A abordagem metodológica da análise multidimensional

Tony Sardinha

Resumo

O presente artigo apresenta a perspectiva metodológica da Análise Multiaspectual Multidimensional (Multi-feature, Multi-dimensional Analysis) ou simplesmente Análise Multidimensional (AMD, BIBER, 1985 et seq.), que é uma das metodologias em uso na Linguística de Corpus (SINCLAIR, 1991; MCENERY e WILSON, 1996; BIBER, CONRAD e REPPEN, 1998; BERBER SARDINHA, 2004; 2005; TEUBERT e KRISHNAMURTHY, 2007; BAKER, 2009; BERBER SARDINHA, 2009; LÜDELING e KYTÖ, 2009). A AMD viabiliza a análise em larga escala de variação de corpora eletrônicos e permite chegar a uma classificação de registros ou gêneros em termos de dimensões, que são padrões de coocorrência de elementos lexicogramaticais que subjazem aos textos de uma língua (BIBER, 2009). Como tal, as dimensões capturam o espaço de variação dos textos, sintetizam-no e mostram a proximidade ou distância entre os textos investigados. A AMD apoia-se na análise estatística de co-ocorrência de grupos de variáveis linguísticas, anotadas de modo automático ou semiautomático.

Palavras-chave: *Linguística de Corpus, Análise Multidimensional, Variação*

Introdução

A Linguística de Corpus é a área de investigação linguística responsável pela coleta e análise de corpora eletrônicos, com o auxílio de ferramentas computacionais (SINCLAIR, 1991; MCENERY e WILSON, 1996; BIBER, CONRAD e REPPEN, 1998; BERBER SARDINHA, 2004; 2005; TEUBERT e KRISHNAMURTHY, 2007; BAKER, 2009; BERBER SARDINHA, 2009; LÜDELING e KYTÖ, 2009). Corpora, por sua vez, são grandes quantidades de texto, coletadas criteriosamente e mantidas em formato de computador, com o propósito de servirem à pesquisa linguística (BERBER SARDINHA, 2004). São formados, por definição, por textos de um ou mais gêneros (ou outra variante discursiva).

O presente artigo enfoca uma das metodologias da Linguística de Corpus, mais especificamente, a Análise Multiaspectual Multidimensional (Multi-feature, Multi-dimensional Analysis) ou simplesmente Análise Multidimensional (AMD, BIBER, 1985 *et seq*.). A AMD viabiliza a análise de variação em corpora e permite chegar a um classificação detalhada e abrangente de registros ou gêneros e das relações que estabelecem entre si. Tal classificação é operacionalizada por meio da anotação automática e semi-automática de variáveis relevantes para a caracterização dos gêneros, com o subsequente agrupamento e interpretação dessas variáveis em fatores, que são conjuntos de textos que possuem padrões de coocorrência de variáveis definidos estatisticamente.

A AMD opera com o conceito de registro, que significa 'uma variedade linguística definida por aspectos situacionais, incluindo o propósito do falante, a relação entre falante e ouvinte, e o contexto de produção'[1] (BIBER, 2009 , p. 823), podendo indicar desde gêneros específicos, como artigos acadêmicos, quanto variedades mais gerais, como 'documentos oficiais' ou 'discurso acadêmico'. Dimensões de variação são padrões de coocorrência de elementos lexicogramaticais que subjazem aos registros de uma língua (BIBER, 2009). Como tal, capturam o espaço de variação dos textos, sintetizam-no e mostram a proximidade ou distância entre os registros investigados. Um exemplo de dimensão de variação (da língua inglesa) é 'Interação versus Informatividade' (BIBER, 1988), que indica que todos os textos dessa língua possuem essas características essenciais, que são a interação, de um lado, e a informatividade, de outro: textos interativos tendem a ser menos informativos e vice-versa.

A AMD se caracteriza como uma 'abordagem metodológica baseada em corpus destinada a (i) identificar os padrões de coocorrência salientes da linguagem (...) e (ii) comparar registros no espaço linguístico definido por tais padrões.'[2] (BIBER, DAVIES, JONES *et al*., 2006 , p. 5). Como tal, busca revelar as dimensões de variação entre os *registros* de uma língua. Registro é o termo usado na AMD para se referir a 'uma variedade linguística definida por

[1] 'Register is used here as a cover term for any language variety defined by its situational characteristics, including the speaker's purpose, the relationship between speaker and hearer, and the production circumstances.'

