Calidoscópio
Vol. 7, n. 3, p. 192-198, set/dez 2009

Lúcia Pacheco de Oliveira
luciapo@openlink.com.br

Maria Carmelita Padua Dias
carmelitapdias@gmail.com

Compilação de corpus: representatividade e o CORPOBRAS[1]

Corpus compilation: Representativeness and the CORPOBRAS

RESUMO – O objetivo deste trabalho é discutir a importância do parâmetro da representatividade no design e compilação de corpus, mostrando a sua relevância para estudos que visem desenvolver uma descrição abrangente da língua. Este trabalho também apresenta um corpus do português do Brasil, o CORPOBRAS, em desenvolvimento, cujo parâmetro norteador tem sido a representatividade. Este corpus compõe-se, atualmente, de 27 gêneros do discurso oral, discurso escrito e discurso escrito para ser falado. Com a finalidade de ilustrar possíveis usos do CORPOBRAS, no final do trabalho, são listadas algumas pesquisas que utilizaram dados do corpus em suas análises.

Palavras-chave: CORPOBRAS, corpus do português do Brasil, linguística de corpus, variação entre gêneros discursivos, representatividade, discurso oral e escrito.

ABSTRACT – This paper discusses an important parameter in corpus design and compilation: representativeness. This parameter is related to the need to include in corpora texts that represent several uses of the language so that comprehensive descriptions can be developed. The paper also presents a corpus of Brazilian Portuguese – CORPOBRAS – that comprises 27 discourse genres and is guided by the representativeness parameter. The paper finally lists several corpus-based studies that draw upon CORPOBRAS data.

Key words: CORPOBRAS, corpus linguistics, genre variation, representativeness, oral and written discourse.

Introdução

Este artigo trata da compilação de um corpus de português do Brasil, tendo como objetivo principal, na coleta dos dados, a representatividade para estudos de gêneros discursivos. Insere-se, assim, dentro da área de Linguística de Corpus, que propõe uma nova maneira de olhar a linguagem e de fazer pesquisa, ou seja, a linguagem é vista como um fenômeno social e analisada a partir de atos concretos de comunicação (Teubert, 1996, p. vi), isto é, textos reais, sendo a pesquisa desenvolvida com o auxílio de ferramentas computacionais e buscando extrair evidências linguísticas do corpus. Outra característica da Linguística de Corpus é que ela pode trazer contribuições tanto para pesquisas teóricas, através de novas descrições linguísticas, como para pesquisas aplicadas, encontrando-se em interface com outras áreas do conhecimento que, como ela, buscam o significado onde ele é negociado, ou seja, no discurso (Oliveira, 2009), tais como a Linguística Aplicada (Grabe, 2004; Hunston, 2002; Kaplan, 2002) e a Linguística Sistêmico-Funcional (Thompson e Hunston, 2006).

O desenvolvimento da área de Linguística de Corpus pode ser relacionado, inclusive no Brasil, às novas perspectivas que possibilita em relação à lexicografia, incluindo a elaboração de dicionários, à léxico-gramática, incluindo o desenvolvimento de gramáticas a partir de corpora (Biber *et al.*, 1999; Carter e McCarthy, 2006; Neves, 1999), aos estudos da variação linguística em gêneros discursivos (Conrad e Biber, 2001) e estudos interculturais (Oliveira, 1997). A Linguística de Corpus é, portanto, uma área que permite o aprofundamento sobre o conhecimento empírico de diferentes línguas, a partir de diferentes paradigmas, tanto o funcionalista (Halliday e Matthiessen, 2004; Biber, 1995) como o gerativista (Tycho Brahe, 2008), podendo levar a novas concepções teóricas

[1] Este texto é a uma versão revista, modificada e atualizada de um trabalho apresentado durante a jornada 'Metodologia sobre a Compilação e Sistematização de Corpus', realizada na PUC-Rio, em 2006, e divulgado na internet pela União Latina, organizadora daquele evento.

sobre a linguagem e podendo ser considerada como "a face moderna da linguística empírica" (Teubert, 1996, p. vi).

