Lúcia Pacheco de Oliveira, PUC-Rio
Maria Carmelita Padua Dias, PUC-Rio

1. 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 Lingüística de Corpus, que pode ser vista tanto como uma abordagem que propõe uma nova maneira de olhar a linguagem, quanto como a face moderna da lingüística empírica (Teubert, 1996).

Cada vez mais os estudos lingüísticos teóricos e aplicados vêm se beneficiando do uso de corpora para a descrição de fenômenos lingüísticos ou para a verificação de hipóteses acerca dos mesmos (Grabe, 2004, Kaplan, 2002, Biber, Conrad & Reppen, 1998).

Pesquisas baseadas em corpus são predominantemente aplicadas e ligadas à análise do discurso, incluindo o estudo de gêneros e a variação lingüística diacrônica e sincrônica; à lexicografia, incluindo a elaboração de dicionários, como o COBUILD; aos estudos léxico-gramaticais para a produção de material de ensino, tais como gramáticas baseadas em corpora (Biber, Johansson, Leech, Conrad & Finegan 1999). Para que esses estudos sejam conduzidos, é preciso que grandes quantidades de dados lingüísticos sejam compilados e sistematicamente organizados, de modo a serem posteriormente analisados com o auxílio de ferramentas computadorizadas (Biber, Conrad & Reppen, 1998; Sardinha, 2000).

Por suas características ligadas à tecnologia e, ao mesmo tempo, a estudos lingüísticos focalizando usos da linguagem, a Lingüística de Corpus tem uma interface com a Lingüística Computacional e com a Lingüística Aplicada, beneficiando-se dos métodos e resultados dessas duas áreas. No entanto, apesar das facilidades trazidas pela tecnologia, 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 lingüístico 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 raras e de difícil coleta, é indispensável para caracterizar mudanças e usos lingüísticos.

2. 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 lingüísticos ou características de variedades lingüísticas 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. Apesar de terem sido tomadas algumas iniciativas, algumas extremamente bem-sucedidas, como o corpus do NILC (http://www.nilc.icmc.usp.br), para a compilação de corpora em português, ainda não contamos com um corpus de dimensões abrangentes, que seja representativo e organizado de acordo com convenções aplicadas internacionalmente, como no American National Corpus (ANC: http://americannationalcorpus.org/about.html) e o British National Corpus (http://www.natcorp.ox.ac.uk).

Para que um corpus seja realmente representativo, alguns aspectos devem ser considerados:

a) A extensão do corpus - Em geral, acredita-se que quanto maior um corpus, melhor ele será. Obviamente, um corpus extenso contém mais amostras de usos lingüísticos. 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, Conrad & Reppen (1998:249), em estudos de freqüência de traços

lingüísticos, por exemplo, 10 amostras de textos de um gênero podem representar uma categoria lexical ou sintática e textos com 1.000 palavras garantem resultados relativamente estáveis quanto ao uso da maioria dos traços lingüísticos. Segundo os autores, entretanto, para estudos lexicográficos, deve-se contar com corpora mais extensos, já que algumas palavras ou colocações são pouco freqüentes 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.titania.bham.ac.uk/docs/about.htm). 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, aquilo que Connor (2005) chama de Tertium Comparationes, 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. 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 lingüísticos 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, principalmente, pertencentes a diferentes gêneros discursivos, para representar a maior variedade possível de ações sociais.

3. O CORPOBRAS PUC-Rio

3.1. Características gerais

O CORPOBRAS PUC-Rio é um projeto de compilação de 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 (CNPq, processo 480143/2004-8; http://www.let.puc-rio.br/corpobras.htm).
Atualmente com aproximadamente 550.000 palavras, o CORPOBRAS almeja atingir 1 milhão de palavras em 2007. 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 seguintes parâmetros: modo ou modalidade; tempo; finalidade; autoria; seleção; conteúdo;
Em termos de modo, o CORPOBRAS não só contempla as modalidades oral e escrita, mas também procura equilibrar o número de amostras de cada uma delas. Atualmente, o corpus já conta com 21 gêneros discursivos: 17 gêneros do discurso escrito e 4 do discurso oral. É importante mencionar aqui que estamos caracterizando gênero não só em termos de forma e conteúdo, mas como um processo social com funcionalidade própria (Dias e Quental, 2005; Martin, 1997)
Como o CORPOBRAS pretende fornecer resultados de estudos sincrônicos da língua portuguesa do Brasil, o corpus 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 (1990-2006). 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 é fornecer subsídios para o estudo de diversos gêneros do discurso oral e escrito, com o auxílio de ferramentas computacionais. De modo a manter a autenticidade dos usos de língua, a autoria dos textos que compõem o CORPOBRAS está circunscrita a falantes nativos do português. 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 lingüística, 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.

