LINGUÍSTICA DE CORPUS: TEORIA, INTERFACES E APLICAÇÕES

Lúcia Pacheco de Oliveira
(PUC-Rio/FAPERJ)

RESUMO

O objetivo deste trabalho é apresentar uma visão geral da Linguística de Corpus, caracterizando-a como uma área do conhecimento; levando em consideração sua interface com outras áreas; e ilustrando suas aplicações, com foco mais específico no português do Brasil. Para atingir este objetivo, este artigo discute características da Linguística de Corpus que a distinguem de outras áreas, tais como: (1) a perspectiva de linguagem que adota e a forma de fazer pesquisas empíricas, com auxílio de ferramentas computacionais e com base em evidências linguísticas extraídas de corpora; (2) a possibilidade de trazer contribuições teóricas para os estudos da linguagem, através de novas descrições de diferentes usos da língua; (3) as interfaces de pesquisa com outras áreas, tais como Linguística Sistêmico-Funcional, Linguística Aplicada e Linguística Computacional; (4) o desenvolvimento da área, inclusive no Brasil, devido às novas perspectivas que possibilita em relação à lexicografia, léxico-gramática, estudos da variação linguística em gêneros discursivos e estudos interculturais. Através da discussão dos pontos acima, espera-se indicar que a Linguística de Corpus é uma área que permite o aprofundamento sobre o conhecimento empírico de diferentes línguas estudadas, levando a novas concepções teóricas sobre a linguagem, não podendo ser considerada, portanto, apenas como uma metodologia de análise. No final do trabalho, serão brevemente apresentadas três pesquisas que incluem dados da língua portuguesa, visando exemplificar aplicações da Linguística de Corpus para o estudo do uso do português. Esses trabalhos foram desenvolvidos a partir do CORPOBRAS PUC-Rio, compilado com o objetivo de ser um corpus representativo do português do Brasil[1].

PALAVRAS-CHAVE: linguística de corpus, teoria e corpus, pesquisa empírica, corpus do português do Brasil, CORPOBRAS PUC-Rio.

1. Linguística de Corpus: caracterização da área

A Linguística de Corpus pode ser considerada como "a face moderna da linguística empírica" (TEUBERT, 1996, p. vi), sendo a linguagem vista como um fenômeno social e analisada a partir de atos concretos de comunicação, isto é, textos reais, buscando o significado onde este é negociado, ou seja, no discurso. Esta perspectiva própria sobre a linguagem, fenômeno que estuda, e uma maneira específica de fazer pesquisa, ou seja, através do estudo de textos reais, com o auxílio de programas de computador, visando extrair evidências linguísticas do corpus, levam-nos a considerar este campo de estudos como uma área do conhecimento com suas próprias bases teóricas e uma maneira específica de fazer análises linguísticas.

Esta área representa uma nova abordagem filosófica para os estudos da linguagem. Svartvik (1996) concorda com Leech, que afirma que "a linguística de corpus não define somente uma metodologia emergente para o estudo da linguagem, mas uma nova maneira de fazer pesquisa, e de fato uma nova abordagem filosófica para este assunto. O computador, como uma ferramenta tecnológica de poder indiscutível, tornou este novo tipo de linguística possível" (LEECH, 1992, p. 106 citado em SVARTVIK, 1996, p. 12). Entretanto, cabe aos linguistas, com suas próprias intuições sobre a língua, instruir estes programas para extrair as evidências linguísticas com as quais irão trabalhar.

Um corpus linguístico de base computacional corresponde a coleções de textos que ocorrem naturalmente na língua, organizadas sistematicamente para representar áreas de uso da língua, e das quais podemos extrair novas informações (BIBER, 1995, p. 31). Hunston (2002, p. 23) diz que "a corpus can offer evidence, but can not give information", isto é, um corpus pode oferecer evidências, mas não pode dar informações. São os linguistas que produzirão novas informações, teóricas ou aplicadas, a partir do corpus.

Por outro lado, análises feitas com auxílio de programas de computador podem também levar a novas descobertas sobre aspectos linguísticos até então não considerados como relevantes pelos pesquisadores, visto que evidências não esperadas podem emergir dos dados. Para que isso possa acontecer, ou seja, para que estas evidências sejam percebidas, alguns linguistas envolvidos com estudos de corpus têm enfatizado que é preciso confiar no texto – "trust the text" (SINCLAIR,

1994), para observá-lo da forma mais isenta possível, deixando que os dados sejam a base para novas descrições e análises, que poderão levar a novas descobertas teóricas.

Em 1993, Halliday já havia se surpreendido com alguns pesquisadores que faziam uma oposição entre a linguística de corpus e a linguística teórica, como se fossem duas espécies distintas. Para ele, naquela época, a Linguística de Corpus já era considerada como uma empreitada altamente teórica:

> o trabalho baseado em corpus já começou a modificar nosso pensamento sobre o léxico, sobre padrões no vocabulário das línguas; e ele está agora começando a causar impacto nas nossas idéias sobre a gramática. No meu ponto de vista, este impacto será completamente benéfico. A linguística de corpus traz recursos novos e poderosos para as investigações teóricas sobre a linguagem. Uma consequência do desenvolvimento de corpora modernos é que agora podemos, pela primeira vez, desenvolver um sério trabalho quantitativo no campo da gramática (HALLIDAY, 1993, p. 1).

Recentemente, Halliday e Matthiessen (2004, p. 34) reafirmam esta posição ao dizer que "o corpus é fundamental para a empreitada de teorizar sobre a linguagem". Para estes autores, entretanto, muitos linguistas especializados em estudos de corpus referem-se a si mesmos, intencionalmente, como *'meros compiladores de dados'*, embora estejam conscientes da importância teórica do que estão fazendo e do que estão descobrindo[2]. Como novos dados que surgem a partir do corpus podem criar problemas para as teorias, alguns preferem manter a dicotomia teoria – dados, quando seria mais adequado considerar uma complementariedade entre teoria e dados, cada lado constantemente alimentando e redefinindo o outro (idem, p. 35-36).

Além disso, para alguns pesquisadores que não conhecem bem a Linguística de Corpus esta se restringe a resultados numéricos extraídos do corpus! Há também pesquisadores de corpus que apresentam resultados estatísticos sem discussões complementares ou confrontações com resultados anteriores. Estes dois grupos estão equivocados ao pensarem que bastam os números ou as estatísticas para descrever fatos linguísticos, já que, para interpretar os dados , com base no corpus, muitas vezes temos que levar também em conta o cotexto e os aspectos sócioculturais que estão ligados aos textos. Segundo McCarthy (1998, p. 1), por exemplo, os seus trabalhos de corpus baseiam-se ocasional-

mente em dados quantitativos, mas na maioria das vezes, este pesquisador observa os dados do corpus qualitativamente, porque é nesta abordagem que vê o maior potencial para reunir *insights* pedagógicos, que fazem parte de seu foco de estudo.

Por outro lado, os estudos de corpus caracterizam-se pela busca de tendências, probabilidades ou padrões de ocorrência ao lidarem com grande quantidade de dados. Nesses casos, os números servem de base para que estes padrões possam ser identificados e, então, interpretados pelos pesquisadores. Os resultados quantitativos produzidos com base no corpus são assim indicadores numéricos que devem ser discutidos à luz de diferentes posicionamentos teórico-metodológicos, para serem compreendidos. Da mesma forma que o corpus oferece apenas evidências linguísticas, e não informações, os números extraídos dos dados linguísticos não são ainda informações em si mesmos, precisando ser interpretados pelo pesquisador para que possam servir de apoio para novas descrições linguísticas ou para a proposta de novas perspectivas teóricas.

Se considerarmos que uma teoria pode ser entendida como uma perspectiva sob a qual um fenômeno é observado, entenderemos facilmente o porquê de existirem múltiplas teorias de linguagem, que correspondem a diferentes maneiras de se olhar esse mesmo objeto de estudo. Para Bernstein (1996, p. 93) "uma teoria deve ser capaz de oferecer uma descrição explícita e não ambígua dos objetos de sua análise... a teoria deve especificar o que será investigado e como os dados serão investigados e descritos". Hasan (1999, p. 13) observa que há dois tipos de teorias: endofóricas e exofóricas. Uma teoria endofórica está centrada no seu objeto de estudo, isolando-o dos diversos universos da experiência humana; uma teoria exofórica, por outro lado, não está limitada dentro das fromteiras de seu objeto de estudo, vendo-o em relação a outros universos da experiência humana, alterando-se e sendo alterada através de sua relação com outros domínios. Hasan acrescenta que, "como consequência dessas constantes trocas, o objeto de estudo em teorias exofóricas parece estar sempre em movimento, apresentando uma faceta diferente de acordo com cada mudança de ponto de vista por parte do observador" (HASAN: 1999, p. 13).