[2] 'a corpus-based methodological approach to, (i) identify the salient linguistic co-occurrence patterns in a language, in empirical/quantitative terms, and (ii) compare registers in the linguistic space defined by those co-occurrence patterns.'

aspectos situacionais, incluindo o propósito do falante, a relação entre falante e ouvinte, e o contexto de produção' (BIBER, 2009 , p. 823). O termo possui considerável amplitude (BIBER, 1994 , p. 32), podendo especificar tanto gêneros específicos, como cartas de instituições de caridade (ANTHONY e GLADKOV, 2007) e artigos acadêmicos de bioquímica (KANOKSILAPATHAN, 2007), quanto variedades mais gerais, como 'conversação' (BIBER, 2004), 'documentos oficiais' e 'humor' (BIBER, 1988).

Todos sabemos que há incontáveis textos em circulação em uma língua como o português do Brasil. Por sua vez, diversas teorias sustentam que os textos não variam livremente, mas na verdade relacionam-se estreitamente ao contexto cultural, situacional, de produção e de recepção, além de compartilharem recursos lexicogramaticais, e propõem conceitos como gênero e registro para explicar essa variação (BRONCKART, 1985; BAKHTIN, 1986; SWALES, 1990; BHATIA, 1993; EGGINS, 1994; FERGUSON, 1994; EGGINS e MARTIN, 1997; BRONCKART, 1999; MAINGUENEAU, 2002; BHATIA, 2004; HALLIDAY e MATTHIESSEN, 2004; MACHADO, 2005; MAINGUENEAU, 2005; MEURER, BONINI e MOTTA-ROTH, 2005; MARTIN e ROSE, 2008). Muitas pesquisas em áreas como Análise de Discurso, Análise de Gênero, Linguística Sistêmico-Funcional, Interacionismo Sócio-Discursivo, Estilística e Linguística de Corpus, entre outras, enfocam um ou mais gêneros, tipos textuais, registros e estilos (BIBER e CONRAD, 2009) e mostram tanto as semelhanças e diferenças entre eles quanto sua constituição e organização interna.

Tais teorias e estudos empíricos são fundamentais para entendermos questões importantes da constituição da língua e do discurso. A AMD é uma metodologia que propicia um olhar *em larga escala* sobre essas questões, na medida em que enfoca muitos textos de vários registros ou gêneros ao mesmo tempo.

Assim, dado que há uma profusão de registros e textos na sociedade, surgem questões chave da variação em larga escala, tais como: (1) quais são os parâmetros de variação subjacentes aos muitos registros conhecidos, ou, em outras palavras, como podemos chegar a uma síntese dos elementos centrais dessa variação? (2) qual a variação entre textos de uma mesma variedade textual (por exemplo, entre uma dezena de dissertações de mestrado, ou entre dissertações de áreas distintas, como engenharia e letras)? (3) quais aspectos lexicogramaticais (voz passiva, expansão do grupo nominal, metáfora gramatical, etc.) distinguem registros próximos como comunicação oral e artigo científico, ou reportagem e notícia? Tais perguntas podem ser respondidas por meio da AMD.

Pesquisas anteriores promoveram a identificação de dimensões de diversas línguas, como o inglês (BIBER, 1988; LEE, 1999; DE MÖNNINK, BROM e OOSTDIJK, 2003; CROSSLEY e LOUWERSE, 2007), o coreano (KIM e BIBER, 1994), o somali (BIBER e HARED, 1994), o nukulaelae tuvalan (BESNIER, 1988), o gaélico (LAMB,

2008) e o espanhol (BIBER, DAVIES, JONES *et al*., 2006; PARODI, 2007); contudo, a variação dimensional da língua portuguesa ainda não foi realizada.

Poucos são os estudos de AMD já realizados no Brasil; até onde pudemos determinar, são os seguintes: Oliveira (1997), que investigou a variação entre composições de alunos de inglês e de falantes nativos; Shimazumi (1998), que também enfocou a escrita de estudantes de inglês como língua estrangeira, porém sob a perspectiva da Linguística Sistêmico-Funcional; Silveira (1997), que pesquisou a linguagem de negócios; Conde (2002), que comparou a escrita de alunos de uma escola bilíngue de inglês com a de alunos advindos de institutos de idioma; Shergue (2003), que contrastou comunicações em congresso e os artigos acadêmicos da área médica; Kauffmann (2005), que mapeou a variação na escrita jornalística de um jornal brasileiro; Oliveira (2007) e Oliveira et al. (2009), que coletaram um corpus voltado à AMD; e Bértoli-Dutra (2010), que descreveu a variação entre letras de música popular anglo-americana.