Cada vez mais os estudos linguísticos teóricos e aplicados vêm se beneficiando do uso de *corpora* para a descrição de fenômenos linguísticos ou para a verificação de hipóteses acerca dos mesmos. Para que esses estudos sejam conduzidos, é preciso que grandes quantidades de dados linguísticos sejam compiladas e sistematicamente organizadas em corpora, para serem posteriormente analisadas com o auxílio de ferramentas computacionais (Biber *et al.*, 1998; Sardinha, 2004). Apesar das facilidades trazidas pela tecnologia quanto ao armazenamento, acesso e análise de dados, nos estudos de corpus ainda persistem algumas questões relativas aos parâmetros que devem ser seguidos na compilação de um *corpus.*

Neste trabalho, tratamos de um desses parâmetros: a representatividade, que está ligada a três aspectos básicos: (a) todos os usos da língua, ou uma boa parte deles, devem ser contemplados na compilação do material que compõe o *corpus,* para que as análises possam ter um caráter amplo; (b) a predominância de algum aspecto linguístico específico no *corpus* – como um tipo de discurso, um estilo textual, ou um dialeto regional – pode acarretar um viés determinado na análise dos dados ou desvio nos resultados; (c) a inclusão de amostras do discurso oral no *corpus*, ainda que mais raras e de difícil coleta, é indispensável para caracterizar mudanças e usos linguísticos.

Representatividade e o *corpus*

Estudos baseados em *corpora* buscam identificar e analisar, em diferentes línguas, padrões de uso em textos que ocorrem naturalmente na língua. Esses estudos têm investigado traços linguísticos ou características de variedades linguísticas e têm contribuído com um maior aprofundamento sobre o conhecimento empírico da língua em uso, bem como novas concepções teóricas acerca da língua estudada.

A montagem de um corpus representativo de uma língua requer o armazenamento de amostras de vários gêneros do discurso oral e escrito[2]. Várias iniciativas bem-sucedidas para a compilação de corpora em português vêm sendo tomadas ao longo dos últimos anos, com finalidades diversas, em diferentes regiões do país e no exterior. Dentre esses corpora podemos citar alguns, tais como: o corpus do NILC – Núcleo Interinstitucional de Linguística Computacional (USP/UFScar/UNESP), criado em 1993 para o desenvolvimento de pesquisas e projetos na área de linguística computacional e processamento de linguagem natural; o corpus da Linguateca (2009), composto inicialmente por uma coleção de textos de português europeu, visando o processamento computacional da língua portuguesa, e agora também incorporando os textos do NILC; o corpus da PUC-SP, contendo textos de comunicação no contexto de negócios (Projeto DIRECT, 2005); o corpus do Projeto NURC, com a fala culta de diferentes regiões brasileiras colhida em situações pré-estabelecidas; o *Corpus Histórico do Português Tycho Brahe (2008)*, composto de textos escritos por autores nascidos entre 1380 e 1845, desenvolvido junto à área de sintaxe gerativa diacrônica da UNICAMP; o Corpus do Português desenvolvido por pesquisadores da Brigham Young University e da Georgetown University (http://www.corpusdoportugues.org/x.asp), também com textos dos anos 1300s aos anos 1900s, contendo amostras em português do Brasil e de Portugal, orais, de ficção, de jornal e acadêmicos.

Embora já possamos contar com esses e outros corpora do português, ainda não dispomos de um corpus de dimensões abrangentes, que seja representativo e organizado de acordo com convenções aplicadas internacionalmente, como, por exemplo, o American National Corpus (2009) e o British National Corpus (2009). Quando tratamos da modalidade oral, a limitação de corpora em português fica ainda mais acentuada, e 'infelizmente, como se sabe, não há disponível, no Brasil, nenhum banco de dados representativo da língua falada contemporânea' (Neves, 1999, p. 14).