A Linguística de Corpus, como já mencionado, apresenta a sua própria perspectiva de linguagem, em que essa é vista sob seu aspecto de uso, observada em textos reais e analisada empiricamente. Podemos considerar que esta área também vê seu objeto de estudo, a linguagem,

sempre em movimento, como *'um sistema dinâmico aberto'* (LEMKE, 1993 citado em HASAN, 1999, p. 13) que se relaciona com diferentes domínios e está sujeito a diferentes pontos de vista, dependendo do pesquisador. Assim, propomos que a Linguística de Corpus seja considerada como uma teoria exofórica, que se completa com os pontos de vista de outras teorias, também exofóricas e com as quais estabelece interfaces, já que todas elas vêm a linguagem relacionada a diversos universos da experiência humana.

2. Linguística de Corpus: Interfaces

A Linguística de Corpus situa-se na interdisciplinaridade e na complementaridade, relacionando-se com outras áreas do conhecimento, teorias ou abordagens linguísticas, que ao somarem conhecimentos, poderão contribuir para um melhor conhecimento do seu objeto comum de estudo que é a linguagem. Assim, podemos observar pontos de contato entre Linguística de Corpus, Linguística Sistêmico-Funcional (LSF), Linguística Aplicada (LA), Linguística Computacional (LC), dentre outras áreas.

A relação entre a Linguística de Corpus e a Linguística Sistêmico-Funcional (HALLIDAY, 1994, HALLIDAY e HASAN, 1989, HALLIDAY e MATTHIESSEN, 2004) pode ser observada na abordagem teórica e metodológica das duas áreas. Em termos teóricos, o aspecto social da linguagem é privilegiado em ambas, sendo valorizado o seu uso e sua funcionalidade. Além disso, para ambas as áreas, a análise deve ser feita a partir de textos. Na LSF, o contexto situacional assume papel determinante tanto para a produção como para a análise textual. Na Linguística de Corpus, trabalha-se com textos reais, ou seja, textos que ocorrem naturalmente na língua, os quais, no corpus, entretanto, estão fora de seu contexto, sendo apenas oferecido aos analistas, geralmente, o seu co-texto (HUNSTON, 2002, p. 23). Em alguns casos, corpora bem documentados, que incluem informações ou classificações complementares em relação ao assunto ou época de produção dos textos, autores ou participantes em interações[3], permitem ao pesquisador recuperar parcialmente o contexto situacional e/ou cultural em que os textos se desenvolveram, mas isso nem sempre é possível. Contudo, esse não parece ser um problema teórico relevante para a Linguística de Corpus, cujas preocupações estão mais voltadas para a identificação de padrões do que para as descrições de usos particulares da língua em situações específicas.

Na LSF, a noção de sistema faz com que seja possível considerar que um falante/escritor, em determinadas condições, possa fazer certas escolhas paradigmáticas e não outras, dentro das possibilidades oferecidas; nesse caso podemos dizer que haverá probabilidades de escolha por um ou outro elemento do sistema. Considerando-se aspectos metodológicos, de modo semelhante, na Linguística de Corpus há interesse em identificar, por exemplo, as probabilidades de colocação de algumas palavras com outras em determinados contextos de uso da língua, sendo para isso utilizados programas computacionais específicos, como os *concordancers*[4]. Há também outras ferramentas computacionais, que visam analisar corpora com base na teoria sistêmico-funcional e que podem fazer investigações no nível da léxico-gramática [5].

A complementaridade entre as duas áreas pode ser notada em algumas pesquisas, em sua abordagem teórica e na análise de dados, conforme exemplificaremos, brevemente, no final deste artigo. Vários estudos de corpus têm sido desenvolvidos usando a teoria sistêmico-funcional como base para a explicação de evidências linguísticas trazidas pelo corpus. Estes estudos têm focos variados, embora a maioria dos trabalhos tome como ponto de partida as evidências lexicais ou léxico-gramaticais.

A relação entre a Linguística de Corpus e a Linguística Aplicada (LA) vem sendo enfatizada de maneira recorrente por linguistas aplicados. Em 1992, em sua *Introduction to Applied Linguistics*, Robert Kaplan e William Grabe incluíram um capítulo de autoria de Douglas Biber sobre as aplicações do computador na linguística aplicada, no qual vários trabalhos de corpus são descritos (BIBER, 1992). No mesmo volume, Grabe (1992, p. 294) afirma que para se tornar um linguista aplicado um pesquisador deve conhecer bem a linguística e outras áreas afins, mas que para funcionar bem na sua própria área deve também ter conhecimentos no uso de computadores e familiaridade com habilidades ligadas à quantificação, para poder desenvolver bases de dados e análises de corpus (GRABE e KAPLAN, 1992, p. 294). Recentemente, Kaplan (2002) afirmou que a Linguística de Corpus está ligada aos desenvolvimentos futuros da LA, prevendo para essa última uma maior ligação com a linguística descritiva (idem, p. 514). Para ele o desenvolvimento da Linguística de Corpus

> está revelando fatos a respeito do uso da linguagem e da variação entre registros que são essenciais para se lidar com questões práticas

> mas que são, muitas vezes, não compatíveis com a maioria dos modelos teóricos da Linguística. Os linguistas aplicados, que devem estar ancorados em uma 'linguística realista', que seja baseada no discurso e comprovada por ocorrências, provavelmente se deslocarão para a análise de novos dados, ao invés de continuarem a argumentar por uma nova teoria, apesar do fato de que a construção de novas teorias possa não só ser possível, mas desejável em uma abordagem descritiva (KAPLAN, 2002, p. 514).

Outros autores e outras publicações têm também mostrado a relação entre a Linguística de Corpus e a LA, tais como Martin Bygate (2004, p. 7), ao incluir a Linguística de Corpus nas futuras tendências de pesquisa da Linguística Aplicada; William Grabe (2004, p. 110), ao incluir a Linguística de Corpus como uma área de pesquisa da LA, que, nesta posição, vem se destacando há mais de 15 anos; Ulla Connor e Thomas Upton (2004), ao organizar o volume *Applied corpus linguistics: a multidimensional perspective*, que inclui capítulos sobre estudos de corpus voltados para a análise do discurso oral e escrito e aplicações pedagógicas de corpora; Susan Hunston (2002), ao publicar o livro *Corpora in Applied Linguistics,* que tem foco na relação entre as duas áreas, e mais especificamente no ensino de línguas; e Michael McCarthy (1998), ao reunir seus trabalhos sobre corpora no livro *Spoken Language & Applied Linguistics*, que está baseado em pesquisas a partir do *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE).

A interface entre a Linguística de Corpus e a LA deve-se também à relação existente entre as subáreas dessa última com a primeira. Neste sentido, por exemplo, o ensino e aprendizagem de línguas, envolvendo setores como língua estrangeira, língua para fins específicos, letramento em língua materna e estrangeira, linguagem e cultura, etc, têm gerado pesquisas de corpus ligadas à análise aplicada do discurso, gramáticas, e materiais de ensino, dentre outras.

Aplicações pedagógicas de estudos de corpus podem ir além das descrições linguísticas, tendo impacto direto no planejamento de currículos e nas práticas pedagógicas ligadas ao ensino de línguas. Estas aplicações ilustram a interface entre a Linguística de Corpus e a Linguística Aplicada e trazem à tona, por exemplo, pontos mais relevantes e realistas da gramática para o estudo em sala de aula. Outro aspecto que também vem sendo discutido é o uso do corpus diretamente com os alunos em sala de aula. Embora haja aqueles que aconselhem cautela quanto a essa prática, uma vez que consideram que dados da 'língua

externalizada' não deveriam ser sempre privilegiados nas situações de ensino e aprendizagem (WIDDOWSON, 2000, 2003 citado em GRABE 2004), muitos outros autores têm produzido materiais cuja finalidade é o uso do corpus para o ensino, ou a discussão de questões que relacionam ensino e corpus (SINCLAIR, 2003, 2004; WICHMANN et al 1997).