Procedimentos metodológicos da Análise Multidimensional

Faz-se necessário explicitar como a Análise Multidimensional é levada a cabo. Para ilustrar, tomaremos como base a extração das dimensões da língua inglesa realizada por Biber (1988).

Primeiramente, foi selecionado um *corpus* de textos, disponível na época, que representasse a variedade de registros encontrada no inglês. Os *corpora* escolhidos foram o LOB, de textos escritos em inglês britânico e o London-Lund, de transcrições de eventos falados, também da variedade britânica. Foram retiradas porções desses *corpora* e adicionados outros dois registros (variedades de cartas) e obteve-se um total de 481 textos, somando 960 mil palavras.

Em segundo lugar, foi feito um levantamento das principais variáveis que, segundo a literatura existente na época, seriam relevantes para a descrição da língua inglesa. Foram elencadas 67 variáveis, de cunho lexical e estrutural. Os 481 textos foram etiquetados com essas variáveis por meio de um etiquetador especificamente desenvolvido para o estudo (conhecido por Biber Tagger). Parte da etiquetagem foi feita manualmente.

Em terceiro lugar, partiu-se para a Análise Fatorial, a qual identificou sete fatores como sendo a melhor solução. Fez-se então o mapeamento de quais textos estavam presentes em cada fator. Os fatores foram inspecionados um por um e decidiu-se eliminar o sétimo fator porque era composto de variáveis cujo peso era maior em outros fatores.

Em quarto lugar, fez-se, então, a computação dos escores de cada texto em cada dimensão. Os escores consistiam de somas relativas às quantidades das variáveis existentes em cada fator.

Para exemplificar o método de cálculo, tomemos o fator 2. Este fator inclui como variáveis de peso positivo as seguintes características: verbos no tempo passado, verbos no aspecto perfeito, pronomes pessoais de terceira pessoa, verbos 'públicos' (concordar/agree, reclamar/complain, negar/deny, etc.), orações reduzidas, e negações sintéticas (formadas por 'no', 'neither' ou 'nor'). Supondo-se que um dos textos tenha a seguinte contagem destas características: 113 verbos no tempo passado, 124 verbos no aspecto perfeito, 30 pronomes pessoais de terceira pessoa, 14 verbos 'públicos', 5 orações reduzidas, e 3 negações sintéticas, seu escore no fator 2 seria 289, isto é, a soma de 113 + 124 + 30 + 14 + 5 + 3. Na verdade, a computação dos escores não foi feita por meio das contagens brutas, mas sim através de contagens padronizadas com base na média e desvio padrão, a fim de se evitar que o tamanho diferente dos textos influísse nos escores. Estes valores padronizados podem assumir valores negativos, pois indicam quão acima ou abaixo da média cada valor está. Por isso, os escores dos textos podem ser negativos também.

Desse modo, cada texto possuía um valor que indicava sua participação em cada dimensão. Depois fez-se o cálculo dos escores de dimensão para cada registro, por meio de uma média aritmética. Por exemplo, se houvesse três textos de um registro específico na dimensão 2, e eles tivessem os escores 16, 12 e 11, somar-se-iam os três valores, o que resultaria em 39, e dividir-se-ia este total por 3, o que daria 13. O valor 13 seria então o escore de dimensão deste registro na dimensão 2. É possível haver escores de dimensão negativos. Isto acontece quando há uma maioria de escores negativos de cada texto individual.

