção

pública de notícias e na relação entre mídia e sociedade. Em seguida, propõe-se um método simples e generalizável para auxiliar na identificação dos períodos de estabelecimento e obsolescência de itens linguísticos em um corpus diacrônico baseado na frequência desses itens no corpus. Esse método pode ser empregado para a análise de qualquer coleção de itens linguísticos, independentemente da língua ou período histórico. Sua aplicabilidade é demonstrada a partir da utilização de dados lexicais do Corpus of Historical American English (COHA), para o qual são fornecidos estudos de caso de caráter quantitativo e qualitativo sobre as palavras que aparecem ou desaparecem desse corpus em diferentes períodos. Finalmente, descreve-se como corpora diacrônicos podem ser usados para a investigação linguística quantitativa, apresentando um método centrado na pesquisa lexical por meio de uma abordagem diacrônica que emprega métodos complementares da linguística de corpus e do processamento de língua natural. A aplicabilidade desse método é demonstrada por meio da análise de caso do termo *fake news*, em que investigam-se sua percepção e conceitualização na mídia tradicional utilizando dados coletados de fontes da mídia brasileira e de língua inglesa. Com essas contribuições, espera-se colaborar para as pesquisas sobre linguística de corpus diacrônica e sobre métodos computacionais para análise linguística.

Palavras-chave: linguística de corpus; linguística assistida por computador; estudos diacrônicos.