Os estudos tradutórios também podem ser vistos em interação com a Linguística Aplicada (KAPLAN e GRABE, 1992, p. 22) e em muito têm se beneficiado da Linguística de Corpus, especialmente através de estudos de lexicografia. Muitos corpora vêm sendo compilados para serem usados como apoio 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*. Além disso, os tradutores brasileiros podem se beneficiar de corpora do português, como o da Linguateca desenvolvido em Portugal, e que abriga corpora também de português do Brasil. Corpora paralelos também são de grande utilidade na pesquisa de soluções terminológicas ou gramaticais, assim como corpora especializados, que podem ser muito úteis em traduções técnicas em áreas específicas, tal como um corpus de textos de Química (UFRGS).

Quanto à Linguística Computacional (LC), esta se relaciona à Linguística de Corpus por ambas basearem-se no corpus para buscar evidências linguísticas; por suas características ligadas à tecnologia; e por focalizarem o uso de linguagem em seus estudos linguísticos. Entretanto, seus objetivos são diferentes, já que a "Linguística Computacional explora relações entre as áreas de linguística e informática, tornando possível a construção de sistemas com a capacidade de reconhecer e produzir informação apresentada em língua natural" (VIEIRA e STRUBE DE LIMA, 2001). Como muitos trabalhos nessa área estão voltados para o processamento da linguagem natural, isto é, construção de programas capazes de interpretar e/ou gerar informações em linguagem natural, a Linguística Computacional utiliza os corpora para poder ter acesso ao material que necessita estudar, ou seja, grande quantidade de textos que ocorrem naturalmente na língua.

No Brasil, a maioria desses programas vem sendo desenvolvida por pesquisadores da área de informática, interessados em pesquisas sobre inteligência artificial, em colaboração, muitas vezes, com linguistas da área de linguística computacional. Trabalhos que visam o estudo do português têm sido desenvolvidos, por exemplo, em algumas

instituições acadêmicas no Brasil (UNICAMP, USP, UFRS, UFMG) com focos variados no léxico, ortografia, léxico-gramática, etiquetagem, ou análise sintática. Entretanto, segundo Vieira e Strube de Lima (2001) há ainda no Brasil uma carência de pesquisas, ferramentas e recursos para o desenvolvimento da área, que conta com mais trabalhos voltados para o inglês, espanhol, alemão e francês, do que para o português.

3. Linguística de Corpus: desenvolvimento da área

A área de Linguística de Corpus vem se desenvolvendo há mais de 40 anos, quando os primeiros corpora foram compilados. O primeiro deles, o *Brown Corpus,* que data do início dos anos 60, foi desenvolvido na Universidade de Brown, nos Estados Unidos e contém 1 milhão de palavras de inglês americano. Um corpus de inglês britânico, o *Lancaster-Oslo/Bergen Corpus (LOB*), de tamanho e formato compatíveis com o americano, foi desenvolvido na Inglaterra, em Lancaster, e na Noruega, em Oslo e Bergen, e começou a ser usado em meados da década de 70.

Entretanto, foi a partir dos anos 80 que a área expandiu-se devido a condições favoráveis em diferentes aspectos: sócio-históricos, acadêmicos, tecnológicos e pragmáticos. O engajamento de importantes linguistas britânicos e americanos na organização de corpora foi um dos principais motivos da expansão da área. Pesquisadores como Geoffrey Leech, Jan Svartvik, John Sinclair, Randolph Quirk e Douglas Biber, foram alguns dos linguistas responsáveis pelo desenvolvimento, respeitabilidade e divulgação da área no meio acadêmico. Muitos desses eminentes linguistas são também, e não por acaso, gramáticos da língua inglesa, podendo-se imediatamente depreender as inúmeras possibilidades que os corpora podem abrir às descrições gramaticais e ao desenvolvimento de teorias gramaticais a partir de novas evidências da língua em uso.

Outro componente importante no desenvolvimento da Linguística de Corpus foi o avanço da tecnologia, que permitiu o uso de computadores e de programas específicos para a análise de corpus, criando a possibilidade de armazenar, acessar e analisar grandes quantidades de dados linguísticos. O trabalho dos gramáticos que adotam o corpus como fonte de dados passou de fichas guardadas em caixas (SVARTVIK, 1996), nos anos 60, com exemplos de usos de palavras e estruturas, geralmente extraídas de textos escritos, para máquinas possantes capa-

zes de armazenar e processar, no século XXI, corpora de mais de 100 milhões de palavras, como o *British National Corpus (BNC),* composto de textos escritos e transcrições de textos orais.

A possibilidade de análise de grandes quantidades de dados que ocorrem naturalmente na língua, baseada na observação do uso da língua em contextos sociais e linguísticos diversos, tem aberto novas perspectivas para estudos aplicados de diferentes naturezas como estudos lexicográficos, léxico-gramaticais, tradutórios e de gêneros discursivos. Através de estudos lexicográficos com base em corpus, pode-se acompanhar o surgimento ou 'nascimento' de palavras em uma língua, como, por exemplo, aquelas ligadas à tecnologia, como 'deletar', já usada com bastante frequência em português. Os estudos tradutórios muito têm se beneficiado de corpora paralelos, como o corpus COMPARA, com textos em português e inglês[6]. Novas descrições gramaticais para fenômenos já bastante estudados, como o diminutivo em português, têm sido embasadas em corpus, evidenciando funções pragmáticas que se mostraram mais frequentes do que as semânticas, apresentadas em gramáticas tradicionais (TURUNEN, 2009). Há também contribuições para o ensino de línguas estrangeiras, por exemplo, através da descrição do uso dos auxiliares modais em um corpus de textos de alunos universitários brasileiros, onde os aprendizes de inglês como língua estrangeira parecem usar o modal *'can'* como um substituto genérico para vários outros modais do inglês, atribuindo-lhe uma função modalizadora 'guarda-chuva' (VIANA, 2008). Nos estudos de gêneros discursivos, a variação sincrônica e diacrônica em inglês tem sido descrita (BIBER e FINEGAN, 1989), bem como a variação intercultural em gêneros discursivos em português e inglês (OLIVEIRA, 2007).

No Brasil, o desenvolvimento da área de Linguística de Corpus aconteceu, principalmente, a partir dos anos 90, quando surgiram pesquisadores interessados em desenvolver estudos baseados em corpus e quando começam a aparecer algumas iniciativas para a organização de corpora do português. Em 2004, com a publicação no Brasil do primeiro livro sobre a área e a divulgação de informações sobre corpora e suas características, bem como das metodologias utilizadas para análise de corpus, os estudos nesta área ganharam força (SARDINHA, 2004). Entretanto, uma maior compreensão da área de Linguística de Corpus, em termos das contribuições teóricas que pode trazer para o conhecimento da linguagem e para a descrição do português do Brasil, parece

estar surgindo apenas nos últimos anos, em que pesquisadores e gramáticos interessados na descrição do português estão se voltando para o corpus de forma mais sistemática (NEVES, 1999; AZEREDO, 2008). Na medida em que a disciplina Linguística de Corpus vem sendo também incluída em programas de pós-graduação no Brasil[7], teses e dissertações que se baseiam nos conhecimentos da área estão aparecendo, muitas delas voltadas para o estudo do português do Brasil[8].

O desenvolvimento de corpora do português, no Brasil e em Portugal, também tem sido intenso, o que vem possibilitando o crescimento da área. De maneira geral, os corpora podem ser classificados como gerais ou especializados, sendo que os primeiros visam representar a língua de forma ampla e servir de base para pesquisas variadas; eles caracterizam-se pela sua variedade em relação aos gêneros discursivos que incluem, à variedade de registros, assuntos e autores. Os corpora especializados são coletados para objetivos específicos de pesquisa e consistem, muitas vezes, em coleções de textos de gêneros ou discursos específicos.

Todo corpus é uma amostragem de uma população da qual não conhecemos o tamanho (SARDINHA, 2004, p. 23), ou seja, o corpus representa uma porção limitada da língua, que é vista como um sistema potencial de significados (HALLIDAY, 1994). Como não se tem uma medida da proporção de usos de textos e discursos em uma comunidade de falantes/escritores da língua, cada corpus passa a ter apenas uma pequena parte do total de amostras potenciais da língua. Por isso, temos que considerar o corpus como um fragmento de língua, mas que, mesmo assim, representa o seu sistema global (ou parte dele) e que, mesmo incompleto e fragmentado, pode refletir as possibilidades de ocorrência de usos linguísticos potenciais (OLIVEIRA e DIAS, 2006).