Por fim, o conjunto de variáveis linguísticas de cada fator foi interpretado funcionalmente e discursivamente, levando ao estabelecimento das dimensões. Cada dimensão é, na verdade, uma escala em que são dispostos todos os registros incluídos na análise, de acordo com seus escores de dimensão. A escala geralmente compreende dois polos opostos, de tal modo que as dimensões são geralmente descritas como 'polo A *versus* polo B'. Quanto mais distantes estão os registros na escala, mais distintos são. Na terminologia da AMD, emprega-se os termos 'positivo' e 'negativo' para se referir a esses polos, sendo que o polo A recebe o nome de 'positivo' e o B de 'negativo'. Contudo, tal denominação não implica em juízo de valor; ambos polos são igualmente relevantes e complementares. Os termos refletem a análise fatorial, na qual são mostradas variáveis com sinal positivo e sinal negativo. Isso significa que, em um mesmo texto, quando uma variável positiva ocorre, uma negativa tende a não ocorrer ou a ocorrer em menor número. Por exemplo, as variáveis positivas de maior peso do primeiro fator são: verbos 'particulares' ('private verbs', e.g. doubt, forget, guess), apagamento de 'that' e contrações. E as principais negativas são: substantivos, palavras longas e propo-

sições. Desse modo, nos textos em que ocorram verbos 'particulares', e apagamento de 'that', há uma tendência de aparecimento também de contrações. Por outro lado, nos textos em que existem verbos 'particulares', apagamento de 'that' e contrações, há uma tendência de escassez ou ausência de substantivos, palavras longas e proposições. Em alguns casos, quando a análise fatorial mostra não haver variáveis negativas, a dimensão é formada por um polo apenas (como a dimensão 6 de Biber 1988).

Assim, ainda em relação ao fator 1, decidiu-se que as variáveis com peso positivo tinham como parâmetro subjacente o que se convencionou chamar de 'produção interativa'. Já o conjunto de características com peso negativo revelavam um parâmetro que se chamou de 'produção informacional'. Por isso, o rótulo adotado para a dimensão 1 foi 'produção interativa *versus* produção informacional'. Na Fig. 1 aparece a escala referente à dimensão 1.

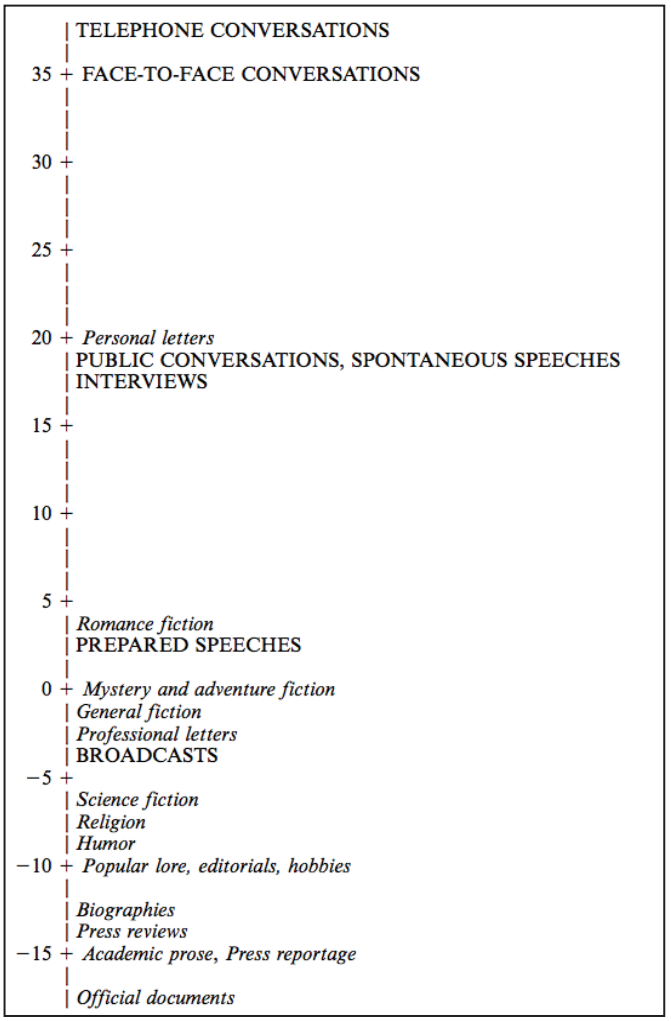

Fig. 1: Dimensão 1 de Biber (1988 , p. 128; 2009 , p. 833): Produção interativa versus informacional (ou Interação versus Informatividade). A parte superior da escala indica o polo 'interativo' e o inferior o 'informacional'. Os números representam o escore de dimensão de cada registro. Os registros grafados em letras maiúsculas são falados, enquanto os em letra minúscula são escritos.