Para que um corpus seja realmente representativo, Sardinha (2004, p. 22-25) sugere que alguns aspectos devam ser considerados, tais como:

(a) A extensão do corpus: em geral, acredita-se que quanto maior um corpus, melhor ele será. Obviamente, um corpus grande contém mais amostras de usos linguísticos. No entanto, assim como uma pesquisa de opinião não considera uma população inteira, mas extratos dela, também os corpora representativos devem obedecer a padrões de extensão de acordo com a pesquisa a ser desenvolvida. Para Biber *et al.* (1998, p. 249), em estudos de frequência de traços linguísticos, por exemplo, 10 amostras de textos podem representar bem um gênero. Quanto ao tamanho das amostras em um corpus, Biber (1990 *in* Biber *et al.*, 1998, p. 249) indica que amostras de 1.000 palavras mostram resultados relativamente estáveis quanto a muitos traços gramaticais bastante usuais. Em outros casos, entretanto, quando o item gramatical é pouco usual, são necessárias amostras bem maiores em estudos quantitativos. Segundo os autores, para estudos lexicográficos, deve-se contar com corpora mais extensos, já que algumas

[2] Gêneros discursivos são vistos neste trabalho na perspectiva sistêmico-funcional, considerando-os como processos sociais, que se desenvolvem em fases e têm um objetivo definido (tradução do original: 'genres are staged, goal-oriented social processes'; Martin, 1997, p. 13). De acordo com esta definição os gêneros discursivos têm mais de um estágio em seu desenvolvimento e são formulados considerando os ouvintes ou leitores a quem se dirigem. Os gêneros representam o sistema de processos sociais através dos quais os indivíduos de uma dada cultura vivem a sua vida.

palavras ou colocações são pouco frequentes e somente um grande corpus viabilizará o seu estudo.

(b) O objetivo da compilação: corpora podem ser compilados com uma série de objetivos em vista. A utilização mais (re)conhecida é o apoio à lexicografia e à confecção de dicionários voltados para o uso da língua, como foi o caso do dicionário de inglês Collins-Cobuild, produzido a partir do corpus de Birmingham, atualmente denominado como o Bank of English (http://www.collins.co.uk/books.aspx?group=153). A elaboração de dicionários específicos (para estudantes, por exemplo) pode também dirigir a escolha dos textos a serem coletados. Para aplicações genéricas, em que um conteúdo variado é privilegiado, a compilação de textos jornalísticos pode ser suficiente, já que em jornais e revistas está representada uma ampla e variada coleção de textos de diferentes gêneros. Para a elaboração de glossários ou ferramentas terminológicas, é preciso restringir os textos a um domínio, bem como a um nível de profundidade (mais informativo versus mais técnico). Muitas vezes, faz-se necessário incluir tanto trechos de discurso oral quanto escrito. Outras vezes, para um levantamento de padrões gramaticais, apenas uma das modalidades é necessária ou suficiente. Além desses objetivos, há ainda outros que poderiam ser descritos, incluindo ainda questões relativas à pesquisa acadêmica ou a usos comerciais.

(c) A adequação aos interesses do pesquisador: quando um corpus for compilado com o objetivo de servir de base para uma pesquisa acadêmica, também deverão ser levados em conta os interesses do pesquisador. Pesquisas de cunho diacrônico, por exemplo, demandam a coleta de textos assemelhados de épocas diferentes (Biber e Finegan, 1989). Pesquisas sobre padrões recorrentes em discurso científico podem incluir textos de áreas de conhecimento diferentes, mas os textos devem pertencer a um mesmo gênero. Assim, em estudos de base comparativa, deve-se buscar, desde a compilação do corpus, o que Connor e Moreno (2005) chamam de *Tertium Comparationis*, ou seja, uma plataforma comum de comparação. E pesquisas de gêneros discursivos exigem uma gama o mais variada possível de trechos de textos de diferentes usos da língua, ainda que as amostras possam ser curtas em extensão. Logo, independente de como a coleta for feita, o corpus deve ser organizado de tal maneira que o pesquisador possa retirar dele aquilo que mais lhe interessa.

(d) A função representativa de todos os corpora: qualquer que seja a forma como foi compilado, um corpus pode ser considerado representativo em maior ou menor grau (Leech *in* Sardinha, 2004, p. 22). Assim, um corpus poderá conter apenas textos de um autor, escritos em uma determinada época de sua vida, mas será representativo do estilo daquele autor em um determinado período. Da mesma maneira, um corpus pode conter apenas textos de um único gênero discursivo ou de uma modalidade (oral ou escrita), podendo tornar-se representativo deste gênero ou desta modalidade.