No Brasil alguns corpora foram compilados, mas vários deles são especializados, como o da PUC-SP, de textos de comunicação no contexto de negócios, do Projeto DIRECT; e o corpus do Projeto NURC, com a fala culta de diferentes regiões do país, colhida em situações pré-estabelecidas. Apesar de terem sido tomadas outras iniciativas para a compilação de corpora em português, algumas extremamente bem-sucedidas, como o corpus do Núcleo Inter-institucional de Linguística Computacional – NILC (USP-São Carlos/ UFSCar/ UNESP), ainda não contamos com um corpus de dimensões abrangentes, que seja um corpus geral e representativo do português do Brasil.

Na PUC-Rio, em 2002, começamos a empreitada de montar um corpus que fosse representativo do português do Brasil, o CORPOBRAS PUC-Rio. Ao longo dos últimos oito anos como coordenadora deste projeto, pude contar com o auxílio de agências de fomento (ver nota explicativa 1), mas o corpus desenvolveu-se, principalmente, graças ao trabalho e contribuições de dados de alunos e professores do Departamento de Letras da PUC-Rio e colegas de outras instituições [9].

Em 2008, o CORPOBRAS ultrapassou a meta de 1.000.000 (hum milhão) de palavras, equiparando-se a corpora considerados como médio-grandes (SARDINHA, 2004, p.26), em relação ao seu tamanho[10]. Atualmente, o corpus é composto por 27 (vinte e sete) gêneros discursivos, distribuídos em: 20 (vinte) gêneros do discurso escrito, 5 (cinco) gêneros do discurso oral, e 2 (dois) gêneros do discurso escrito para ser falado[11]. O corpus totaliza 1.361 textos e 1.149.600 palavras, e contém, até o momento, os seguintes gêneros: artigos científicos, cartas ao editor, cartas de reclamação, cartas de recomendação, cartas pessoais, cartas profissionais, cartas profissionais acadêmicas, circulares, contos, crônicas, dissertações, editoriais, e-mails acadêmicos, e-mails pessoais, notícias de jornal, redações de alunos ensino médio, redações de alunos universitários, redações de vestibular, romances, teses, conversas cariocas, conversas de crianças, entrevistas acadêmicas, grupos de enfoque, atendimento ao cliente, discursos políticos e roteiros cinematográficos.

O objetivo do CORPOBRAS é que ele possa servir a uma descrição ampla da língua ou a análises específicas. Por isso, tivemos cuidados especiais em fazê-lo representativo do português do Brasil, levando em conta que a montagem de um corpus representativo de uma língua requer o armazenamento de amostras de vários gêneros do discurso oral e escrito. Para criarmos um corpus representativo do português do Brasil, acreditamos que devemos considerar, principalmente, que os textos devem ser: reais, 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; e, principalmente, distribuídos em gêneros discursivos variados para representar a maior variedade possível de ações sociais (OLIVEIRA e DIAS, 2006).

No Brasil, onde a pesquisa linguística tem se desenvolvido com muita rapidez, esperamos que o CORPOBRAS PUC-Rio possa servir de base tanto a estudos linguísticos teóricos como aplicados para a descrição do discurso oral e escrito em português[12]. Entretanto, sabemos que ainda há muito trabalho a ser feito, em termos de compilação e organização dos dados já coletados!

4. Aplicações: Estudos de corpus

Nos estudos de corpus, muitas vezes, o pesquisador utiliza o corpus para ajudar a estender uma descrição linguística, mas, ao fazê-lo, deixa abertas as possibilidades de mudanças na teoria, podendo as evidências do corpus tornarem-se mais importantes do que as categorias teóricas ou descritivas anteriores. Por isso, acredito que não seja necessário classificar as pesquisas de corpus em "baseadas em corpus" (*'corpus based'*) e aquelas "dirigidas por corpus"(*'corpus driven')* (TONIGNI-BONELLI, 2001). Ao invés de dividi-las em dois grupos, considero mais adequado aceitar as duas perspectivas como misturadas, sem que haja, portanto, a necessidade de classificar os estudos de corpus em uma ou outra perspectiva, já que, em estudos de corpus, podemos chegar a conclusões sobre uma proposição descritiva, com consequências teóricas.

Cabe ainda ressaltar algumas outras características gerais de estudos de corpus. Uma delas é que eles podem ser desenvolvidos de acordo com abordagens metodológicas diversas que visam acessar, analisar ou contrastar dados em corpora. Muitas abordagens podem ser aplicadas ao corpus, dependendo do objetivo e do escopo da pesquisa, incluindo, por exemplo, o cálculo da frequência de palavras, colocações, prosódia semântica, fraseologia, etc. Dentre as metodologias de estudo de corpus podemos mencionar a Análise Multidimensional (BIBER, 1988; CONRAD e BIBER, 2001). Vários são os estudos multidimensionais: estudos diacrônicos e sincrônicos (BIBER e FINEGAN, 1989; GRABE, 1987); estudos em uma língua, como o inglês (BIBER, 1988), coreano (KIM e BIBER, 1994), somali (BIBER e HARED, 1994), nukulaelae tuvaluan (BESNIER,1988); ou contrastivos (OLIVEIRA, 1997; BIBER, 1995)[13].

Gostaria novamente de enfatizar que acredito que a Linguística de Corpus não pode ser considerada, ela mesma, apenas como uma metodologia de análise. Com base no fato de que há diferentes

metodologias que podem ser usadas em estudos de corpus; que os estudos de corpus desenvolvem pesquisas empíricas com características próprias e apresentam maneiras variadas para a descrição de fenômenos linguísticos, as quais podem gerar teorias, podemos afirmar que a Linguística de Corpus é muito mais do que uma metodologia, constituindo-se em uma área do conhecimento com suas próprias características teóricas e aplicações práticas.

A maioria dos estudos desenvolvidos a partir de *corpora* toma como base o léxico (KENNEDY, 1998, p. 90) , ou seja, baseiam-se em palavras isoladas, grupos de palavras, ou em sua relação com outras (ex: colocações, *chunks,* palavras chave). Esta tendência pode ser atribuída ao fato de que há maior disponibilidade de programas que auxiliam neste tipo de análise (*'concordancing'*); por outro lado, a etiquetagem, ou identificação automática de classes das palavras *('tagging'*), e a análise da função sintática das palavras (*'parsing'*) são mais complexas, e por isso custaram mais a serem viabilizadas. Entretanto, nos últimos anos, foram desenvolvidos e disponibilizados vários programas capazes de fazerem a marcação gramatical automática de um corpus, alguns capazes de desenvolver análises do português, como o Unitex (PAUMIER, 2006) e Palavras (BICK, 2002).

Um problema enfrentado pela Linguística de Corpus é que ela designa uma empreitada coletiva, compreendendo vários trabalhos independentes, ou seja, há coleções de trabalhos independentes que descrevem diferentes aspectos das línguas, mas que não estão sistematicamente organizados (KENNEDY, 1998, p. 88). Entretanto, se postos todos juntos, formam já um corpo bastante representativo de conhecimentos gramaticais em diferentes línguas.

Um exemplo importante de uma descrição gramatical abrangente do inglês, a partir de corpus, é a *Longman Grammar of Spoken and Written English* (BIBER, JOHANSON, LEECH, CONRAD & FINEGAN, 1999), que se baseia em um corpus de 40 milhões de palavras, representando quatro variedades da língua: conversas face-a-face, textos de jornais, ficção e prosa acadêmica. O objetivo desta gramática é descrever, a partir de pesquisas empíricas, o uso real de traços gramaticais, aí incluídas as classes gramaticais, estruturas frasais, componentes oracionais e outras categorias gramaticais. A frequência e distribuição de traços linguísticos nas variedades linguísticas selecionadas servem de base para explicações sobre o uso desses traços, sendo também con-

siderados elementos do contexto situacional, como a finalidade da comunicação, o modo oral ou escrito, e outras condições de produção (BIBER et al, 1999, p. 5).