Como se percebe na figura, os registros mais fortemente associados com a interação (na parte superior da escala), no inglês, são as conversas, sejam elas por telefone ou face a face; já os registros mais diretamente ligados à informação (na parte inferior) são documentos oficiais (relatórios, etc.) e a escrita acadêmica e a jornalística.

O mesmo procedimento foi levado a cabo em relação aos outros fatores. Os registros foram então dispostos numa escala, de acordo com seu escore de dimensão. A nomeação dos fatores resultou na identificação de seis dimensões, que são:

1. produção interativa versus produção informacional;[3]
2. preocupações narrativas versus não-narrativas;
3. referências explícitas versus referências dependente do contexto;
4. expressão explícita de persuasão versus não-explícita;
5. informação abstrata versus não-abstrata;
6. elaboração informacional 'on-line'.

Os registros mais típicos de cada dimensão são:

- Dimensão 1 - produção interativa versus produção informacional: os registros que melhor representam o modo de produção com interação são as conversas, tanto ao telefone quanto face a face; os registros que melhor representam a produção informacional são documentos oficiais, reportagem jornalística e prosa acadêmica.
- Dimensão 2 - preocupações narrativas versus não-narrativas: os registros que melhor demonstram uma preocupação com a narração são os registros de ficção, enquanto que os que melhor exprimem uma orientação não narrativa são os registros de rádio e TV, passatempos e documentos oficiais.
- Dimensão 3 - referências explícitas versus referências dependente do contexto: os registros que apresentam referência explícita em maior grau são documentos oficiais, cartas profissionais, resenhas jornalísticas e prosa acadêmica. Já os registros de rádio e TV, conversas telefônicas e cara a cara e ficção romântica exprimem referência dependente da situação.
- Dimensão 4 - expressão explícita de persuasão versus não-explícita: os registros de caráter mais persuasivo são as cartas profissionais, os editorias e a ficção romântica. Por outro lado, os registros nos quais a persuasão é menos explícita são os de rádio e TV, resenhas jornalísticas e ficção de aventura.

[3] Os nomes das dimensões podem ser sintetizados, de tal forma que a primeira dimensão poderia ser chamada de 'Interação versus Informatividade'.

- Dimensão 5 - informação abstrata versus não-abstrata: os registros que veiculam informação mais abstrata são os acadêmicos, os documentos oficiais e os religiosos. Já as conversas telefônicas, face a face e ficção romântica apresentam informação menos abstrata.
- Dimensão 6 - elaboração informacional on-line: os registros nos quais a elaboração da informação é mais imediata são palestras preparadas, entrevistas e palestras espontâneas, enquanto que os registros nos quais a informação é elaborada de antemão são os de ficção (mistério, aventura, científica e geral.)

As dimensões mostram uma inter-relação entre registros escritos e falados. Alguns registros escritos possuem características em comum com registros falados e vice-versa. Por exemplo, de acordo com a dimensão 1, cartas pessoais, palestras espontâneas e entrevistas possuem como característica comum o fato de serem produzidas com interação entre escritor ou falante de um lado e leitor ou ouvinte do outro.

Apesar disso, persiste uma diferenciação básica entre os registros falados e escritos na metade das dimensões. Nas dimensões 1, 3 e 5, os registros escritos ocupam majoritariamente um dos polos e os registros falados o outro. O registro que predomina no polo onde se concentram os textos escritos é a escrita acadêmica. Já os registros que se concentram no polo falado destas dimensões são os conversacionais.

Principais estudos com base em Análise Multidimensional

Conforme colocado na introdução, dimensões de variação são padrões de coocorrência de elementos lexicogramaticais que subjazem aos textos de uma língua (BIBER, 2009). Como tal, capturam o espaço de variação dos textos, sintetizam-no e mostram a proximidade ou distância entre os registros investigados. Conforme define Berber Sardinha (2004 , p. 304-305):

> 'Dimensão é o estatuto que um fator assume assim que é interpretado do ponto de vista de sua função comunicativa. Uma dimensão permite visualizar características em comum partilhadas por uma porção significativa dos dados. A interpretação do fator leva em conta tanto as características linguísticas quanto as características partilhadas pelos registros que estão representados no fator. As dimensões permitem redefinir o quadro inicial de registros.'

A metodologia de identificação das dimensões foi introduzida por Biber (1985) e posteriormente refinada por Biber (1988), para a língua inglesa. O termo 'multidimensional' deriva do fato de a análise pressupor a existência de múltiplas dimensões no espaço de variação intertextual.

