

# **Projeto: Compilação de um corpus representativo do português do Brasil e análise multidimensional da variação entre gêneros discursivos**

**Nome do Proponente:** Lúcia Pacheco de Oliveira

**Instituição proponente:** Pontifícia Universidade Católica do Rio de Janeiro

**Financiamento:** Conselho Nacional de Desenvolvimento Científico

## **Resumo**

O problema:

Análises empíricas de grandes corpora lingüísticos com o auxílio de ferramentas computadorizadas têm sido desenvolvidas na área de Lingüística de Corpus. Para que estes estudos sejam conduzidos, é preciso que grandes quantidades de dados lingüísticos sejam compilados e sistematicamente organizados, para serem posteriormente analisados com o auxílio de ferramentas computadorizadas (Biber, Conrad & Reppen, 1998). A montagem de um corpus representativo de uma língua requer o armazenamento de amostras de vários gêneros do discurso oral e escrito. Apesar de terem sido tomadas algumas iniciativas isoladas para a compilação de corpora com textos em português em alguns centros acadêmicos, até o momento, ainda não contamos com um corpus de dimensões abrangentes, que seja representativo do Português do Brasil e organizado de acordo com convenções aplicadas internacionalmente. Pesquisas desenvolvidas no Mestrado e Doutorado na PUC-Rio, na área em Estudos da Linguagem (Lanziotti, 2002, Almeida, 2002, Amarante, 2002), e IBIC/CNPq, possibilitaram a coleta de um corpus que hoje conta com cerca de 500.000 palavras. Entretanto, se comparada a outros corpora tais como British National Corpus (BNC), American National Corpus (ANC), Longman-Lancaster Corpus, COBUILD/Birmingham Corpus, Helsinki Corpus, esta amostra é ainda relativamente pequena, e deverá ser ampliada para tornar-se mais representativa do Português do Brasil e servir de base para novas pesquisas lingüísticas.

**Objetivos:** Visando a organização, disponibilização e análise de um corpus lingüístico representativo do português do Brasil a ser utilizado para pesquisas lingüísticas e futuramente para a produção de materiais, este projeto apresenta os seguintes objetivos: a) ampliar o corpus atual através da compilação e incorporação ao corpus de novos textos do discurso oral e escrito, para atingir a meta inicial de 1.000. 000 (um milhão ) de palavras, tal como importantes corpora (Brown Corpus, LOB e London-Lund Corpus); b) organizar e codificar o corpus seguindo normas internacionais já estabelecidas para a montagem de corpora ; c) disponibilizar o corpus em CD-ROM e via Internet; d) analisar a variação entre gêneros discursivos do corpus de acordo com a abordagem multidimensional.

**Metodologia:** Os dados a serem incorporados ao corpus deverão ser escaneados, digitados ou transcritos dependendo da situação, local ou meio em que forem coletados. Cada texto deverá também ser codificado de acordo com convenções internacionais e todos os textos deverão ser gravados em CD ROM e o corpus disponibilizados na internet. A coleta de novos textos para serem incorporados aos gêneros que já fazem parte do corpus será feita visando-se padronizar o número de textos em cada gênero. Também serão incorporados ao corpus novos gêneros discursivos para aumentar sua representatividade e tornar sua distribuição comparável a corpora em outras línguas (Biber,1995). Os dados serão analisados de acordo com a Análise Multidimensional (Biber, 1988), que identifica parâmetros de variação, interpretados como dimensões, ao longo das quais os

textos e os gêneros variam. Esta metodologia permite caracterizar as semelhanças e diferenças entre textos e gêneros discursivos.

Resultados: Este projeto contribuirá para os estudos da linguagem através da compilação de um corpus que possibilitará novos estudos sobre o discurso oral e escrito em português. A descrição mais ampla da português em uso no Brasil possibilitará o desenvolvimento de pesquisas teóricas e aplicadas sobre a nossa língua materna. Os resultados da pesquisa contribuirão diretamente para duas áreas em crescimento no Brasil: Estudos de Gêneros e Lingüística de Corpus.

## 1. Caracterização do Problema

---

A Lingüística de Corpus caracteriza-se pela análise empírica de grandes corpora lingüísticos com o auxílio de ferramentas computadorizadas. Para que estudos baseados em corpora sejam conduzidos, é preciso que grandes quantidades de dados lingüísticos sejam compilados e sistematicamente organizados, para serem posteriormente analisados com o auxílio de ferramentas computadorizadas (Biber, Conrad & Reppen, 1998, Sardinha, 2000).