(e) O corpus como amostragem de uma população de tamanho desconhecido: a maioria dos corpora disponíveis costumam ser compostos de textos de jornais e revistas. Essa característica se deve à maior facilidade de compilação desse tipo de material. No entanto, como não se tem uma medida da proporção de usos de textos e discursos numa comunidade falante e que faz uso da escrita, cada corpus passa a ter apenas uma pequena parte do total de amostras potenciais de língua.

(f) A linguagem como um sistema global e probabilístico: aliado ao aspecto anterior, é mister se pensar que todo corpus é um fragmento de língua, mas que, mesmo assim, representa o sistema global de uma língua (ou parte dele) e que, mesmo incompleto e fragmentado, pode refletir as possibilidades de ocorrência de usos linguísticos potenciais.

Além dos aspectos vistos acima, para criarmos um corpus representativo do português do Brasil, acreditamos que devemos considerar, principalmente, que os textos devem ser: autênticos, refletindo a real 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; e pertencentes a diferentes gêneros discursivos, visando representar a maior variedade possível de ações sociais.

O CORPOBRAS

Características gerais

O CORPOBRAS é um corpus representativo do português do Brasil, em fase de desenvolvimento, e que pretende fornecer dados e subsídios para uma análise multidimensional da variação entre gêneros discursivos[3]. Atualmente com aproximadamente 1.170.000 palavras, o corpus compreende 27 (vinte e sete) gêneros discursivos: artigo científico, carta ao editor, carta de reclamação, carta de recomendação, carta pessoal, carta profissional, carta profissional acadêmica, circular, conto, crônica, disser-

[3] O projeto CORPOBRAS contou com auxílio do CNPq (2004 a 2007) através de Edital Universal (processo 480143/2004-8). Atualmente está vinculado ao projeto 'Escrita e inclusão social: análise de corpus e a metáfora gramatical no Ensino Médio', que tem apoio da FAPERJ, através do Edital de Humanidades (processo E-26/112.269/2008).

tação e tese (introduções e conclusões), editorial, e-mail acadêmico, e-mail pessoal, notícia de jornal, redação de alunos de ensino médio, redação de alunos universitários, redação de vestibular, romance, conversa carioca, conversa de crianças, entrevista acadêmica, grupo de enfoque, atendimento ao cliente, discurso político e roteiro cinematográfico (ver Tabela, Anexo I).

Como uma das principais metas do CORPOBRAS é manter um nível significativo de representatividade, as suas características sempre se adaptam a essa meta e podem ser resumidas de acordo com os parâmetros de: *modo ou modalidade; tempo; finalidade; autoria; seleção; e conteúdo.*

Em termos de *modo*, o CORPOBRAS não só contempla as modalidades oral e escrita , como a modalidade escrita para ser falada.

Quanto ao parâmetro *tempo*, o CORPOBRAS se debruça apenas no tempo contemporâneo, considerando textos de domínio acadêmico, comercial e jornalístico (artigos científicos, circulares, notícias, editoriais, etc) da última década do século passado e os primeiros anos deste século (1990-2009). Já no caso do domínio literário e pessoal, ou seja, romances, contos, crônicas, cartas pessoais, o corpus considera um escopo maior, mas ainda dentro da contemporaneidade: de 1901 a 2001.

A *finalidade* do CORPOBRAS, como mencionado anteriormente, é fornecer subsídios para o estudo de diversos gêneros do discurso oral e escrito, com o auxílio de ferramentas computacionais.

A *autoria* dos textos que compõem o CORPOBRAS está circunscrita a falantes nativos do português, de modo a manter a autenticidade dos usos de língua. No entanto, não há limitações quanto ao status dos escritores, regiões geográficas ou áreas de conhecimento. Assim, temos, por exemplo, cartas redigidas por escritores profissionais ou usuários não especialistas da língua; textos de diferentes regiões de país, como editoriais e notícias de jornais do Distrito Federal e dos estados do Rio de Janeiro, São Paulo, Espírito Santo, Alagoas e Rio Grande do Norte; e textos de diferentes áreas de conhecimento, como artigos científicos de linguística, nutrição, etc.