Diversos estudos já empregaram a AMD, voltados a uma ampla gama de situações. Entre os estudos disponíveis na literatura, encontramos aqueles que se ocupam da variação geral de uma língua inteira (BIBER, 1985; BESNIER, 1988; BIBER, 1988; BIBER e HARED, 1994; KIM e BIBER, 1994; JANG, 1998; LEE, 1999; LAMB, 2002; BIBER, DAVIES, JONES et al., 2006; BIBER e TRACY-VENTURA, 2007; PARODI, 2007), bem como outros que investigam registros específicos, tais como o discurso universitário (BIBER, 2006), a linguagem da música pop (BÉRTOLI DUTRA, 2010), composições de aprendizes de inglês (REPPEN, 1994; SHIMAZUMI, 1998; CONDE, 2002) e textos jornalísticos (KAUFFMANN, 2005), entre outros.

Existem dois tipos básicos de pesquisa em AMD. No primeiro tipo, é realizada a identificação das dimensões, por meio de análise fatorial, que podemos chamar de 'completa' ('full MD study' segundo BIBER, 2009 , p. 844), muito embora os corpora analisados não precisem representar uma língua por completo, podendo ser específicos de um registro apenas. Neste primeiro grupo é que se encaixa a presente proposta. No segundo tipo, não são extraídos fatores, mas são utilizadas dimensões obtidas em pesquisas anteriores e são mapeados os dados sobre essas dimensões; podemos chamar essa modalidade de 'aplicação de dimensões' ('applying dimensions', BIBER, 2009 , p. 844).

O segundo grupo, como foi dito, utiliza-se de dimensões já identificadas (geralmente as que se referem à língua como um todo) e serve para descrever a variação de corpora que não estavam presentes nelas. Como exemplos desse segundo tipo (todos referentes ao inglês e tendo como base as dimensões relatadas por Biber (1988)) temos os estudos de Biber et al. (2002), que investiga registros do contexto universitário, como aulas, orientações e livros didáticos; Biber (1987), que compara registros escritos de inglês americano a semelhantes do inglês britânico; Helt (2001), que confronta registros falados do inglês britânico a seus semelhantes do inglês americano; Conrad (1996), que contrasta artigos de pesquisa, livros didáticos e trabalhos estudantis de duas áreas de conhecimento (biologia e história); Biber e Finegan (1989), que mapeiam as mudanças diacrônicas em diversos registros; Atkinson (1992; 2001), que apresenta as mudanças ao longo do tempo no discurso acadêmico; Connor e Upton (2003), que focalizam variação em cartas comerciais; Quaglio (2009), que verifica a semelhança entre o seriado de TV Friends e a conversação face a face; e Rey (2001), que observa as mudanças nos padrões dos diálogos dos personagens masculinos e femininos da série de TV Star Trek.

Estão disponíveis na literatura quatro análise multidimensionais de registros específicos do português, sendo duas completas e duas que aplicam dimensões existentes. Os dois estudos que efetuaram uma análise com extração de fatores da língua portuguesa são Oliveira (1997) e Kauffmann (2005). Nenhum

desses estudos, contudo, enfocou a variação da língua portuguesa em geral. Oliveira enfocou composições escritas por estudantes, enquanto Kauffmann investigou a variação de registros jornalísticos da Folha de S. Paulo. Os dois estudos que aplicaram dimensões são Santos (2003) e Berber Sardinha (2003). Santos analisou um manual de gestão de negócios, extraiu palavras-chave (SCOTT, 2000; BERBER SARDINHA, 2009) e as mapeou sobre as dimensões do inglês obtidas por Biber (1988). Berber Sardinha (2003) teve como corpus de estudo uma reunião de negócios e também levantou as palavras-chave desse corpus, encaixando-as nas dimensões do inglês previamente extraídas por Biber (1988). Esses dois estudos têm a limitação séria de empregarem dimensões de uma língua (inglês) para caracterizar dados de uma outra (português), o que não é recomendável, visto que língua diferentes geralmente possuem dimensões distintas. Caso houvesse na época uma pesquisa que tivesse extraído as dimensões do português, esses dois estudos poderiam se servir delas e mapear seu corpus nessas dimensões, o que seria mais apropriado.