3.2. A representatividade no CORPOBRAS PUC-Rio

Quando falamos de um corpus representativo, temos de considerar três questões (Sardinha, 2005):

1.Do que ?

2.Para que?

3.Para quem?

Primeiramente, de que representatividade estamos falando, ou seja, o que está sendo representado? 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 incluir ainda gêneros que não foram contemplados, mantendo-se um equilíbrio semelhante ao do British National Corpus (10% discurso oral e 90% discurso escrito).

A segunda questão se refere à finalidade: representatividade para quê? Como já mencionamos acima, a finalidade 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 lingüísticos em gêneros discursivos.

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: 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, podem ser representados por grupos de textos que incluem 10 ou mais amostras. Até o presente momento, os gêneros discursivos que estão representados no CORPOBRAS são os seguintes:

- Artigos científicos
- Atendimentos de serviço
- Cartas de recomendação
- Cartas pessoais
- Cartas profissionais
- Cartas profissionais acadêmicas
- Circulares
- Contos
- Conversas
- Crônicas
- Discursos políticos
- Editoriais
- E-mails acadêmicos
- E-mails pessoais

- Entrevistas acadêmicas
- Notícias de jornal
- Redações de alunos (ensino médio)
- Redações de alunos (universitários)
- Reuniões de negócios
- Romances
- Roteiros cinematográficos

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 lingüísticos (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, tais 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. Entretanto, muitas dúvidas ainda persistem e, em alguns casos, para solucionar certas situações que parecem híbridas, como por exemplo no caso do 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 esses textos de seus corpora por considerarem-nos de difícil classificação (modalidade oral ou escrita?) ou por preferirem ater-se, no caso de 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: 9).

No CORPOBRAS, decidimos incluir “textos escritos para serem falados”, mas até o momento ainda estamos trabalhando com a divisão dos gêneros em modalidade oral e escrita, de acordo com os canais de produção dos textos (Halliday & Hasan, 1989). A seguir, nas Figuras 2 e 3, apresentamos a distribuição das amostras do corpus de acordo com as duas modalidades discursivas.