A *seleção* dos textos é realizada visando-se manter uma amostragem equilibrada dos textos de diferentes gêneros que compõem o corpus. Quanto ao conteúdo dos textos, visa-se a variedade de temas e aproximação com a diversidade discursiva.

Por seguir todos os parâmetros mencionados acima, o CORPOBRAS apresenta, como uma de suas características mais marcantes, uma ampla variedade de modalidades, de gêneros discursivos e de regiões, assuntos e autores. Em sua atual configuração, o corpus contém 347.769 palavras em gêneros do discurso oral, 783.204 em gêneros do discurso escrito e 39.931 em gêneros do discurso escrito para ser falado. É importante mencionar aqui que estamos caracterizando gênero não só em termos de forma e conteúdo, mas como um processo social com funcionalidade própria (Martin, 1997)

Cabe, entretanto, mencionar que a classificação dos gêneros em um corpus tem-se mostrado como tarefa difícil, já que, até o momento não há um consenso sobre o conceito de gênero na área de estudos linguísticos (Connor, 1996; Johns, 2002; Marcuschi, 2002). Em 1964, quando acabou de ser compilado, o Brown Corpus, apresentava 15 gêneros, sendo que, dentre os grupos de textos que incluía, estavam 'cultura popular', 'humor', 'religião', etc. Atualmente, estas categorias não seriam incluídas como gêneros discursivos ou textuais (embora permaneçam assim indicadas no Brown Corpus e no LOB Corpus), já que tem havido um refinamento maior nas classificações de gêneros discursivos. Entretanto, muitas dúvidas ainda persistem e, em alguns casos, para solucionar certas situações que parecem híbridas, como, por exemplo, no caso dos discursos políticos, peças teatrais ou roteiros, alguns pesquisadores têm criado categorias específicas em seus corpora, como por exemplo 'textos escritos para serem falados'. Outros pesquisadores têm excluído estes textos de seus corpora por considerarem-nos de difícil classificação (modalidade oral ou escrita?) ou por preferirem ater-se, em corpora de discurso oral, como o CANCODE (Cambridge and Nottingham Corpus of Discourse in English), a textos orais não ensaiados que reproduzem a fala não formal (McCarthy, 1998, p. 9). No CORPOBRAS, decidimos incluir 'textos escritos para serem falados' e classificá-los como um grupo separado, devido às suas características específicas que os diferenciam dos demais gêneros da modalidade oral e escrita, classificados de acordo com os canais de produção dos textos (Halliday e Hasan, 1989).

A representatividade no CORPOBRAS

Quando falamos de um corpus representativo, temos de considerar três questões (Sardinha, 2004): *Do quê? Para quê? Para quem?*

Primeiramente, *o que* está sendo representado, ou seja, de que representatividade estamos falando? No caso do CORPOBRAS, trata-se da representação de amostras do português do Brasil, com o maior número possível de gêneros discursivos. A meta é aumentar o número de textos em alguns gêneros já compilados e incluir ainda gêneros que não foram contemplados, tais como conversas em fóruns de discussão, perfis em plataformas de cursos a distância, narrativas de histórias pessoais, narrativas em sala de aula, etc.

A segunda questão se refere à finalidade: representatividade *para quê*? Como já mencionamos acima, a finalidade inicial do CORPOBRAS é o estudo da variação em gêneros do discurso oral e escrito, assim como o estudo com abordagem estatística para verificar a co-ocorrência de traços linguísticos em gêneros discursivos. Entretanto,

o corpus poderá ser usado também em estudos linguísticos com objetivos diversos, como já tem ocorrido com subcorpora, extraídos do CORPOBRAS (Turunen, 2009).