O arcabouço empregado para a descrição multidimensional do inglês foi aplicado a uma série de outras línguas. Até o presente, foram descritas multidimensionalmente por Biber e outros pesquisadores os seguintes idiomas: inglês (BIBER, 1988; LEE, 1999), nukulaelae tuvalan (BESNIER, 1988), coreano (KIM e BIBER, 1994), somali (BIBER e HARED, 1994), taiwanês (JANG, 1998), gaélico (LAMB, 2008) e espanhol (BIBER, DAVIES, JONES et al., 2006; PARODI, 2007). No que se segue, são apresentados os resultados das análises que trataram de línguas oficiais, portanto o inglês, nukulaelae tuvalan, coreano, somali e espanhol, concentrando-se nas dimensões encontradas e em seus registros mais salientes.

Para a descrição dessas línguas usou-se, em cada estudo, um corpus específico. A quantidade de variáveis, registros e textos também variou consideravelmente. O quadro a seguir resume os elementos centrais de cada corpus empregado.

A primeira língua a ser enfocada pela AMD foi o inglês, conforme dito acima, por Biber (1985 et seq.), que empregou uma mescla de corpora de textos escritos e falados de inglês britânico e americano. As dimensões da língua inglesa reveladas nesse estudo já foram apresentadas e discutidas acima. O inglês também foi foco de outros dois estudos, o de Lee (1999) e o de Crossley e Louwerse (2007). Lee replicou e testou a metodologia proposta por Biber com outro corpus, retirado do British National Corpus, de inglês britânico apenas. Seu estudo mostrou que é preciso ter muito rigor na análise estatística fatorial a fim de garantir a confiabilidade dos resultados. Lee, entretanto, não chegou a interpretar os fatores e propor dimensões. Scott e Louwerse utilizaram um conjunto de corpora britânicos e americanos, alguns com situações simuladas, com o TRAINS corpus, que são diálogos simulados

(*role playing*) entre um informante que se faz o papel de passageiro de trem e outro que se faz passar por funcionário do serviço de informação da estação ferroviária (https://192.5.53.208/research/speech/trains.html). Embora haja textos autênticos (os corpora London Lund, LOB e Brown) na coletânea de corpus empregada, a presença de textos artificiais é passível de crítica, visto que uma das razões de ser da AMD é lidar com a variação existente em um contexto autêntico de uso da língua, e não em dados fabricados para a pesquisa linguística. Mesmo assim, do ponto de vista metodológico, o estudo merece destaque, pois autores inovaram ao empregar variáveis lexicais (mais especificamente, bigramas, que são sequências de duas palavras) como variáveis, em vez de morfológicas, estruturais e sintáticas, que são normalmente usadas em AMD. Os resultados foram promissores, pois sugeriram que a ocorrência de pacotes lexicais frequentes (como os bigramas) seja capaz de distinguir registros. Bértoli-Dutra (2010) também empregou pacotes lexicais na análise multidimensional de músicas pop britânicas e americanas e notou que essas variáveis podem ser úteis na caracterização dos registros. Pretendemos testar o uso de n-gramas (bi, tri ou quadrigramas) na pesquisa a fim de aferir a sua viabilidade para complementar as variáveis lexicogramaticais. Entretanto, devido à presença de textos artificiais, as dimensões encontradas pelos autores ficaram comprometidas, pois não refletem a distribuição de registros autênticos da língua inglesa, não sendo assim comparáveis às encontradas por Biber.

Quadro 1: Dimensões de corpora usados em AMD

Língua	Variáveis	Registros (escritos / falados)	Textos	Textos por registro	Total de palavras
Inglês (1)	67	23 (17/6)	481	6 a 80	960.000
Inglês (2)	39, 58, 63, 65	66 (41/25)	430	2 a 133	2.006.093
Inglês (3)	84	22 (4/18)	Não informado	Não informado	6.287.734
Nukulaelae Tuvaluan	6	7 (2/5)	222	12 a 70	152.771
Coreano	42	22 (12/10)	150	5 a 10	135.500
Somali	58	33 (23/10)	604	3 a 49	600.000
Espanhol (4)	85	19 (8/11)	4049	16 a 791	20.301.847
Espanhol (5)	65	3 (2/1)	90	4 a 74	1.466.744

(1) Biber (1988)

(2) Lee (1999). Seu estudo testou diversos números de variáveis.

