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

**Alunos: Márcia Gonzaga de Brito[1]**
**Rubiane Guilherme Valério**

**Orientadora: Lúcia Pacheco de Oliveira**

## Introdução

Esta pesquisa, desenvolvida no âmbito do PIBIC, está vinculada ao projeto 'Compilação de um corpus representativo do português do Brasil e análise multidimensional da variação entre gêneros discursivos: CORPOBRAS PUC-Rio', em andamento com apoio do Edital Universal CNPq n°019/2004[2]. A pesquisa compreende a coleta de textos de gêneros do discurso oral, escrito e escrito para ser falado, visando a ampliação do corpus e sua disponibilização através de um banco de dados *on-line*. Nesta etapa do projeto, foram acrescentados ao corpus 8 (oito) gêneros discursivos, totalizando 27 (vinte e sete gêneros no corpus, assim distribuídos: 20 (vinte) gêneros do discurso escrito, 5 (cinco) gêneros do discurso oral, e 2 (dois) gêneros do discurso escrito para ser falado. A análise de textos do gênero '' também foi desenvolvida nesta fase do projeto, sob a perspectiva da Lingüística Sistêmico-Funcional (Halliday, 1994), enfocando o estudo da metáfora gramatical na escrita acadêmica.

## Objetivos

Um dos objetivos desta fase do projeto foi a ampliação do CORPOBRAS, prioritariamente, através da seleção e da compilação de textos do discurso oral, o que possibilitou que o corpus atingisse a meta de 1.000.000 (hum milhão) de palavras.

Outro objetivo do projeto concentrou-se na 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) e desenvolvida a análise do uso de nominalizações, a fim de verificar como seu uso influencia a escrita de textos acadêmicos.

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

## Aspectos Teórico – Metodológicos

Nesta etapa do projeto, foram compilados 8 novos gêneros: conversa carioca, conversa de crianças, grupos de enfoque, entrevistas, atendimento ao cliente, cartas ao editor, dissertações de mestrado e teses de doutorado (compilação em andamento). Os textos compilados foram coletados em fontes diversas e alguns deles tiveram que ser escaneados, revistos ou redigitados. Alguns 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).

---

[1] Márcia Gonzaga de Brito é bolsista FAPERJ e Rubiane Guilherme Valério é bolsista PIBIC/CNPq.
[2] Processo 480143/2004-8

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 um código 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. A coluna 'Código' da Figura 1, a seguir, ilustra o tipo de codificação utilizada no corpus:

**Figura 1 – Relatório do Gênero 'Discurso Político'**

| Código | Tema | Fonte | Cargo | Nome | Ano | Nº. de palavras |
|---|---|---|---|---|---|---|
| DISPOL17PORT1 | 1° de Maio | Tese da Professora Del Carmen (UERJ) | Presidente | Getúlio Vargas | 1954 | 2000 |
| DISPOL18PORT1 | 1° de Maio | Tese da Professora Del Carmen (UERJ) | Presidente | JK | 1960 | 1112 |
| DISPOL19PORT1 | 1° de Maio | Tese da Professora Del Carmen (UERJ) | Presidente | Jango | 1963 | 1524 |
| DISPOL20PORT1 | 1° de Maio | Tese da Professora Del Carmen (UERJ) | Presidente | Castelo Branco | 1966 | 3502 |
| DISPOL21PORT1 | 1° de Maio | Tese da Professora Del Carmen (UERJ) | Presidente | Costa e Silva | 1968 | 1130 |

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, para cada gênero do corpus, relatórios em forma de tabelas, constando o código, o número total de palavras e informações relevantes para cada gênero, conforme Figura 2, abaixo:

**Figura 2 - Relatório do Gênero 'Conversa Carioca'[3]**

| Código | Tema | Sexo | Idade | Profissão | Zona residencial | Duração da conversa | Data do registro | Total de palavras |
|---|---|---|---|---|---|---|---|---|
| CONCAR1PORT1 | Alimentação | Masculino | 29 anos | Professor de biologia | Suburbana | 47 min. | 26 de setembro de 1972 | 7213 |
| CONCAR2PORT1 | Alimentação | Feminino | 30 anos | Advogada | Sul | 46 minutos | 14 de agosto de 1972 | 6526 |
| CONCAR3PORT1 | Alimentação | Masculino | 44 anos | Professor de desenho | Suburbana | 52 minutos | 23 de agosto de 1972 | 8435 |
| CONCAR4PORT1 | Alimentação | Feminino | 37 anos | Professora de psicologia | Suburbana | 43 minutos | 12 de julho de 1976 | 7470 |
| CONCAR5PORT1 | Alimentação | Masculino | 55 anos | Administração pública | Suburbana | 48 minutos | 25 de setembro de 1972 | 6830 |
| CONCAR6PORT1 | Alimentação | Feminino | 44 anos | Professora de filosofia | Sul | 43 minutos | 18 de outubro de 1971 | 5511 |
| CONCAR7PORT1 | Alimentação | Masculino | 57 anos | Dentista | Suburbana | 52 minutos | 11 de julho de 1972 | 10252 |

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, tendo como modelo interfaces de bancos de dados on-line internacionais, como, o V*iew,* que possibilita a seleção de textos para análise de acordo com múltiplos parâmetros, conforme ilustrado na Figura 3, a seguir:

[3] Dados originalmente coletados pelo Projeto NURC.

**Figura 3 – Interface do *View***


Em relação à análise lingüística de textos acadêmicos desenvolvida nesta fase do projeto, foram selecionadas 3 diferentes versões de introduções e conclusões de dissertações de mestrado na área de Estudos da Linguagem, escritas em momentos diferentes do desenvolvimento do texto de seis alunas, sendo uma versão a inicial, outra intermediária, ou seja, produzida próxima ao final do processo de escrita da dissertação, e a terceira, que corresponde à versão final da introdução ou conclusão de cada trabalho. Nestas amostras, os sufixos que indicam nominalizações foram identificados com o auxílio da ferramenta *Concordance*, um gerador de linhas de concordância do software Wordsmith Tools, o qual disponibiliza listas de palavras a partir da indicação de itens indicados para buscas em contexto. Nesta pesquisa foram geradas listas de palavras através da busca de sufixos: -ção, -ssão, -mento, -cia, -dor e seus respectivos plurais. O uso destes sufixos foi observado já que eles são formadores de nominalizações as quais podem indicar processos de transformação de idéias mais concretas em mais abstratas, através de metáforas gramaticais.

## Resultados

Com a compilação de 8 novos gêneros, o corpus ultrapassou a marca de 1.000.000 (hum milhão) de palavras, equiparando-se a corpora internacionais como o Brown Corpus, o LOB e London Lund Corpus.

**Figura 4 – Distribuição dos gêneros e textos do CORPOBRAS**

| **CORPOBRAS PUC-Rio 2007** | | |
|---|---|---|
| **Língua Portuguesa-L1** | | |
| **Discurso Escrito** | | |
| **Gêneros** | **Número de Textos** | **Número de palavras** |
| Artigos científicos | 12 | 63.818 |
| Cartas ao editor | 18 | 1.054 |
| Cartas de reclamação | 136 | 21.417 |
| Cartas de recomendação | 31 | 6.012 |
| Cartas pessoais | 16 | 7.829 |
| Cartas profissionais | 16 | 3.166 |
| Cartas profissionais acadêmico | 15 | 3.529 |
| Circulares | 16 | 2.608 |
| Contos | 14 | 15.253 |
| Crônicas | 26 | 17.434 |
| Dissertações de mestrado | | em andamento |
| Editoriais | 16 | 7.931 |
| E-mails acadêmicos | 15 | 1.816 |
| E-mails pessoais | 16 | 1.858 |
| Notícias de jornal | 99 | 40.409 |
| Redações de alunos | 16 | 3.416 |
| Redações de alunos universitários | 91 | 25.065 |
| Redações de vestibular | 139 | 28.646 |
| Romances | 28 | 27.061 |
| Teses de doutorado | | em andamento |
| | **Total de Palavras:** | **278.322** |
| **Discurso Oral** | | |
| Conversas cariocas | 53 | 353.678 |
| Conversas de crianças | 94 | 84.573 |
| Entrevistas (acadêmicas) | 17 | 88.769 |
| Grupos de enfoque | 7 | 40.513 |
| Atendimento ao cliente | 393 | 215.671 |
| | **Total de Palavras:** | **783.204** |
| **Discurso Escrito para ser Falado** | | |
| Discursos políticos | 27 | 22.751 |
| Roteiros cinematográficos | 18 | 17.180 |
| | **Total de Palavras:** | **39.931** |
| **Total de Palavras no Corpus em 2007** | | **1.101.457** |