Em relação ao português, um exemplo de descrição gramatical baseada em dados de uso real da língua é a *Gramática de Usos do Português,* desenvolvida por Maria Helena Moura Neves (1999) que, a partir do uso da língua em textos, descreve as funções gramaticais de outras unidades. Outro trabalho mais recente, também com base no uso da língua, é a *Gramática Houaiss de Língua Portuguesa* de José Carlos de Azeredo (2008), que descreve a variedade escrita do português, a partir de um corpus de textos de escritores, jornalistas ou autores brasileiros. Para Azeredo, ela é uma "fonte de informações sistematizadas sobre o português padrão do Brasil. Por isso, fazemos o registro da oscilação de usos correntes do corpus, deixando a escolha a critério do leitor/usuário que busca a informação" (AZEREDO, 2008, p. 26). Ainda outro trabalho considerado como relevante para a descrição gramatical do português é a *Gramática do Português Falado* (CASTILHO,1990), que é formada por "um conjunto expressivo de estudos.... descritivos da língua portuguesa" (AZEREDO, 2008, p. 36). Entretanto, as pesquisas contidas nos vários volumes que compõem essa última publicação não apresentam uma descrição sistemática da gramática do português falado, o que seria uma grande contribuição para os estudos da língua portuguesa. Mas, para que isso pudesse vir a acontecer, necessitaríamos também de um corpus abrangente e representativo do discurso oral em português do Brasil, que, infelizmente, ainda não está compilado, devido à dificuldade que tal empreitada representa, em termos de coleta e transcrição de dados.

Um outro aspecto relevante em relação aos estudos de corpus é que, como afirmamos anteriormente, estes estudos são primordialmente geradores de evidências linguísticas. Entretanto, na Linguística de Corpus, o uso da intuição linguística não está totalmente descartado (OLIVEIRA, 2007) e, por isso, nos vemos diante de um dilema: até que ponto podemos confiar em nossas intuições linguísticas para explicar algumas questões relativas ao uso da língua, e em que ocasiões as evidências linguísticas são essenciais? Algumas questões mais simples poderão ser respondidas com base apenas em nossas intuições, mas para responder outras mais complexas, entretanto, necessitaremos, sem dúvida, de recorrer às evidências linguísticas trazidas pela pesquisa

empírica desenvolvida com base em corpus. Como afirmamos acima, o corpus nos fornece as evidências, mas caberá ao linguista usar suas intuições e conhecimentos linguísticos para explicá-las. Assim, Conrad (2002), conclui que

> os estudos de corpus frequentemente são desenvolvidos a partir de questões que surgem de intuições ou observações casuais sobre a língua, e as interpretações dos achados extraídos do corpus frequentemente também incluem impressões intuitivas sobre o impacto de escolhas linguísticas específicas. Entretanto, o foco principal é empírico, baseado no que é observado no corpus (CONRAD, 2002, p. 77).

As evidências trazidas pelos dados reais de uso da língua podem chegar a provocar mudanças relevantes nos conhecimentos teóricos. Mas para que isso possa acontecer é preciso também que os estudos de corpus sejam desenvolvidos por pesquisadores com um sólido embasamento de conhecimentos linguísticos, teóricos e aplicados, para que possam perceber e demonstrar que conhecimentos produzidos anteriormente são incompletos, inadequados ou incorretos. Talvez seja por isso que linguistas aplicados, por exemplo, devam se aproximar mais da linguística descritiva e das teorias gramaticais para embasar seus trabalhos com corpus, conforme enfatizado por Kaplan (1992), e já mencionado neste trabalho.

Podemos resumir, então, algumas características dos estudos baseados em corpora: constituem-se em investigações da língua em uso; baseiam-se em coleções de textos selecionados de acordo com certos critérios; usam computadores para a análise automática ou interativa; incluem análises quantitativas e/ou interpretações qualitativas para descreverem padrões; possibilitam a análise de textos longos e variados; possibilitam o uso de um mesmo corpus para verificar ou procurar novos resultados; podem trazer subsídios para linguistas teóricos e aplicados; proporcionam maior precisão e credibilidade às análises quantitativas.

5. Estudos de corpus: aplicações a partir do CORPOBRAS

Neste trabalho vamos ilustrar, através de três trabalhos, a pesquisa desenvolvida a partir de corpus, com base em trabalhos ligados ao CORPOBRAS PUC-Rio, os quais são baseados em descrições diversas de uso do português do Brasil, dois deles em interface com a Linguísti-

ca Sistêmico-Funcional. Estes trabalhos estão ligados à lexicografia, léxico-gramática, gêneros discursivos, estudos inter-culturais e da variação linguística sincrônica. Alguns destes trabalhos fizeram uso mais extenso do CORPOBRAS e de análises automáticas com auxílio do computador. Outros utilizaram coleções de textos extraídas do corpus, formando subcorpora, e fizeram uso menos intenso do computador para extrair as evidências linguísticas, mesmo assim produzindo resultados quantitativos e qualitativos a partir do corpus. Dois destes trabalhos foram desenvolvidos como dissertações de mestrado (LANZIOTTI, 2002 e CALDEIRA, 2006), no Departamento de Letras da PUC-Rio. Em um dos trabalhos exemplificado abaixo (OLIVEIRA, 2006) a descrição apresentada foi contrastada com o inglês[14].

1 - *Variação de gêneros discursivos: a explicitação do contexto em um corpus do português escrito* (LANZIOTTI, 2002)

Este trabalho tem como foco o estudo da variação sincrônica de gêneros escritos da língua portuguesa, com abordagem multidimensional (ver nota explicativa 13). Esta abordagem foi também utilizada por (OLIVEIRA, 1997), para o estudo de um corpus de 270 redações de alunos universitários, produzidas em dois contextos culturais diversos, no Brasil e nos Estados Unidos, e divididas em 3 grupos: inglês (L1), português (L1) e inglês como língua estrangeira (L2). Uma das dimensões de variação que foram identificadas nesse corpus foi a *Explicitação do Contexto* (OLIVEIRA, 2002), a qual LANZIOTTI retomou e desenvolveu em sua pesquisa com 11 gêneros do Português escrito.

O corpus da pesquisa de Lanziotti compõe-se de 176 textos, sendo 16 amostras de 11 gêneros do Português escrito, que fazem parte do CORPOBRAS PUC-Rio. Os gêneros selecionados para formar o subcorpus da pesquisa foram: e-mail, carta pessoal, carta profissional, redação de aluno, artigo científico, editorial, notícia, circular, discurso político, romance e crônica. O corpus analisado totaliza aproximadamente 76.000 palavras. As evidências linguísticas consideradas são os sintagmas nominais em que o núcleo, ou o modificador, constituem referências culturais, históricas e geográficas; e sintagmas nominais em que o núcleo ou o modificador constituem referências sociais, econômicas e políticas (OLIVEIRA, 1997). Estas referências foram identificadas nos textos selecionados através de nomes próprios, identificados manualmente, e através de substantivos comuns, estes últi-

mos identificados com o auxílio do programa de buscas em contexto, *MonoConc Pro* (BARLOW, 1999).

Na pesquisa de LANZIOTTI, após o cálculo da frequência dos traços linguísticos no corpus e de sua normatização, médias e testes estatísticos foram aplicados ao corpus. Os resultados da pesquisa mostram que houve uma variação significativa dos gêneros escritos ao longo do contínuo *Explicitação do Contexto vs. Não-Explicitação do Contexto*, sendo que os gêneros notícia, editorial e discurso político estão mais próximos do pólo da Explicitação do Contexto, enquanto o e-mail, a crônica e a redação de aluno de ensino médio se aproximam da não-explicitação. Os resultados apontam para uma correlação entre a explicitação do contexto e o público alvo a que os textos se destinam, sendo mais explícitos quando o público é mais abrangente, havendo, portanto, menor compartilhamento de conhecimentos.

2 - *A redação do vestibular como gênero: configuração textual e processo social.* (CALDEIRA, 2006)

Este trabalho tem como foco o estudo de um gênero específico, com abordagem discursiva. O corpus da pesquisa compõe-se de redações de vestibular (N= 135) de quatro instituições, compiladas entre 2004 e 2005. As evidências linguísticas examinadas quantitativamente a partir do corpus de aproximadamente 30.000 palavras foram itens lexicais com referências exofóricas; nominalizações em –mento, - cão e -(c)ia; processos de diferentes tipos (HALLIDAY, 1994); e marcas de subjetividade, como pronomes pessoais de primeira pessoa. Estes itens foram identificados e quantificados com a ajuda do software *MonoConc Pro,* que faz buscas em contexto. Os resultados quantitativos da pesquisa ajudaram a caracterizar os significados ideacionais, textuais e interpessoais criados nas redações, mostrando que o mundo nelas representado é mais caracterizado por processos materiais e relacionais, onde predominam ações e relações; a baixa frequência de processos mentais pode indicar que o mundo representado nos textos é também mais objetivo e menos reflexivo. As nominalizações foram menos frequentes do que os processos, indicando que os textos dos alunos vestibulandos estão em pouca consonância com o discurso acadêmico, onde, segundo Basílio (1999, p. 25 citado em CALDEIRA, 2006), o processo da nominalização, entendido como o *'enquadramento do verbo em uma estrutura nominal'*, é recorrente. Por outro lado, a baixa ocorrência das marcas de subjetividade

nas redações pode estar indicando a aproximação da produção textual dos alunos em direção ao discurso acadêmico, onde as marcas interpessoais são deixadas de lado, muitas vezes por recomendação do ensino da escrita na escola.

