ELABORAÇÃO DE UM CORPUS DE LINGUAGEM ORAL DO PORTUGUÊS DO BRASIL

Aluno: Rodrigo Segges Ferreira Baros
Orientadora: Maria Carmelita P. Dias

Introdução

Este projeto pretende ajudar a elaborar um corpus significativo de linguagem oral do português do Brasil, em especial do estado do Rio de Janeiro. Assim, estão sendo compiladas e organizadas amostras de língua falada coletadas em contexto, ou seja, em situações reais de uso. Diversos gêneros da linguagem oral estão sendo contemplados, tal como entrevistas, reuniões de trabalho, atendimentos de balcão e aulas. Tais amostras fazem parte de teses e dissertações defendidas no Departamento de Letras nos últimos anos.

A importância dessa pesquisa se deve à importância que existe hoje, nos estudos lingüísticos, de estudos baseados em *corpus*. *Corpus* é uma palavra de origem latina que se refere a qualquer amostra de texto, isto é, um corpo de texto. No entanto, ultimamente, quando falamos em *corpus* nos referimos a volumes consideráveis de dados coletados de uma determinada língua, mantendo-se principalmente uma garantia de representatividade (Biber et al. 1998; Sardinha 2004). Dentre as principais carências de *corpora*, podemos citar amostras de linguagem oral, devido às dificuldades de coleta.

Objetivos

O objetivo deste projeto é elaborar um *corpus* equilibrado e representativo do português do Brasil em sua modalidade oral. Por equilibrado, entende-se composto de textos de situações de uso oral, configurando cada grupo de textos como um gênero, e mantendo-se o mesmo número de palavras para cada gênero. Como representativo, entende-se composto de textos típicos de vários tipos diferenciados de situações de uso de língua oral. Um objetivo relacionado é a proposta de partição de gêneros de linguagem oral, com grupos mais abrangentes e mais específicos, levando em conta, entre outros critérios, a formalidade da situação, a intimidade e a hierarquia dos participantes e as particularidades da situação. Pretende-se também etiquetar os textos, de forma que possam ser usados para pesquisa lingüística.

Metodologia

A dificuldade de coleta de textos de linguagem oral impõe certas restrições de tempo para a consecução dos objetivos propostos. Foi feita uma seleção de exemplos de linguagem oral constantes de teses e dissertações defendidas no Departamento de Letras. Todo o material teve de ser digitalizado. O material foi escaneado, sendo que algumas partes tiveram de ser digitadas. A codificação dos textos foi feita seguindo o padrão do *corpus* de língua inglesa oral da Universidade de Michigan. Estão, assim, computadas não só a origem e algumas características de cada falante, mas também marcações de pausas, hesitações, entonações peculiares e mudanças de turno. Essas medidas servem para pesquisas que lidem com diversidades sociolinguisticas e diferenciações entre gêneros. Posteriormente, foi feita uma "limpeza" nos textos para que pesquisas lingüísticas que envolvem apenas dados léxicais e gramaticais também possam ser realizadas. Foram compiladas 10 teses e/ou dissertações, algumas das quais possuem mais de uma amostra da língua falada, que correspondem a diferentes situações de uso e, portanto, a diferentes gêneros. Os extratos lingüísticos coletados

a partir dessas teses serviram como base de dados para o projeto e estão computadas cerca de 4.000 palavras, do gênero entrevista.

Uma outra providência foi transformar todos os arquivos de formato doc ou txt também para unicode, de forma que possam ser tratados por diferentes tipos de etiquetadores.

Conclusões

O projeto encontra-se em andamento, mas foram encontradas certas dificuldades no que diz respeito às decisões quanto às normas de transcrição utilizadas nas amostras selecionadas. Parte dos textos não segue convenções de transcrição de linguagem oral; tais textos aproximam-se mais da língua escrita na medida em que utilizam a pontuação (pontos, vírgulas, letras maiúsculas) conforme as regras desta modalidade. A solução encontrada foi criar um novo grupo de textos, de caráter semi-orais. Assim, com os mesmos textos, serão possíveis vários tipos de pesquisas, com ferramentas diferenciadas.

Referências:

Biber, D., Conrad, S. & Reppen, R .**Corpus Linguistics: Investigating Language Structure and Use**. Cambridge University Press, 1998.

Sardinha, T.B. **Lingüística de Corpus**. Barueri, SP: Manole. 2004.