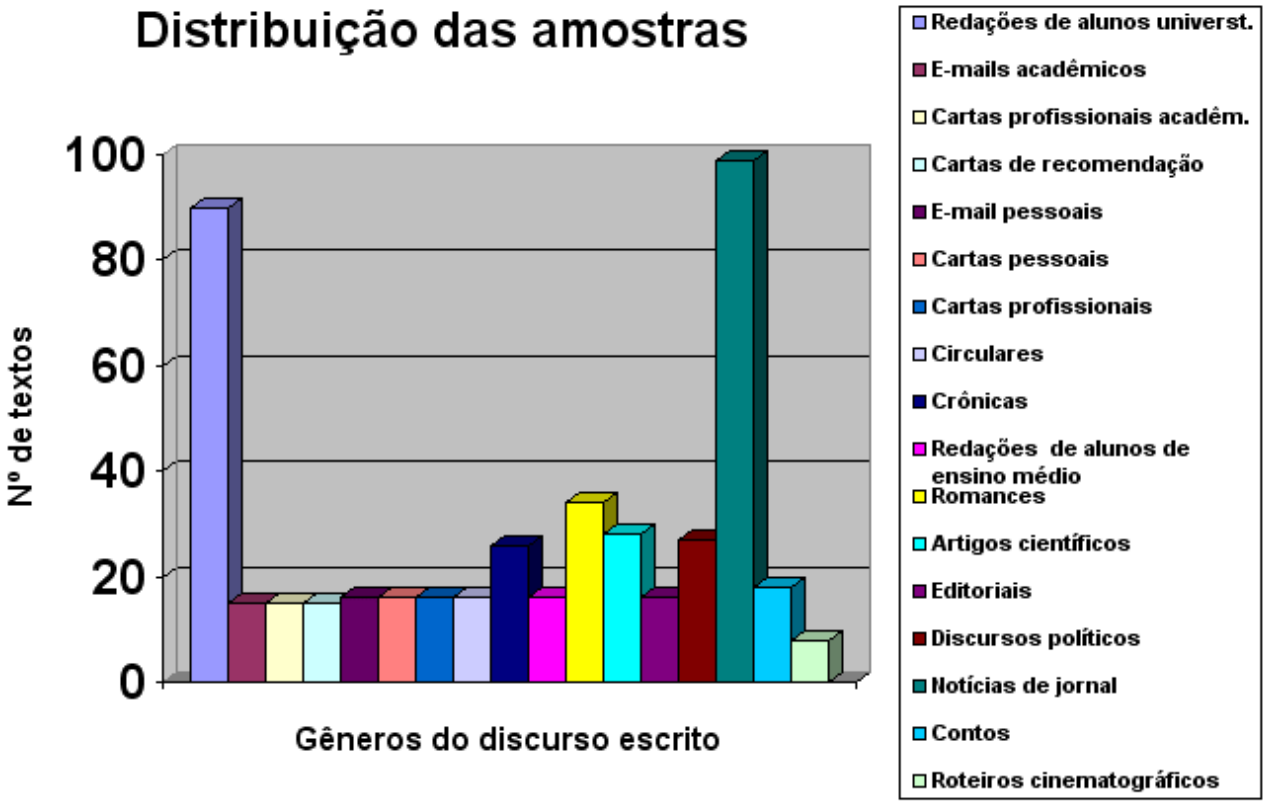

Figura 2: Distribuição dos textos no CORPOBRAS: Modalidade escrita

Figura 2: Distribuição dos textos no CORPOBRAS: Modalidade escrita

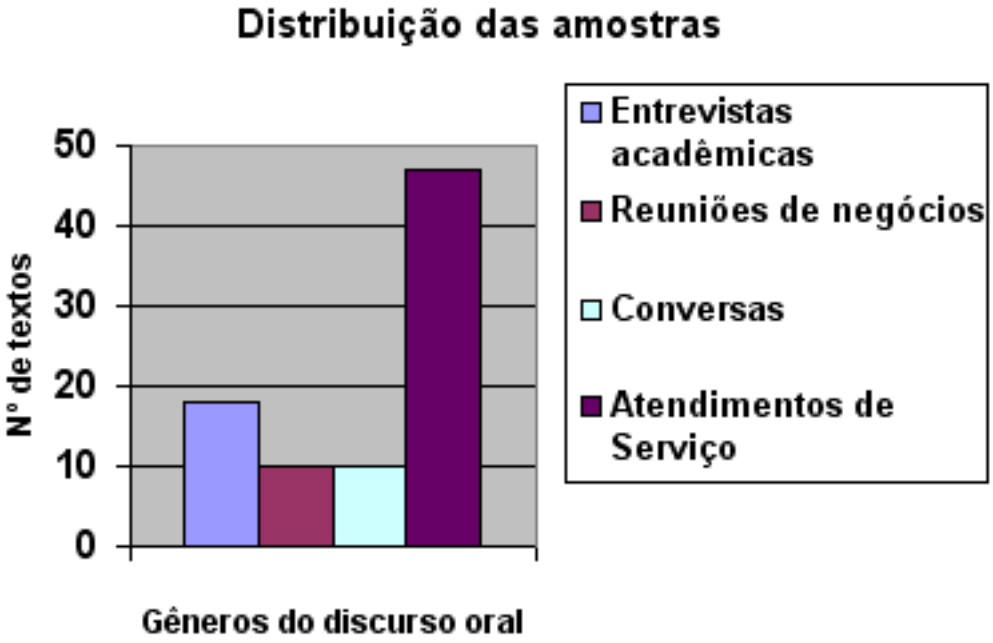

Figura 3: Distribuição dos textos no CORPOBRAS: Modalidade oral

4. 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 e melhorar o acesso, a análise e o contraste entre corpora lingüísticos, 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), capaz de caracterizar a variação lingüística em grandes corpus de dados, com o auxílio de medidas estatísticas. Esse tipo de metodologia foi também utilizado para a análise de textos em português (Oliveira, 1997, 2001; Lanziotti 2002).