3. *Grammatical metaphor in research articles: Linguistic and disciplinary contrasts* (OLIVEIRA, 2006)

Neste trabalho de corpus o foco é na léxico-gramática, em uma abordagem descritiva e interface com a teoria sistêmico-funcional. O corpus é composto de artigos de pesquisa em português e em inglês, totalizando 24 amostras de aproximadamente 1000 palavras cada uma, selecionadas de periódicos científicos nas áreas de Linguística e Nutrição. As evidências empíricas pesquisadas foram as nominalizações, consideradas como grupos nominais que podem funcionar como realizações metafóricas de configurações processuais, em lugar de orações, que seriam as formas mais congruentes (HEYVAERT, 2003). A análise dos dados incluiu o cálculo da frequência de nominalizações, identificadas no corpus através de buscas de palavras em contexto, ou concordâncias, com o auxílio do programa WordSmith Tools (SCOTT, 1999). Os sufixos formadores de nominalizações em português e em inglês (ex: -tion/ção, ssão; -ance,ence/-cia; -ment/mento; -er/dor), serviram de base para as buscas em contexto. A frequência dos textos foi normatizada para 1000 palavras e médias calculadas para o uso de cada sufixo, em cada língua. Testes estatísticos (MANOVA e ANOVA) foram calculados para verificar se a variação entre as médias obtidas para os grupos de textos, em relação à disciplina e à língua, era significativa. Os resultados da pesquisa indicam que os artigos de pesquisa produzidos por acadêmicos nas duas áreas variam quanto à frequência no uso de nominalizações. A variação entre as duas línguas mostra que os acadêmicos brasileiros tendem a usar mais nominalizações do que os americanos, especialmente na área de Linguística, em português, a qual apresentou mais ocorrências de nominalizações. Estes resultados podem ser relacionados com outros anteriores (MORAES, 2005) que mostraram que há uma maneira discursiva diferenciada entre as duas áreas, Linguística e Nutrição, de construir conhecimento, sendo os trabalhos de nutrição mais factuais e os de linguística mais voltados para as idéias, o que pode ser confirmado pelo uso de nominalizações.

Vários outros trabalhos têm sido desenvolvidos a partir do CORPOBRAS, alguns dos quais estão indicados a seguir: ALMEIDA, 2002; AMARANTE, 2002, 2008; CORRÊA, 2004; MORAES, 2005; OLIVEIRA, 1997, 1999, 2002, 2007, 2008; OLIVEIRA et al, 2009[15]; TURUNEN, 2009; VIANA, 2008. Estas pesquisas formam já um conjunto de informações extraídas de um corpus do português, algumas vezes em contraste com o inglês, que poderão contribuir para um conhecimento mais amplo da língua em uso.

6. Considerações finais

A Linguística de Corpus é uma área em expansão. Sua história ainda é recente, se comparada a outras subáreas da Linguística. Há, entretanto, fatores que poderão acelerar ou retardar o seu desenvolvimento. A seu favor está o fato de a área estar altamente relacionada ao uso de computadores. Como a tecnologia vem se desenvolvendo de maneira acelerada, em breve poderemos contar com máquinas ainda mais robustas, capazes de armazenar quantidades cada vez maiores de dados, tornando os corpora cada vez mais completos. Contudo, para analisá-los precisaremos de programas cada vez mais sofisticados e estes dependerão, para sua criação e desenvolvimento, que pesquisadores de diferentes áreas trabalhem em colaboração, o que é muitas vezes difícil, já que cada profissional é bastante exigido dentro de sua própria esfera de interesse e a interdisciplinaridade é, em muitos casos, ainda, uma proposta e não uma realidade.

Temos também que considerar o fato de a Linguística de Corpus ser uma ciência empírica, inserida em uma área maior do conhecimento, Letras e Linguística, onde a tendência, durante muitos anos, foi o foco em estudos teóricos. É preciso ainda convencer a muitos que precisamos de novos dados sobre a linguagem em uso para descrevê-la de forma mais adequada, de maneira a conhecer melhor o nosso objeto de estudo, e poder ensinar a língua de maneira mais eficiente aos seus aprendizes. Seria para isso necessário deixar de pensar que a Linguística de Corpus se restringe à compilação e coleta de dados, já que ao contribuir para a geração de novas descrições das línguas ela contribui também para que possamos conhecer novas gramáticas, que por sua vez nos levam a entender melhor a experiência humana tal como é construída na linguagem.

Uma teoria gramatical deveria ser sistemática, ou seja, ela deveria dar conta da língua em sua totalidade. Infelizmente, entretanto, não podemos dizer que a Linguística de Corpus tenha conseguido chegar a realizar este intento. Até o momento, temos uma série de estudos, alguns mais completos do que outros, que descrevem aspectos específicos das línguas. Podemos argumentar, entretanto, que esta área, ao desenvolver uma lógica direcionada pelos dados, uma observação meticulosa dos fatos ou evidências linguísticas, leva a avanços em direção à elaboração de uma teoria gramatical (TURUNEN, 2009), a qual poderá vir a ser proposta à medida que as pesquisas de corpus se consolidarem ou se organizarem em torno de um propósito descritivo mais sistemático.

Há ainda outros fatores que podem facilitar ou dificultar o percurso da área. A seu favor podemos mencionar o fato de que, em várias partes do mundo, ela tem ganhado notoriedade e que muitos corpora, em diferentes línguas, têm sido compilados. Contudo, estes projetos são trabalhosos e de longa duração; um corpus geral de uma língua necessita de muitos anos de trabalho de muitas pessoas, e instituições envolvidas, para ser viabilizado. O apoio financeiro para estes projetos também precisa ser robusto, para cobrir despesas com equipamentos, produtos e recursos humanos especializados. Estas duas condições, uma relativa a recursos humanos e outra a recursos financeiros, são difíceis de satisfazer, especialmente a segunda, já que, nos dias atuais, de maneira geral, os financiamentos para pesquisas na área de ciências humanas são escassos, e os projetos que envolvem o estudo de línguas não são vistos como prioritários. Porém, apesar das dificuldades encontradas, a área está em expansão no Brasil, na esfera acadêmica, onde, em vários centros do país, novos cursos são oferecidos e novos pesquisadores estão se especializando em Linguística de Corpus.

É essencial, entretanto, que a pesquisa em corpus não seja vista apenas como uma metodologia, e sim como uma abordagem teórica que permite múltiplas aplicações, para que conquiste cada vez mais espaços acadêmicos e políticos que possibilitem que ela cresça e continue a exercer a sua função primordial que é contribuir, empiricamente, para o conhecimento mais profundo, abrangente e teórico da linguagem e, em especial, do Português do Brasil.

Recebido em 14/04/09

Aprovado em 04/05/09

ABSTRACT

The purpose of this paper is to present an overview of Corpus Linguistics, characterizing it as an area of research, considering its relations with other areas of study and illustrating its applications with specific focus on Brazilian Portuguese. In order to develop these topics, this research paper discusses Corpus Linguistics characteristics by pointing out some issues that distinguish it from other areas of research, such as: (1) its specific way to define language as well as a particular form to do empirical research on the basis of evidence extracted from linguistic corpora, using computational tools; (2) the possibility to generate theoretical contributions through new descriptions of different language uses; (3) the interfaces it establishes with Systemic-Functional Linguistics, Applied Linguistics and Computational Linguistics; (4) the expansion of the area in many countries, including Brazil, due to new perspectives opened in several fields, such as, lexicography, lexicogrammatical studies, genre and language variation studies as well as cross-cultural studies. The discussion of the topics above should reinforce the argument that Corpus Linguistics cannot be considered only as a methodological approach, but rather as a research area that allows for empirical linguistic knowledge, leading into new theoretical insights about language. In order to illustrate some corpus research done within the scope of Corpus Linguistics using data from the Portuguese language, three empirical studies are briefly described at the end of this paper. These academic works used data from the CORPOBRAS PUC-Rio, a corpus compiled with the purpose of representing Brazilian Portuguese.