Finalmente, devemos considerar para quem o corpus é representativo, ou seja, que pesquisadores ou especialistas reconhecem-no como representativo. No caso do CORPOBRAS, este segue os padrões indicados por Biber (Biber *et al.*, 1998, p. 249), que, após testes estatísticos, comprovou que, no LOB corpus (Lancaster-Oslo/Bergen Corpus), que foi utilizado na descrição da variação entre gêneros do inglês (Biber, 1988), 10 textos representam a variedade de falantes e escritores e as categorias do corpus para a variação de muitos traços gramaticais. Seguindo os mesmos padrões, acreditamos que os gêneros discursivos, no CORPOBRAS, para serem representativos em relação a estudos da frequência de traços linguísticos, devem conter grupos de textos que incluem 10 ou mais amostras.

Estudos baseados em corpora: Análises aplicadas a partir do CORPOBRAS

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 a ajudar a melhor acessar, analisar e contrastar corpora linguísticos, havendo uma ampla gama de ferramentas tecnológicas como suporte para essa tarefa. Dentre as metodologias existentes e produtivas, aparece a Análise Multidimensional (Biber 1988; Conrad e Biber 2001), de base funcional e estatística, capaz de caracterizar a variação linguística em grandes corpora de dados, com o auxílio de testes estatísticos, tais como a Análise Fatorial. Essa metodologia foi utilizada por Biber (1988) para descrever a variação de 23 gêneros do discurso oral e escrito, em inglês, ao longo de dimensões textuais identificadas a partir de um corpus de aproximadamente 900.000 palavras. Futuramente, a Análise Multidimensional será também aplicada a todo o CORPOBRAS, visando-se a identificação de contínuos de variação e a descrição dos gêneros do português do Brasil ao longo destes parâmetros de variação.

Até o momento, alguns estudos multidimensionais já foram desenvolvidos utilizando apenas parcialmente os dados do CORPOBRAS. Alguns destes trabalhos contribuíram para a expansão do corpus, coletando dados que foram posteriormente acrescentados a ele (Oliveira, 1997; Lanziotti, 2002); outros trabalhos contribuíram para o estudo da variação entre alguns gêneros do português ou para o estudo da variação entre textos do CORPOBRAS e textos em inglês (Moraes, 2005; Oliveira, 2001, 2006). Além destes, outros estudos baseados nos dados do CORPOBRAS têm sido desenvolvidos, tomando como base teórico-metodológica a Linguística Sistêmico-Funcional (Caldeira, 2006; Marques, 2006; Nóbrega, 2009).

Considerações finais

Como mostramos, a compilação de um corpus representativo do português do Brasil com gêneros do discurso oral e escrito poderá fornecer material para o estudo de gêneros do discurso pedagógico, profissional e espontâneo, adotando-se diferentes metodologias para estudos baseados em corpora de língua em uso, e para o estudo de itens lexicais inseridos em variados gêneros. Além disso, as pesquisas baseadas em corpus mostram que esse tipo de estudo é promissor e tem aplicações práticas em diversas áreas, como os estudos da variação entre gêneros e a lexicografia.

Em relação à compilação de textos para o CORPOBRAS, pretendemos expandir as amostras de discurso oral, discurso escrito, e discurso escrito para ser falado, bem como ampliar a diversificação de gêneros, mantendo sempre como parâmetro norteador a sua representatividade. A anotação do corpus está planejada para a próxima etapa do projeto, mas dependerá, para sua execução, de financiamento específico, visto que tal empreitada requer apoio robusto, para cobrir despesas com tecnologia e recursos humanos especializados. Por outro lado, em termos de análise, pretendemos intensificá-las, buscando não só itens lexicais, mas igualmente padrões lexicais e variação de traços linguísticos em relação aos gêneros.

Referências

AMERICAN NATIONAL CORPUS. 2009. Disponível em http://americannationalcorpus.org/about.html. Acesso em: 19/11/2009.

BIBER, D. 1988. *Variation across speech and writing*. Cambridge, Cambridge University Press, 299 p.

BIBER, D. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistics Computing*, **5**:257-269.

BIBER, D. 1995. *Dimensions of Register Variation: A cross-linguistic comparison.* Cambridge, Cambridge University Press, 428 p.