Estudos baseados em corpora buscam identificar e analisar, em diferentes línguas, padrões de uso em textos que ocorrem naturalmente na língua. Estes estudos têm investigado traços lingüísticos ou características de variedades lingüísticas, estando inseridos em diferentes áreas como a lexicografia, léxico-gramática, estudos de gêneros discursivos e variação lingüística diacrônica e sincrônica. Estes estudos têm também trazido um maior aprofundamento sobre o conhecimento empírico da língua em uso, bem como novas concepções teóricas sobre as línguas estudadas. Além disso, a existência de corpora lingüísticos tem possibilitado a geração de produtos como dicionários (COBUILD Dictionary) e novas gramáticas descritivas (Biber, Johansson, Leech, Conrad & Finegan, 1999), bem como possibilitado uma ampla gama de pesquisas lingüísticas, qualitativas e quantitativas, desenvolvidas a partir dos dados da língua em uso.

A montagem de um corpus representativo de uma língua requer o armazenamento de amostras de vários gêneros do discurso oral e escrito. Apesar de terem sido tomadas algumas iniciativas isoladas para a compilação de corpora com textos em português em alguns centros acadêmicos portugueses e brasileiros, como por exemplo a USP, NILC/São Carlos, PUC-SP, até o momento, ainda não contamos com um corpus de dimensões abrangentes, que seja representativo do Português do Brasil e organizado de acordo com convenções aplicadas internacionalmente como em vários outros corpora: American National Corpus (ANC), Michigan Corpus of American Spoken English (MICASE), The English-Norwegian Corpus (ENPC) ou Cambridge and Nottingham Corpus for the Description of English (CANCODE).

No Brasil, onde a pesquisa lingüística tem se desenvolvido com muita rapidez, e onde já há uma grande quantidade de dados coletados e já informatizados, parece-me importante que seja organizado e sistematizado um corpus do português oral e escrito, que poderá servir de base tanto a estudos lingüísticos teóricos como aplicados. Um projeto deste porte envolveria, entre outros, profissionais das áreas de lingüística, informática, ensino de línguas, etc. Este esforço conjunto e interdisciplinar poderá resultar em um maior conhecimento da língua portuguesa e dos gêneros do discurso em português. A coleta de dados proposta neste projeto poderá ser considerada como uma contribuição importante para o desenvolvimento de um Corpus do Português do Brasil.

## 2. Objetivos

---

Visando a organização, disponibilização e análise de um corpus representativo do Português do Brasil a ser utilizado para pesquisas lingüísticas e produção de materiais, este projeto apresenta os seguintes objetivos:

- a) ampliação de um corpus já existente através da compilação e incorporação ao corpus de novos textos e gêneros do discurso oral e escrito, para atingir a meta inicial de 1.000.000 (um milhão) de palavras, tal como importantes corpora (Brown Corpus, Lancaster-Oslo/Berger (LOB) e London-Lund Corpus);
- b) organizar e codificar o corpus seguindo normas internacionais já estabelecidas para a montagem de corpora;
- c) disponibilizar o corpus em CD-ROM e via Internet.
- d) analisar a variação entre gêneros discursivos do corpus de acordo com a abordagem multidimensional.

Em fases anteriores da pesquisa de corpus desenvolvida na PUC-Rio pela Coordenadora deste projeto, vários gêneros do discurso oral e escrito foram coletados tais como redações escritas por alunos universitários (Oliveira,1997), artigos acadêmicos (Oliveira,1999), cartas de recomendação, cartas profissionais e e-mails (Oliveira, 2001). Mais recentemente este corpus foi ampliado, incluindo gêneros orais, tais como atendimentos de serviços (*Check-in* de aeroporto, atendimento de balcão de companhia de saúde), entrevistas acadêmicas (com professores universitários de escola pública e particular, bem como entrevistas com alunos destas escolas), conversas face a face (com representantes de diferentes cidades do Brasil: Rio de Janeiro, Natal e Fortaleza) e reuniões de negócios (setor de vendas, imobiliário e publicitário). Gêneros do discurso escrito também foram coletados, como cartas de reclamação, cartas pessoais, editoriais, notícias de jornais, crônicas, romances, peças de teatro e roteiros de cinema.