KEY WORDS: corpus linguistics, theory and corpus, empirical research, Brazilian Portuguese corpus, CORPOBRAS PUC-Rio.

REFERÊNCIAS

ALMEIDA, P.M.C. Atendimento de *check-in* de companhia aérea: Análise sistêmico-funcional de um gênero discursivo do português. Dissertação (Mestrado em Estudos da Linguagem). Departamento de Letras, PUC, RJ, 2002. 193 f.

AMARANTE, R. M. C Começando do princípio: Uma análise do *lead* como subgênero discursivo em português e em inglês. Dissertação (Mestrado em Estudos da Linguagem). Departamento de Letras, PUC, RJ, 2002. 109 f.

AMARANTE, R. M. C. Heróis de papel: Uma abordagem sistêmico-funcional da imagem do jornalista projetada em notícias de guerra e esporte (Título provisório). Trabalho de Qualificação (Doutorado em Estudos da Linguagem). Departamento de Letras, PUC, Rio de Janeiro, 2008. 65 f.

AZEREDO, J.C. *Gramática Houaiss da língua portuguesa*. São Paulo: PubliFolha, 2008.

BADDINI, D.M. Estudos baseados em corpora: design, complementação e disponibilização de um corpus representativo do português do Brasil. *Anais do XII Seminário de Iniciação Científica da PUC-Rio*. Rio de Janeiro: PUC-Rio, 2004.

BADDINI, D.M Gêneros do discurso escrito: complementação e disponibilização de um corpus representativo do português do Brasil. *Anais do XIII Seminário de Iniciação Científica da PUC-Rio*. Rio de Janeiro: PUC-Rio, 2005, p. 423-424.

BARLOW, M. *MonoConc PRO* . Houston: Athelstan, 1998.

BASÍLIO, M.M.P. *Teoria lexical* . São Paulo: Ática, 1999.

BERNSTEIN, B. *Pedagogy, symbolic control and identity: theory, research, critique*. London: Taylor & Francis, 1996.

BESNIER, N. The linguistic relationships of spoken and written nukulaelae registers. *Language* 64, p. 707-736, 1988.

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

BIBER, D. Applied linguistics and computer applications. In GRABE, W. & KAPLAN, R. (eds). *Introduction to applied linguistics*. Reading, Massachusetts: Addison-Wesley, 1992. p. 257-278.

BIBER, D. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press, 1995.

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

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. & FINEGAN, E. *Longman*

grammar of spoken and written English. Essex, England: Pearson Education Limited, 1999.

BIBER, D., CONRAD, S. & REPPEN, R. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press, 1998.

BIBER, D. & HARED, M. Linguistic correlates of the transition to literacy in Somali: Language adaptation in six press registers. In: BIBER, D. & FINEGAN, E. (eds.), *Sociolinguistic perspectives on register*. New York/Oxford: Oxford University Press, 1994. p.182-216.

BRITO, M. G. E VALÉRIO, R. G. (2007). Um corpus do Português do Brasil: variação entre gêneros discursivos. *Anais do XV Seminário de Iniciação Científica da PUC-Rio*. Rio de Janeiro: PUC-Rio. p 525-526.

BICK, E. *The parsing system PALAVRAS: automatic gramatical analysis of Portuguese in a constraint grammar framework*. Aarhus: Aarhus University Press, 2000.

BYGATE, M. Some current trends in applied linguistics: towards a generic view. *AILA Review,* 17, p. 6-22, 2004.

CALDEIRA, J. R. A redação de vestibular como gênero: configuração e processo social. Dissertação (Mestrado em Estudos da Linguagem). Departamento de Letras da PUC, Rio de Janeiro, 2006. 150f.

CASTILHO, A. T. (Org) *Gramática do português falado*. vol.1: A Ordem. Unicamp, 1990.

CONNOR, U. & UPTON, T. *Applied corpus linguistics*: a multidimensional perspective. Amsterdan: Rodopi, 2004.

CONRAD, S. Corpus linguistics approaches to discourse analysis. *Annual Review of Applied Linguistics,*22, p. 75-95, 2002.

CONRAD, S. & BIBER, D. *Variation in English*: multi-dimensional studies. New York: Longman, 2001.

CORRÊA, F. J. A. Cross-cultural rhetorical move analysis: letters to the editor in English and Portuguese. Monografia. Pós-Graduação Lato Sensu em Língua Inglesa. Rio de Janeiro: PUC-Rio, 2004. 85 f.

GRABE, W. Contrastive rhetoric and text type research. In: CONNOR, U. and KAPLAN, R. (eds.), *Writing across languages: analysis of L2 texts,* Reading, MA: Addison-Wesley, 1987. p. 113-137.

GRABE, W. & KAPLAN, R. (eds.) *Introduction to applied linguistics*. Reading, Massachusetts: Addison-Wesley, 1992.

GRABE, W. Becoming an applied linguist. In: GRABE, W. & KAPLAN, R. (eds). *Introduction to applied linguistics*. Reading, Massachusetts: Addison-Wesley, 1992. p. 281-300.

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

HALLIDAY, M. A. K. Quantitative studies and probabilities in grammar. In: HOEY, M. (ed.). *Data, description, discourse: papers on the English language in honour of John McH Sinclair*. London: HarperCollins Publishers, 1993. p.1-25.

HALLIDAY, M. A. K. *An introduction to functional grammar*. London: Edward Arnold, 1994.

HALLIDAY, M. A. K. & MATTHIESSEN, C. M.I.M. *An introduction to functional grammar* (3ª ed.). London: Hodder Arnold, 2004.

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

HASAN, R. Society, language and the mind: the meta-dialogism of Basil Bernstein's theory. In: CHRISTIE, F. (org), *Pedagogy and the shaping of consciousness: linguistic and social processes*. London: Continuum, 1999. p. 10-30.

HEYVAERT, L. Nominalization as grammatical metaphor: on the need for a radically systemic and metafunctional approach. In: SIMON-VANDENBERGEN, A.; TAVERNIERS, M. & RAVELLI, L. (eds.) *Grammatical metaphor: views from systemic functional linguistics*. John Benjamins: Amsterdam, 2003. p. 66-99.

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

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

KENNEDY, G. *An Introduction to corpus linguistics*. London: Longman, 1998

KIM, Y. & BIBER, D. A corpus-based analysis of register variation in Korean. In BIBER, D. & FINEGAN, E. (eds.), *Sociolinguistic perspectives on register*. New York/Oxford: Oxford University Press,1994. p.157-181.

LANZIOTTI, M.G. P. Variação de gêneros discursivos: a explicitação do contexto em um corpus do português escrito. Dissertação (Mestrado em Estudos da Linguagem). Departamento de Letras, PUC, Rio de Janeiro, 2002. 140 f.

MARQUES, G. O. Tecnologia e internet no ensino de língua estrangeira: avaliação discursiva de professores e alunos. Dissertação (Mestrado em Estudos da Linguagem). Departamento de Letras, PUC, Rio de Janeiro, 2006. 162 f.

McCARTHY, M. *Spoken language and applied linguistics*. Cambridge: Cambridge University Press, 1998.

MORAES, L. S. B. O metadiscurso em artigos acadêmicos: variação intercultural, interdisciplinar e retórica. Tese (Doutorado em Estudos da Linguagem), Departamento de Letras, Rio de Janeiro, PUC-Rio, 2005. 183 f.

NEVES, M.H.M. *Gramática de usos do português*. São Paulo: Editora UNESP, 1999.

OLIVEIRA, L. P. Variação intercultural na escrita: contrastes multidimensionais em inglês e português. Tese (Doutorado em Linguística Aplicada). LAEL, PUC, São Paulo, 1997. 358 p.

OLIVEIRA, L. P. Cross-cultural complexity-level variation in written discourse styles. Trabalho apresentado na *American Association for Applied Linguistics Annual Conference (AAAL*), Stanford, Connecticut, 1999.

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

OLIVEIRA, L. P. 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, 2006.