BIBER, D.; CONRAD, S.; REPPEN, R. 1998. *Corpus linguistics: Investigating language structure and use.* Cambridge, Cambridge University Press, 300 p.

BIBER, D.; FINEGAN, E. 1989. Drift and the evolution of English style: a history of three genres. *Language,* **65**(3):487-517.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. 1999. *Longman Grammar of Spoken and Written English*. Essex, Pearson Education Limited, 1204 p.

BRITISH NATIONAL CORPUS. 2009. Disponível em http://www.natcorp.ox.ac.uk. Acesso em: 19/11/2009.

CALDEIRA, J.R. 2006. *A redação de vestibular como gênero: configuração e processo social.* Rio de Janeiro, RJ. Dissertação de Mestrado. Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio, 150 p.

CARTER, R.; MCCARTHY, M. 2006. *Cambridge grammar of English: A comprehensive guide – Spoken and written English grammar and usage*. Cambridge, Cambridge University Press, 973 p.

CONNOR, U. 1996. Genre specific studies in contrastive rhetoric. *In*: U. CONNOR, *Contrastive Rhetoric*: *Cross-cultural aspects of second language writing.* Cambridge, Cambridge University Press, p. 126-149.

CONNOR, U.; MORENO, A. 2005. Tertium Comparationis: A Vital Component in Contrastive Rhetoric Research. *In:* P. BRUTHIAUX;

D. ATIKINSON; W.G. EGGINGTON; W. GRABE; V. RAMANATHAN (eds.), *Directions in Applied Linguistics: Essays in honnor of Robert Kaplan*. Clevedon, Multilingual Matters, p. 153-164.

CONRAD, S.; BIBER, D. 2001. Variation in English: Multi-dimensional studies. New York, Longman, 255 p.

DIRECT. 2005. Disponível em http://www2.lael.pucsp.br/direct/. Acesso em 19/11/2009.

GRABE, W. 2004. Perspectives in applied linguistics: A North American view. *AILA Review,* **17**:105-132.

HALLIDAY, M.A.K.; HASAN, R. 1989. *Language, context, and text: Aspects of language in a social-semiotic perspective.* Oxford, Oxford University Press, 126 p.

HALLIDAY, M.A.K.; MATTHIESSEN, C.M.I.M. 2004. *An introduction to functional grammar.* 3ª ed., London, Hodder Arnold, 689 p.

HUNSTON, S. 2002. *Corpora in applied linguistics*. Cambridge, Cambridge University Press, 241 p.

JOHNS, A. (ed.) 2002. *Genre in the Classroom: Multiple perspectives.* Mahwah, Lawrence Erlbaum Associates, Publishers, 350 p.

KAPLAN, R. (ed.) 2002. *The Oxford handbook of applied linguistics*. Oxford, Oxford University Press, 641 p.

LANZIOTTI, M.G.P. 2002. *Variação de gêneros discursivos: a explicitação do contexto em um corpus do português escrito*. Rio de Janeiro, RJ. Dissertação de Mestrado. Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio, 140 p.

LINGUATECA. 2009. Disponível em http://www.linguateca.pt/. Acesso em: 19/11/2009.

MARCUSCHI, L.A. 2002. Gêneros textuais: definição e funcionalidade. *In:* A.P. DIONÍSIO; A.R. MACHADO; M.A. BEZERRA (orgs.), *Gêneros Textuais e Ensino*. Rio de Janeiro, Editora Lucerna, p. 20-35.

MARQUES, G.O. 2006. *Tecnologia e internet no ensino de língua estrangeira: Avaliação discursiva de professores e alunos*. Rio de Janeiro, RJ. Dissertação de Mestrado, Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio, 162 p.

MARTIN, J.R. 1997. Analysing genre: Functional parameters. *In:* F. CHRISTIE; J.M. MARTIN (eds.), *Genre and Institutions: Social Processes in the Workplace and School*. London, Continuum, p. 3-39.

MCCARTHY, M. 1998. *Spoken language and applied linguistics*. Cambridge, Cambridge University Press, 206 p.

