CORPOBRAS PUC-RIO: DESENVOLVIMENTO E ANÁLISE DE UM CORPUS REPRESENTATIVO DO PORTUGUÊS

Alunos: Márcia Gonzaga de Brito
Rubiane Guilherme Valério
Gabriel Paladino de Almeida
Orientadora: Lúcia Pacheco de Oliveira

Introdução

Este projeto propõe-se à compilação de um corpus representativo do português do Brasil (CORPOBRAS – PUC-Rio) através de uma coleta equilibrada de gêneros do discurso oral, escrito e escrito para ser falado. A ampliação desse corpus tem sido feita através da incorporação de diversos gêneros discursivos com o objetivo de disponibilizá-los através de um banco de dados *on-line*. Nesta etapa do projeto, foram acrescentados ao corpus oito gêneros discursivos, totalizando o número de vinte gêneros do discurso escrito, cinco gêneros do discurso oral, e dois gêneros do discurso escrito para ser falado.

Este projeto, desenvolvido no âmbito do PIBIC, está vinculado ao projeto de pesquisa 'Compilação de um corpus do português do Brasil e análise multidimensional da variação dos gêneros discursivos', em andamento com apoio do Edital Universal CNPq n°019/2004.

Objetivos

Um dos objetivos desta fase do projeto foi a ampliação do corpus, prioritariamente, através da seleção e da compilação de textos do discurso oral, para que se obtivesse uma amostra equilibrada de textos dos diversos gêneros discursivos e o corpus atingisse a meta de 1 milhão de palavras.

Outro objetivo do projeto concentrou-se na compilação e análise de textos do discurso acadêmico na perspectiva da Lingüística Sistêmico-Funcional. Para tanto, foram feitas leituras de textos teóricos com foco na metáfora gramatical (Halliday, 1994) a fim de verificar como seu uso influencia a escrita de textos acadêmicos.

Além disso, pretendeu-se disponibilizar esses textos acadêmicos, juntamente com os demais textos do corpus, através da criação de uma interface digital com base em um programa de banco de dados que possibilitasse o acesso e a disponibilização do corpus via internet.

Metodologia

Nesta etapa do projeto, foram compilados os gêneros: cartas ao editor, dissertações de mestrado, teses de doutorado, conversa carioca, conversa de crianças, grupos de enfoque, entrevistas e atendimento ao cliente. Os textos de alguns destes gêneros, tais como, 'atendimento ao cliente', 'conversa carioca' e 'conversa de criança', foram disponibilizados por docentes que coordenam outros projetos no Departamento de Letras, (Oliveira, 2003-2007; Correa, 2003-2006 e 2001-2003).

Nas teses e dissertações compiladas foram observados processos de transformação de idéias mais concretas em mais abstratas através da análise de traços lingüísticos que representam processos verbais e nominalizações (Ravelli, 2003; Biber, 1988).

Em relação à reorganização dos textos anteriormente compilados, os gêneros 'discursos políticos' e 'redações de alunos universitários' foram reagrupados seguindo a codificação dada aos textos na última etapa do projeto, correspondente a uma sigla que indica o gênero e o número dos mesmos, a língua na qual foram produzidos e se foram redigidos por

um falante nativo ou não. Uma nova categoria discursiva, denominada 'discurso escrito para ser falado', também foi criada para suprir a necessidade de classificação mais adequada dos gêneros 'discursos políticos' e 'roteiros cinematográficos'.

Com a finalidade de tornar acessíveis os dados de identificação dos textos para futuras consultas, e para a elaboração do banco de dados, foram redigidos relatórios em forma de tabelas para todos os gêneros, constando o número total de palavras, a origem de cada texto, e a descrição dos participantes. A partir desses documentos, iniciou-se o levantamento de todas as informações para a formação de um banco de dados e a criação de uma interface digital para acesso dos textos do CORPOBRAS.

Conclusão

Com a compilação de oito gêneros, o corpus atingiu 1 milhão de palavras, equiparando-se a corpora internacionais como o Brown Corpus, o LOB e London Lund Corpus. Os seguintes gêneros estão representados no CORPOBRAS: artigos científicos, cartas ao editor, cartas de reclamação, cartas de recomendação, cartas pessoais, cartas profissionais, cartas profissionais acadêmicas, circulares, contos, crônicas, dissertações de mestrado, editoriais, e-mails acadêmicos, e-mails pessoais, notícias de jornal, redações de alunos, redações de alunos universitários, redações de vestibular, romances, teses de doutorado, conversas cariocas, conversas de crianças, entrevistas, grupos de enfoque, atendimento ao cliente, discursos políticos e roteiros cinematográficos.

Quanto às análises dos textos acadêmicos, concluiu-se que durante o processo de escrita e reescrita houve uma variação no grau de abstração desses textos. Observou-se que os textos sofrem transformações de idéias mais concretas em mais abstratas através do uso de nominalizações em lugar de processos verbais. Nesse sentido, os resultados indicam que o grau de abstração destes textos varia de acordo com o uso da metáfora gramatical (Simon-Vanderbergen, Taverniers & Ravelli, 2003).

Em relação ao banco de dados, o mesmo está em fase de organização quanto à informatização e à disponibilização *on-line* dos dados do CORPOBRAS, sendo futuramente necessárias pesquisas complementares sobre direitos autorais relativos a corpora em ambientes virtuais.

Referências

1. BIBER, D., Conrad, S e Reppen, R. (1998) *Corpus Linguistics*. Cambridge: Cambridge Universits. Press.
2. CORREA, L. M. S. (2003-2006) *Concordância de Gênero e de Número e o Conceito de Interpretabilidade em Teorias do Processamento e da Aquisição da linguagem. CNPq 551491/2002-7* e (2001-2003) *Processamento de relações de concordância e a aquisição do sistema de gênero em português. CNPq 523434/96-0*
3. HALLIDAY, M.A.K (1994). *An introduction to functional grammar*. London: Edward Arnold. 2ª ed.
4. OLIVEIRA, M.C.L. (2002-2004). *Alta tecnologia e trabalho: um estudo da interação atendente - cliente em uma central de atendimento telefônico*. CNPq 521686/94-6.
5. SIMON-VANDENBERGEN, A., Taverniers, M & Ravelli, L. (Eds.) (2003). *Grammatical metaphor: Views from Systemic functional linguistics*. John Benjamins: Amsterdam