OLIVEIRA, L. P. Writing in the academic context: a corpus-based contrastive view. In: ZYNGIER, S.; VIANA, V. e JANDRE, J. (eds), *Textos e leituras: estudos empíricos de língua e literatura*. Rio de Janeiro: Publit, 2007. p 53- 64.

OLIVEIRA, L. P. (aceito para publicação). 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.; DIAS, M. C. P. Representatividade na compilação de corpus: o projeto CORPOBRAS PUC-Rio. Trabalho apresentado na *Jornada de metodologia para recolha e sistematização de corpora para fins dicionarísticos* . Rio de Janeiro: União Latina, 2006.

OLIVEIRA, L. P.; VALÉRIO, R. G.; BRITO, M. G. CORPOBRAS PUC-Rio: Um corpus do português do Brasil e análise do discurso acadêmico. Trabalho apresentado no *VIII Encontro de Ciência Empírica em Letras*. Rio de Janeiro: UFRJ, 2007.

PAUMIER, S. *Unitex, versão 1.2*. University of Marne-la-Vallée, França, 2006

SARDINHA, T. B. *Linguística de corpus*. São Paulo: Manole, 2004.

SCOTT, M. *WordSmith Tools*. Version 3. Oxford: Oxford University Press, 1999.

SINCLAIR, J. Trust the text. In: COULTHARD, M. (ed.), *Advances in written text analysis*. London: Routledge, 1994. p. 12-25.

SINCLAIR, J. *Reading concordances*. London: Pearson/Longman, 2003.

SINCLAIR, J. *How to use corpora in language teaching*. Amsterdam: John Benjamins Publishing Company, 2004.

SVARTVIK, J. Corpora are becoming mainstream. In: THOMAS, J. and SHORT, M. (orgs). *Using corpora for language research*. London and New York: Longman, 1996. p 3-13.

TEUBERT, W. Editorial. *International Journal of Corpus Linguistics*, Vol.1, No. 1. iii-x. 1996.

TONIGNI-BONELLI, E. *Corpus linguistics at work*. Amsterdam: John Benjamins, 2001.

TURUNEN, V. J. A reversão da relevância: aspectos semânticos e pragmáticos de formações diminutivas no português do Brasil. Tese (Doutorado em Estudos da Linguagem), Departamento de Letras. Rio de Janeiro: PUC-Rio, 2009. 198 f.

VALÉRIO, R.V. Um corpus do português do Brasil: variação entre gêneros discursivos. *Anais do XIV Seminário de Iniciação Científica da PUC-Rio*. Rio de Janeiro: PUC-Rio, 2006.

VALÉRIO, R.V. CORPOBRAS PUC-Rio: Desenvolvimento e análise de um corpus representativo do português. *Anais do XVI Seminário de Iniciação Científica da PUC-Rio*. Rio de Janeiro: PUC-Rio, 2008.

VIANA, V.P. Verbos modais em contraste: análise de corpus da escrita de universitários em inglês. Dissertação (Mestrado em Estudos da Linguagem). Departamento de Letras, PUC, Rio de Janeiro, 2008. 230 f.

VIEIRA, R. & STRUBE DE LIMA, V. L. Linguística computacional: princípios e aplicações. In: MARTINS, A.T. & BORGES, D.L. (org.) SBC - *Jornadas de Atualização em Inteligência Artificial* (JAIA). v. 3, p. 47-86, Fortaleza, 2001.

WICHMANN, A. FLIGELSTONE, S. MCENERY, T. & KNOWLES, G. *Teaching and language corpora*. London: Longman, 1997.

NOTAS

[1] Este projeto contou com apoio do CNPq, de 2004 a 2007, através de Edital Universal, (CNPq, processo 480143/2004-8), e de Bolsas de Iniciação Científica do CNPq/PIBIC (2004-2009) e da FAPERJ (2007).

[2] Consideramos que este mal estar teórico pode estar ligado ao fato de muitos pesquisadores da área de Linguística de Corpus não serem gramáticos ou linguistas, tendo sua formação acadêmica em outras áreas do conhecimento, como a Informática, etc. Por isso, muitas vezes, não querem comprometer-se com inovações ou novas descrições teóricas que possam ser contestadas por outros pesquisadores, especificamente da área de linguística.

[3] O Michigan Corpus of Academic Spoken English (MICASE) pode ser um exemplo de corpus bem documentado.

[4] Dentre os diversos programas com esta função, destacamos o WordSmith Tools, (SCOTT, 1999) para a análise de Corpus .

[5] Alguns pesquisadores como Christian Matthiessen, Mike O'Donnell e Tony Sardinha têm contribuído para o desenvolvimento de *software* específicos para

a descrição gramatical, em inglês e português, com base na teoria sistêmico funcional.

[6] O COMPARA, organizado pela Linguateca, em colaboração com Ana Frankenberg-Garcia, é um corpus paralelo bidireccional de português e inglês, ou seja, funciona como uma base de dados com textos originais nestas duas línguas e as suas respectivas traduções, ligadas frase a frase. Ele permite contrastar o português e o inglês através de pesquisas automáticas.

[7] Na PUC-Rio, por exemplo, a disciplina Linguística de Corpus vem sendo oferecida, desde 2005, embora somente a partir de 2010 deva passar a integrar a estrutura curricular do programa de pós-graduação na categoria de 'disciplina teórica'.

[8] Ver sites de diversas universidades que desenvolvam estudos de corpus, como PUC-SP e PUC-Rio, dentre outras.

[9] Graduandos de Letras da PUC-Rio participaram da compilação e organização do corpus, através de bolsas de Iniciação Científica (BADDINI,2004 - 2005; BRITO, 2006-2007; VALÉRIO, 2006-2009). Alunos de pós-graduação cederam os dados que coletaram para suas teses, dissertações ou monografias (ALMEIDA, 2002, AMARANTE, 2002, CALDEIRA, 2006, CORRÊA, 2004, LANZIOTTI, 2002, MARQUES, 2006, MORAES, 2005). Alguns colegas do Departamento de Letras cederam corpora de seus projetos ou dados coletados por seus alunos: Letícia Sicuro Corrêa, Maria do Carmo Leite de Oliveira, Maria das Graças Dias Pereira, dentre outros. Colegas de outras instituições, como Del Carmem Daher, também disponibilizaram dados para o CORPOBRAS., dentre outros.

[10] Os corpora representativos devem obedecer a padrões de extensão de acordo com a pesquisa a ser desenvolvida. Para Biber, Conrad & Reppen (1998, p. 249), em estudos de frequência de traços linguísticos, por exemplo, 10 amostras de textos de um gênero, com aproximadamente 2000 palavras, podem representar uma categoria lexical ou sintática e garantem resultados relativamente estáveis quanto ao uso da maioria dos traços linguísticos. Segundo os autores, entretanto, para estudos lexicográficos, deve-se contar com corpora mais extensos, já que algumas palavras ou colocações são pouco frequentes e somente um grande corpus viabilizará o seu estudo (Oliveira e Dias, 2006).

[11] Para solucionar certas situações em relação à classificação dos gêneros em um corpus, como no caso de discursos políticos e roteiros cinematográficos, alguns pesquisadores têm criado categorias novas em seus corpora, como por exemplo 'textos escritos para serem falados' (McCarthy, 1998, p. 9)

[12] O CORPOBRAS ainda não está disponível em sua totalidade. Atualmente, o corpus está em fase de organização em relação à documentação dos dados,

questões de autorizações autorais e elaboração de relatórios sobre textos e gêneros. Entretanto, subcorpora de diversos gêneros, já documentados, têm sido cedidos para pesquisas acadêmicas.

[13] Visando um estudo da variação linguística 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'. As dimensões são definidas através do agrupamento de traços linguísticos que co-ocorrem com frequência nos textos. Estas dimensões são identificadas estatisticamente através da 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 linguísticos 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).

[14] Para alguns gêneros discursivos do CORPOBRAS existem dados paralelos do inglês, o que vem permitindo o desenvolvimento de pesquisas contrastivas.

[15] Projeto 'Escrita e inclusão social: análise de corpus e a metáfora gramatical no Ensino Médio', que conta com apoio FAPERJ (2009-2010), através do Edital nº 26/2008 na área de Humanidades, processo E-26/112.269/2008. Será compilado e incorporado ao CORPOBRAS um subcorpus de textos de alunos de Ensino Médio a ser analisado com apoio das ferramentas computacionais Unitex e Palavras.