Estas pesquisas, desenvolvidas no âmbito do PIBIC-CNPq, com auxílio de bolsistas de graduação, e por alunos de Mestrado e Doutorado em suas Dissertações e Teses (Lanziotti, 2002, Almeida, 2002, Amarante, 2002, Moraes (Tese em processo de finalização), possibilitaram a compilação de textos para formar um corpus de português, na PUC-Rio. Este corpus hoje conta com cerca de 500.000 palavras, distribuídas em 24 gêneros, sendo 19 gêneros do discurso escrito e 5 do discurso oral. Entretanto, se comparado a outros corpora existentes, em outras línguas, tais como o British National Corpus (BNC), Longman-Lancaster Corpus, COBUILD/Birmingham Corpus, Helsinki Corpus, esta amostra é ainda pequena, e deverá ser ampliada para tornar-se mais representativa do Português do Brasil e servir de base para novas pesquisas lingüísticas e produção de materiais.

A exemplo de estudos baseados em corpora desenvolvidos em outras línguas (Biber, 1988, 1995), pesquisas realizadas a partir do Corpus de Português do Brasil permitirão dar continuidade à análise da variação entre gêneros do discurso em português, buscando-se uma descrição abrangente e comparativa do discurso oral e escrito, com base na língua efetivamente em uso em diferentes contextos educacionais e profissionais (Oliveira, 2002).

### **3. Metodologia**

---

Para a coleta do corpus é preciso considerar-se a representatividade das amostras selecionadas. Por esta razão, os textos a serem incluídos em um corpus representativo do português do Brasil devem, entre outras características, ser: autênticos, refletindo a língua em uso; produzidos por falantes nativos da língua, ou seja brasileiros; produzidos por falantes/escritores únicos, ou seja, cada texto deve ser de um autor/participante

diferente; produzidos em diferentes regiões do país, para representar a variedade regional de forma abrangente; selecionados de forma não aleatória, tendo conteúdo variado; pertencentes a diferentes gêneros discursivos, para representar a maior variedade possível de ações sociais.

Para a organização do corpus, cada texto deve ser identificado em termos das características acima (ex: <www.hti.umich.edu/m/micase/browse.html>. Cada texto é identificado através de uma sigla que indica o gênero e a língua em que o texto foi produzido; os textos são numerados, de maneira ininterrupta, por gênero. Essa organização do corpus facilita a catalogação, tornando-a consistente e unificada para facilitar a pesquisa e a análise dos textos. Os textos, classificados por língua e por gênero, são gravados em arquivos de diferentes formatos. Arquivos gravados em WORD(.doc) são destinados à pesquisa e permitem a leitura direta dos textos. Os arquivos gravados no formato do Texto (.txt) são destinados à análise computacional, visto que algumas ferramentas usadas para esse tipo de análise lêem somente arquivos gravados neste .txt. Os dados são posteriormente armazenados em CD-ROM para possibilitar a arquivagem de uma maior quantidade de dados, de maneira mais segura. Alguns corpora apresentam também uma versão etiquetada, com identificação de classes gramaticais, e uma versão não etiquetada.

Para a análise dos dados diferentes metodologias podem ser utilizadas, já que os estudos baseados em corpora não dispõem de uma metodologia própria e específica, o que tem gerado o aparecimento de diferentes abordagens metodológicas que visam ajudar a melhor acessar, analisar e contrastar corpora lingüísticos, havendo uma ampla gama de metodologias e programas de computador que podem auxiliar nesta tarefa. Dentre as metodologias existentes e produtivas, aparece a Análise Multidimensional (Biber, 1988), capaz de caracterizar a variação lingüística em grandes corpus de dados, com o auxílio de medidas estatísticas. É este tipo de metodologia que será utilizado para estudar a variação dos gêneros discursivos neste projeto.

Visando um estudo da variação lingüística na língua oral e escrita, Biber (1988) utilizou a Análise Multidimensional capaz de analisar um grande corpus de dados (900.000 palavras), composto de diversos gêneros (N=23), através de múltiplos parâmetros de variação a que denominou 'dimensões'. As dimensões são definidas através do agrupamento de traços lingüísticos que co-ocorrem com frequência nos textos. Estas dimensões são identificadas estatisticamente através de um teste estatístico, a Análise Fatorial, e interpretadas de acordo com a função comunicativa compartilhada pelos traços que co-ocorrem nos textos.

A abordagem multidimensional tem base funcional na medida em que considera que os traços lingüísticos têm uma função como marcadores de uma situação, ou seja, atuam para distinguir diferentes aspectos da situação de comunicação (Hymes, 1974, Halliday e Hasan, 1989, Halliday, 1994, Biber,1988). Para identificar as dimensões textuais no corpus de dados, medidas estatísticas são utilizadas para reconhecer os agrupamentos de traços lingüísticos, mostrando que sua co-ocorrência não se dá de forma randômica. Parece então apropriado interpretar por que certos traços se agrupam em textos, ou seja, verificar que função subjacente influencia o seu uso.