Quanto à análise do gênero 'dissertações de mestrado', ainda em fase de desenvolvimento, observou-se que durante o processo de escrita e re-escrita houve uma variação na freqüência de nominalizações, e nas últimas versões das introduções e das conclusões verificou-se que ocorrem maiores índices de uso de sufixos indicadores de nominalizações. Além disso, parece haver maior utilização desse recurso de formação de palavras com a ausência de marcação do plural, ou seja, no singular, à medida que os textos vão sendo re-escritos. Essa recorrência pode sinalizar maior grau de abstração dos textos, uma vez que em português palavras no singular podem indicar maior generalização.

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

O projeto gerou diversos trabalhos neste último período, que foram apresentados pela orientadora em publicações e congressos nacionais e internacionais (Oliveira, 2006; 2007a; 2007b). Ainda como resultado desse projeto, foi submetido, com participação das bolsistas, em 2007, um resumo para o Encontro da Ciência Empírica da Literatura (ECEL), na Universidade Federal do Rio de Janeiro.

Nesta etapa do projeto, visando maior embasamento teórico para a pesquisa, as bolsistas aprofundaram seus estudos frequentando as aulas de Pós-Graduação nas disciplinas 'Lingüística de Corpus' (2006.2), 'Lingüística Aplicada' (2007.1) e 'Linguística Sistêmico-Funcional'(2007.2), ministradas pela orientadora. Ainda em relação ao aprofuntamento da pesquisa, houve o cadastro em um novo Grupo de Pesquisa junto ao Diretório de Grupos de Pesquisa do CNPq: Gr Pesq. 'Lingüística Sistêmico-funcional, lingüística de corpus e análise do discurso', do qual as bolsistas participam. Visando a análise lingüística dos dados do corpus com apoio de ferramentas computacionais, a orientadora e as bolsistas freqüentaram um curso de extensão sobre 'Wordsmith Tools', oferecido na PUC-Rio (dezembro, 2006).

## Referências


BIBER, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics.* Cambridge: Cambridge University Press.

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*

HALLIDAY, M.A.K (1994). *An introduction to functional grammar*. London: Edward Arnold. 2ª ed.

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.

OLIVEIRA, L. P. (2006). *Grammatical metaphor in research articles: Linguistic and disciplinary contrasts*. Trabalho apresentado na American Association for Applied Linguistics and the Canadian Association for Applied Linguistics Conference (AAAL/CAAL), Montreal, Canada.

OLIVEIRA, L. P. (aceito para publicação; 2007a). *Involvement variation in the writing of academics: A cross-cultural analysis of three genres*. International Journal of Corpus Linguistics. Amsterdam: John Benjamins.

OLIVEIRA, L.P. (aceito para publicação; 2007b). *Cross-cultural contrasts in discourse styles: Complexity-level variation in writing. Languages in Contrast*. Amsterdam: John Benjamins.

SIMON-VANDENBERGEN, A., Taverniers, M & Ravelli, L. (Eds.) (2003). *Grammatical metaphor: Views from Systemic functional linguistics*. John Benjamins: Amsterdam.