Assim, visando a um estudo da variação lingüística na língua oral e escrita, Biber (1988) propôs uma metodologia 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”.

O principal objetivo desse método é fornecer descrições abrangentes de padrões de variação entre gêneros discursivos, levando em consideração dois componentes básicos:

- a identificação de parâmetros lingüísticos subjacentes, ou dimensões de variação; e
- a especificação das semelhanças e diferenças entre gêneros e textos em relação a estas dimensões.

A Análise Multidimensional desenvolvida neste tipo de metodologia de corpus compreende as seguintes etapas:

1) Análises preliminares:

- Seleção de traços lingüísticos (itens lexicais e gramaticais);
- Identificação dos traços lingüísticos nos textos;
- Cálculo da freqüência dos traços lingüísticos em cada variável.

2) Análises estatísticas:

- Normatização das freqüências;

- Análise Fatorial para identificação da co-ocorrência de traços lingüísticos no corpus.
- Identificação dos Fatores;
- Interpretação das co-ocorrências das variáveis nos Fatores como Dimensões Textuais, de acordo com sua função discursiva.

3) Análise comparativa da variação:

- Estandardização das freqüências normatizadas, considerando-se a média e o desvio padrão;
- Cálculo de escores correspondentes a cada Fator/ Dimensão;
- Comparação das dimensões textuais.

Vários estudos multidimensionais já foram desenvolvidos utilizando 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 gêneros do português ou para o estudo da variação entre textos do CORPOBRAS e textos em inglês (Moraes, 2005, Oliveira, 2001, 2006), que formam um corpus paralelo para o estudo de influências culturais na escrita.

Dentre estes trabalhos multidimensionais, vamos enfocar dois: o estudo de influências culturais em redações de alunos universitários em português e inglês (Oliveira, 1997) e o estudo do envolvimento – informação, em cartas profissionais, cartas de recomendação e e-mails produzidos por acadêmicos brasileiros e americanos (Oliveira, 2001).

Estudo 1: Redações de alunos universitários em português e inglês

Corpus:

270 redações

- 90 textos de alunos brasileiros em português;
- 90 textos de alunos americanos em inglês
- 90 textos de alunos brasileiros em inglês.

Metodologia:

Análise Multidimensional - identificação de dimensões textuais ou parâmetros de

variação para gêneros discursivos através da co-ocorrência de itens léxico-gramaticais

Dimensões identificadas:

- Dimensão 1: Estilo elaborado vs. estilo reduzido
- Dimensão 2: Orientação para o envolvimento vs. informação
- Dimensão 3: Explicitação do contexto vs. não-explicitação do contexto

Resultados:

A variação dos três grupos de textos nestas dimensões mostrou que, em português, as redações de alunos universitários mostraram um estilo mais elaborado, caracterizado por sentenças mais longas e sintaticamente mais complexas. Já os textos em inglês caracterizaram-se por sentenças mais curtas e maior número de orações simples (cf. Oliveira, 1999). Quanto à segunda Dimensão, as redações em inglês como língua estrangeira foram as que mostraram maior envolvimento e as redações em português mostraram-se mais informacionais. Quanto à explicitação do contexto (cf. Oliveira, 2002) as redações em português mostraram que os alunos universitários brasileiros, em geral, contextualizam o tópico antes de desenvolvê-lo. Por outro lado, os alunos americanos usam uma organização dedutiva, enfocando o tópico logo no início do texto, para desenvolvê-lo logo a seguir. Os três gráficos abaixo ilustram estes resultados.