As dimensões constituem escalas contínuas de variação, em vez de pólos dicotômicos (Biber,1988). Um texto é muitas vezes classificado como: interativo/não-interativo, formal/ informal, elaborado/não-elaborado, etc. Entretanto, os textos podem variar quanto ao grau de interação, formalidade ou elaboração que apresentam, devendo cada uma destas categorias ser considerada ao longo de um contínuo. Ao longo deste contínuo, alguns textos serão vistos como, por exemplo, mais interativos ou menos

interativos, mais formais ou menos formais ou como mais elaborados ou menos elaborados. Nos extremos deste contínuo, estarão os pólos "interativo/não-interativo", "formal/informal", "elaborado/reduzido", etc. Esta noção é quantitativa e permite a descrição da variação de textos ou gêneros textuais ao longo deste contínuo.

Vários pesquisadores tentaram analisar a variação textual através de dimensões (Chafe,1982), especialmente os sociolinguistas, que buscam verificar a distribuição de traços lingüísticos em diferentes grupos sociais ou situações, mas geralmente enfocando apenas uma dimensão. Entretanto, "uma única dimensão não será adequada em si mesma para dar conta da gama de variações na língua; em vez disto, uma análise multidimensional é necessária" (Biber, 1988), para que, através da identificação e combinação de várias dimensões simultaneamente, se possa ver como os textos se caracterizam e como, e até que ponto, eles se aproximam ou se distanciam uns dos outros, levando a uma descrição mais completa da variação lingüística.

### Abordagem multidimensional

#### ***Suporte teórico:***

- conceitos de *variação e relação*:

A variação lingüística engloba várias dimensões, e as relações entre textos são definidas pela comparação das dimensões, que mostram como, e até que ponto, os textos se aproximam ou se distanciam;

- base *situacional e funcional*:

O contexto situacional para os atos de fala é considerado como base para análise das dimensões. Biber (1988:29) distingue oito componentes da situação de fala: papéis e características dos participantes; relações entre os participantes; cenário; tópico; propósito; avaliação social; relações dos participantes com o texto; canal.

A abordagem multidimensional tem base funcional na medida em que considera que os traços lingüísticos têm uma função como marcadores de uma situação, ou seja, atuam para distinguir diferentes aspectos da situação de comunicação (Hymes, 1974, Halliday e Hasan, 1989, Halliday, 1994, Biber,1988). Biber (1988:34 ) distingue sete funções: ideacional, textual, pessoal, interpessoal, contextual, processual, estética.

#### ***Análises preliminares:***

- seleção de classes de traços lingüísticos;
- identificação (semi-automática) dos traços lingüísticos nos textos;
- cálculo da frequência dos traços lingüísticos em cada variável.

#### ***Análises estatísticas:***

##### Macroanálise:

- normatização das frequências dos traços lingüísticos;
- Análise Fatorial para identificação da co-ocorrência de traços lingüísticos no corpus. Método de fatoração: 'Principal Components Analysis'. Rotação oblíqua; (cálculos estatísticos com auxílio de software específico: SPSS – Statistical Package for Social Sciences).
- identificação dos Fatores.

### Microanálise:

- interpretação das co-ocorrências das variáveis nos Fatores como Dimensões Textuais, de acordo com sua função discursiva.

### ***Análise comparativa da variação:***

- standardização das freqüências normatizadas, considerando-se a média e o desvio padrão;
- cálculo de escores correspondentes a cada Fator/ Dimensão; comparação das dimensões textuais em diferentes gêneros do discurso.

Muitas pesquisas já foram desenvolvidas utilizando a abordagem multidimensional, dentre elas: Biber, 1988, 1995; Grabe, 1987; Lux e Grabe, 1991; Oliveira, 1997, 1999, 2001; Sardinha, 2000; Lanziotti, 2002.

### **Estratégias de Ação:**

Localização do projeto: Em sala da Instituição, com três postos de trabalho para as pesquisadoras e apoio técnico. Neste local serão colocados os equipamentos solicitados.

Estrutura do projeto: O projeto se desenvolverá na PUC-Rio e receberá visitas dos colaboradores, em missões de trabalho. As pesquisadoras brasileiras também irão a outros centros de pesquisa em estudos baseados em corpora.

- Coordenadora – Pesquisadora responsável pela equipe
- Núcleo de compilação: 3 pesquisadoras – bolsista PIBIC; mestranda; doutoranda.
- Núcleo de informática: 1 apoio técnico e um colaborador de informática
- Núcleo de análise: 3 pesquisadoras – doutoras ou doutorandas.
- Núcleo de colaboradores: pesquisadores visitantes de outras instituições brasileiras ou estrangeiras.