MORAES, L.S.B. 2005. *O metadiscurso em artigos acadêmicos: variação intercultural, interdisciplinar e retórica*. Rio de Janeiro, RJ. Tese de Doutorado. Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio, 194 p.

NEVES, M.H. de. M. 1999. *Gramática de Usos do Português*. São Paulo, Editora UNESP, 1037 p.

NÓBREGA, A.N.A. 2009. *Narrativas e avaliação no processo de construção do conhecimento pedagógico: abordagem sociocultural e sociossemiótica*. Rio de Janeiro, RJ. Tese de Doutorado. Pontifícia Universidade Católica do Rio de Janeiro-PUC-Rio, 244 p.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL. 2006. Disponível em: http://www.nilc.icmc.usp.br/nilc/index.html. Acesso em: 19/11/2009.

OLIVEIRA, L.P. 1997. *Variação intercultural na escrita: contrastes multidimensionais em Inglês e Português*. São Paulo, SP. Tese de Doutorado, LAEL/PUC-SP, 358 p.

OLIVEIRA, L.P. 2001. Cross-linguistic and cross-genre involvement variation in the writing of academics. *In:* CONFERÊNCIA ANUAL DA AMERICAN ASSOCIATION FOR APPLIED LINGUISTICS, Saint Louis, 2001. [Trabalho apresentado].

OLIVEIRA, L.P. 2006. Influências culturais e contrastes em gêneros do discurso escrito. *In:* J.C.V. DINIZ (ed.), *Diálogos Ibero-Americanos II*. Rio de Janeiro, Editora Galo Branco, p. 72-93.

OLIVEIRA, L.P. 2009. Linguística de corpus: teoria, interfaces e aplicações. *Matraga,* **16**(24):48-76.

SARDINHA, T.B. 2004. *Linguística de Corpus*. São Paulo, Manole, 410 p.

TEUBERT, W. 1996. Editorial. *International Journal of Corpus Linguistics,* **1**(1):iii-x.

THOMPSON, G.; HUNSTON, S. (eds.). 2006. *System and corpus: Exploring connections*. London, Equinox, 326 p.

TURUNEN, V.J. 2009. *A reversão da relevância: aspectos semânticos e pragmáticos de formações diminutivas em português do Brasil*. Rio de Janeiro, RJ. Tese de Doutorado. Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio, 194 p.

TYCHO BRAHE. 2008. Disponível em http://www.tycho.iel.unicamp.br/~tycho/pesquisa/. Acesso em: 19/11/2009.

Submissão: 31/08/2009
Aceite: 06/11/2009

Lúcia Pacheco de Oliveira
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente, 225
22453-900, Rio de Janeiro, RJ, Brasil

Maria Carmelita Padua Dias
Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente, 225
22453-900, Rio de Janeiro, RJ, Brasil

ANEXO 1

CORPOBRAS (2009)		
Discurso escrito		
Gêneros	**Número de textos**	**Número de palavras**
Artigo científico	12	63.818
Carta ao editor	18	1.054
Carta de reclamação	136	21.417
Carta de recomendação	31	6.012
Carta pessoal	16	7.829
Carta profissional	16	3.166
Carta profissional acadêmica	15	3.529
Circular	16	2.608
Conto	14	15.253
Crônica	26	17.434
Dissertação e Tese (Introduções e Conclusões)	32	69.447
Editorial	16	7.931
E-mail acadêmico	15	1.816
E-mail pessoal	16	1.858
Notícia de jornal	99	40.409
Redação de aluno	16	3.416
Redação de aluno universitário	91	25.065
Redação de ensino médio	24	6.238
Redação de vestibular	139	28.646
Romance	28	27.061
	Total de palavras:	**347.769**
Discurso oral		
Conversa carioca	53	353.678
Conversa de criança	94	84.573
Entrevista (acadêmica)	17	88.769
Grupo de enfoque	7	40.513
Atendimento ao cliente	393	215.671
	Total de palavras:	**783.204**
Discurso escrito para ser falado		
Discurso político	27	22.751
Roteiro cinematográfico	18	17.180
	Total de palavras:	**39.931**

Total de Palavras no Corpus em 2009**: 1.170.904**