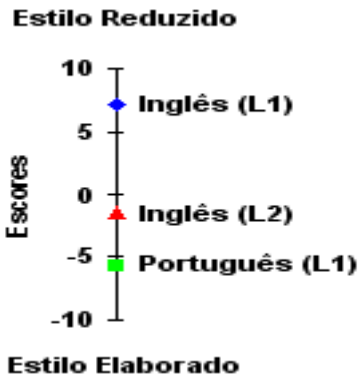

Gráfico 1: Dimensão 1 (“Estilo elaborado vs. estilo reduzido”) em 90 redações de alunos universitários

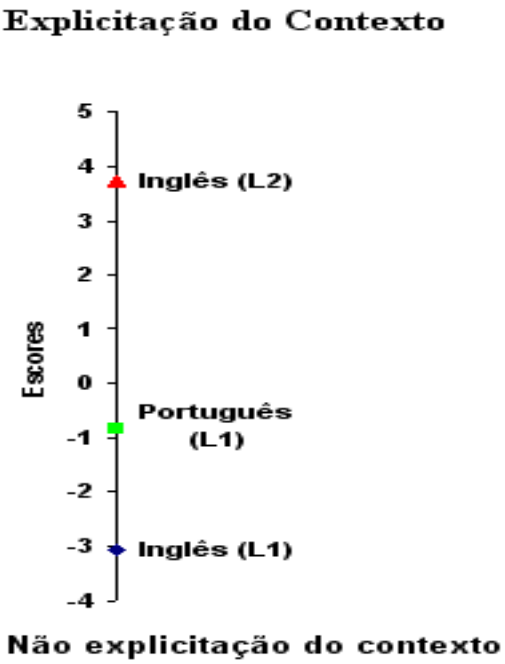

Gráfico 2: Dimensão 2 (“Orientação para o envolvimento vs. informação”) em 90 redações de alunos universitários

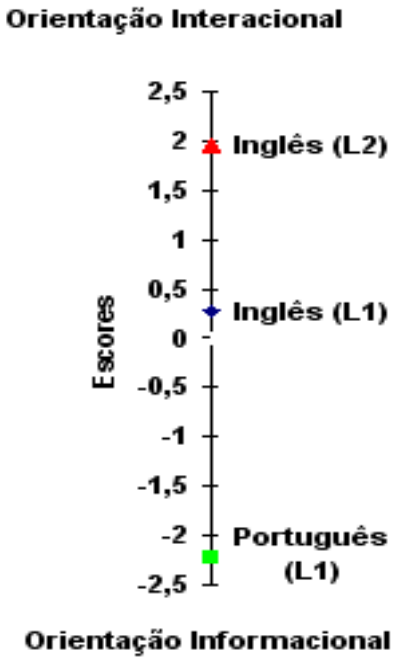

Gráfico 3: Dimensão 3 (“Explicitação do contexto vs. não-explicitação do contexto”) em 90 redações de alunos universitários

Estudo 2: Envolvimento em gêneros produzidos por acadêmicos

Corpus:

- 30 e-mails (15 em português;15 em inglês)
- 30 cartas profissionais (15 em português;15 em inglês)
- 30 cartas de recomendação (15 em português;15 em inglês)

Metodologia:

Análise Multidimensional - investigar como os gêneros variam na dimensão "Orientação para o envolvimento vs. informação", identificada por Biber em várias línguas (1988, 1995) e Oliveira (1997, 2001) em português.

Resultados:

A variação quanto ao envolvimento dos três gêneros estudados nas duas línguas mostrou que as cartas de recomendação e cartas profissionais mostram mais envolvimento em inglês do que em português; e-mails mostram mais envolvimento em português do que em inglês. Em geral, os contrastes entre as línguas indicam que os textos de acadêmicos americanos tendem a mostrar mais envolvimento do que os brasileiros. O gráfico 4, abaixo, ilustra esses resultados.

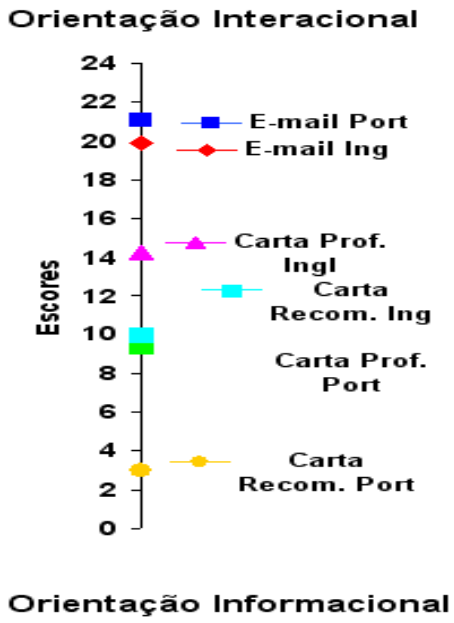

Gráfico 4 – Dimensão "Orientação para o envolvimento vs. informação" em cartas e e-mails produzidos por acadêmicos

5. Considerações finais

Como mostramos, a compilação de um corpus representativo do português do Brasil com gêneros do discurso oral e escrito poderá fornecer material tanto para o estudo de gêneros do discurso pedagógico, profissional e espontâneo, adotando-se diferentes metodologias para estudos baseados em corpora de língua em uso, quanto para o estudo de itens lexicais inseridos em variados gêneros.