## **4. Outros Projetos Financiados Atualmente** (sugerido: máximo de 1 página)

---

Os projetos que envolvem a compilação de corpora, em vários países, têm sido financiados por instituições acadêmicas, juntamente com editoras e outras empresas interessadas. Um exemplo é o American National Corpus, atualmente ainda em fase de desenvolvimento, em torno do qual formou-se um consórcio do qual participam as Universidades de Northern Arizona e da Pensilvânia, bem com editoras e a Microsoft. Em outros países, o governo tem financiado projetos, como em Portugal, através do seu Ministério de Ciência e Tecnologia. No Brasil, alguns centros de pesquisa têm sido contemplados com financiamentos para a coleta de corpora, mas estes são muitas vezes focados em gêneros específicos do discurso.

Para o projeto aqui proposto, que visa a compilação e análise de um corpus representativo do português do Brasil, não foi solicitado financiamento a nenhuma outra agência de fomento até o momento, o que será feito durante o próximo ano. Também pretendemos sensibilizar a iniciativa privada para o projeto, uma vez que, para editoras,

por exemplo, o corpus é de grande importância para a produção de dicionários, itens altamente vendáveis, ou de gramáticas. A pesquisa sobre corpus, na PUC-Rio, tem contado a colaboração das seguintes bolsistas PIBIC/ CNPq, orientandas da Coordenadora do projeto: Dafne Malheiros Baddini (2004-2005); Tacila Moura (2002-2003); Clarissa dos Santos Soares (2001-2002).

## 5. Principais Referências Bibliográficas (sugerido: máximo de 20 referências)

---

### BIBLIOGRAFIA

- Amarante, R. M. C. (2002). Começando do princípio: Uma análise do *lead* como subgênero discursivo em português e em inglês. Dissertação de Mestrado, Estudos da Linguagem, PUC-Rio.
- Almeida, P.M.C.(2002). Atendimento de *check-in* de companhia aérea: Análise sistêmico-funcional de um gênero discursivo do português. Dissertação de Mestrado, Estudos da Linguagem, PUC-Rio.
- Biber, Douglas (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, Douglas (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D. , Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Essex, England: Pearson Education Limited.
- Chafe, W.(1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: exploring orality and literacy*. Norwood, New Jersey: Ablex.
- Conrad, S. & Biber, D. (Eds.) (2001). *Variation in English: Multi-Dimensional Studies*. London/NewYork: Longman.
- Grabe, W. (1987). Contrastive Rhetoric and Text Type Research. In U. Connor and R. Kaplan (Eds.), *Writing Across Languages: Analysis of L2 Texts*, pp. 113-137. Reading, MA: Addison-Wesley.
- Halliday, M. A. K.(1994). *An Introduction to Functional Grammar*. London: Edward Arnold. 2ª edição.
- Halliday, M. A.K. e Hasan, R. (1989). *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective*. Oxford: Oxford University Press.
- Hymes, Dell (1974). *Foundations in Sociolinguistics*. Philadelphia: University of Pennsylvania Press.
- Lanziotti, M.G. P. (2002). Variação de gêneros discursivos: A explicitação do contexto em um corpus do português escrito. Dissertação de Mestrado, Estudos da Linguagem, PUC-Rio.

- Lux, P. e Grabe, W. (1991). Multivariate Approaches to Contrastive Rhetoric. *Linguas Modernas*, 18, 133-160.
- Oliveira, Lúcia P. (1997). Variação Intercultural na Escrita: Contrastes Multidimensionais em Inglês e Português. Tese de Doutorado, LAEL/PUC-SP, São Paulo.
- Oliveira, Lúcia P. (1999). Cross-cultural complexity-level variation in written discourse styles. Trabalho apresentado na American Association for Applied Linguistics Annual Conference (AAAL), Stamford, Connecticut.
- Oliveira, Lúcia P. (2001). Cross-linguistic and cross-genre involvement variation in the writing of academics. Trabalho apresentado na Conferência Anual da American Association for Applied Linguistics, Saint Louis, EUA.
- Oliveira, L. P. (2002). Explicitação do contexto em textos de alunos brasileiros e americanos. *Palavra*, 8, 102-116.
- Sardinha, T. B. (2000a). Análise multidimensional. *D.E.L.T.A.*, 16 (1), 99-127.
- Sardinha, T. B. (2000b). Lingüística de Corpus: Histórico e problemática. *D.E.L.T.A.*, 16 (2), 323-367.

//