As análises exemplificadas aqui mostram que esse tipo de estudo é promissor e tem aplicações práticas em diversas áreas, como os estudos da variação entre gêneros, a lexicografia, o ensino de línguas, a tradução e a lingüística computacional.

Para o futuro, em termos da coleta de dados, pretendemos expandir as amostras de discurso oral bem como ampliar a diversificação de gêneros. Por outro lado, em termos de análises, pretendemos realizar uma avaliação dos gêneros incluídos, levando em conta a modalidade. Buscamos também analisar não só itens lexicais, mas igualmente padrões lexicais em relação aos gêneros.

Referências:

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (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. & 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, England: Pearson Education Limited.

Connor, U. (1996). *Genre specific studies in contrastive rhetoric. In Contrastive Rhetoric: Cross-cultural aspects of second language writing.* Cambridge: Cambridge University Press.

Connor, U. & Moreno, A. (2005). *Tertium Comparationes: A Vital Component in Contrastive Rhetoric Research.* In Bruthiaux et al (Eds).
Directions in Applied Linguistics: Essays in Honnor of Robert Kaplan, pp. 153-164. Clevedon: Multilingual Mattres.
Conrad, S. & Biber, D. (2001). *Variation in English: Multi-Dimensional Studies*. New York: Longman.
Dias e Quental, (2005). *Novas tecnologias, velhos paradoxos: a internet em/como sala de aula.* Calidoscópio Vol. 3, No. 1, janeiro/abril 2005, 31-39
Grabe, W. (2004). *Perspectives in applied linguistics: A North American view*. AILA Review, 17, p. 105-132.
Halliday, M. A.K. e Hasan, R. (1989). *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective*. Oxford: Oxford University Press.
Johns, A. (Ed.) (2002). *Genre in the Classroom: Multiple perspectives.* Mahwah, New Jersey: Lawrence Erlbaum Associates, PublishersKaplan, R. (Ed.) (2002). The Oxford handbook of applied linguistics. Oxford: Oxford University 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.
Marcuschi, L. A. (2002). *Gêneros textuais: definição e funcionalidade.* In A. P. Dionísio, A. R. Machado, & M. A . Bezerra, M. A. (Orgs.). Gêneros Textuais e Ensino, pp. 20-35. Rio de Janeiro: Editora Lucerna.
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, pp. 3-39. London: Continuum.
McCarthy, M. (1998). *Spoken language and applied linguistics.* Cambridge: Cambridge University Press.
Moraes, L. S. B. (2005). *O metadiscurso em artigos acadêmicos: Variação intercultural, interdisciplinar e retórica.* Tese de Doutorado, Departamento de Letras, Rio de Janeiro, PUC-Rio.
Oliveira, L. P. (1997). *Variação Intercultural na Escrita: Contrastes*

Multidimensionais em Inglês e Português. Tese de Doutorado, LAEL/PUC-SP, São Paulo.

Oliveira, L. 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, L. P. (2001). *Cross-linguistic and cross-genre involvement variation in the writing of academics.* Trabalho apresentado na American Association for Applied Linguistics Annual Conference (AAAL), Saint Louis, Missouri, EUA.

Oliveira, L. P. (2002). *Explicitação do contexto em textos de alunos brasileiros e americanos.* Palavra, 8,102-116.

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

Sardinha, A.B. (2000). *Lingüística de Corpus: Histórico e Problemática.* D.E.L.T.A., Vol. 16, Nº 2, 323-367.

Sardinha, T. B. (2005). *Lingüística de Corpus.* Manole: São Paulo.Teubert, W. (1996). Editorial. International Journal of Corpus Linguistics, Vol.1, No. 1. iii-x.