(3) Crossley e Louwerse (2007). O artigo relata dois estudos com os mesmos dados, mas com números de variáveis diferentes. Referimo-nos ao estudo 2, considerado mais robusto.

(3) Biber et al. (2006; 2007)

(4) Parodi (2007)

Cronologicamente, a segunda língua a ser descrita por meio da AMD foi o nukulaelae tuvalan, falado em Tuvalu, um arquipélago localizado no Pacífico. O corpus usado para a descrição do nukulaelae tuvalan consistiu de pouco mais de 150 mil palavras, compreendendo 222 textos de sete registros diferentes. As três dimensões extraídas são apresentadas no quadro abaixo, juntamente com os registros mais característicos de cada uma.

Quadro 2: Dimensões da língua nukulaelae tuvalan (BESNIER, 1988)

Dimensão		**Registros mais característicos**	
		Polo positivo	**Polo negativo**
1	Discurso atitudinal *versus* abalizado	Atitudinal: Discursos em contextos particulares e comícios	Abalizado: Sermões escritos e cartas pessoais
2	Referência interpessoal *versus* informacional	Interpessoal: Cartas pessoais e conversações	Informacional: Programas de rádio e TV e sermões escritos
3	Construção textual em grupo *versus* monológica	Em grupo: Conversação, sermões escritos	Monológica: Cartas pessoais e programas de rádio e TV

A terceira língua cuja descrição multidimensional foi publicada é o coreano. O corpus empregado possuía cerca de 135 mil palavras, incluindo 22 registros. As seis dimensões extraídas aparecem no quadro a seguir.

Quadro 3: Dimensões da língua coreana (KIM e BIBER, 1994)

Dimensão		**Registros mais característicos**	
		Polo positivo	**Polo negativo**
1	Interação informal *versus* elaboração explícita	Interação: Conversas particulares e dramaturgia televisiva	Elaboração: Crítica literária, livros didáticos para faculdade
2	Coesão explícita *versus* implícita	Explícita: Contos e conversação	Implícita: Documentos legais e oficiais e notícias de rádio e TV
3	Expressão explícita de posicionamento interpessoal	Mais explícita: Dramaturgia televisiva e conversas particulares	Menos explícita: Reportagem jornalística e documentos legais e oficiais
4	Discurso narrativo *versus* não-narrativo	Narrativo: Dramaturgia televisiva e contos	Não-narrativo: Documentos legais e oficiais e crítica literária
5	Relato presente (‘on-line’) de eventos	Mais presente: Transmissões esportivas e discursos públicos preparados	Menos presente: Reportagem jornalística e notícias de rádio e TV
6	Honorificação	Mais honorífico: Cartas pessoais e conversas em público	Menos honorífico: Documentos legais e oficiais e crítica literária

A quarta língua descrita multidimensionalmente foi o somali. O corpus que serviu de base para a descrição possuía 33 registros, o que somava por volta de 600 mil palavras. À semelhança do coreano, foram extraídas seis dimensões, as quais aparecem no quadro a seguir.

Quadro 4: Dimensões da língua somali (BIBER e HARED, 1994)

Dimensão		**Registros mais característicos**	
		Polo positivo	**Polo negativo**
1	Elaboração estrutural: envolvimento *versus* exposição	Envolvimento: conversações e reuniões de família	Exposição: Introduções de livros e editoriais
2	Elaboração lexical: produção em tempo real (‘on-line’) *versus* planejada	Em tempo real: Transmissões esportivas e palestras universitárias	Planejada: Discursos políticos publicados e editoriais
3	Apresentação argumentativa *versus* relatada de eventos	Argumentativa: reuniões de família e reuniões formais	Relatada: Reportagens jornalísticas e estórias populares
4	Organização discursiva narrativa *versus* não-narrativa	Narrativa: Histórias populares e histórias seriadas	Petições e anúncios
5	Interação distanciada e diretiva	Mais distanciada e diretiva: Cartas pessoais e reuniões de família	Menos distanciada e diretiva: Reportagem jornalística e transmissões esportivas
6	Persuasão pessoal	Mais persuasão: Petições e cartas pessoais	Menos persuasão: reportagem jornalística e transmissões esportivas

A quinta língua cuja variação foi mapeada pela AMD é o espanhol. Há duas pesquisas diferentes referentes a esse idioma, quais sejam Biber et al. (2006; BIBER e TRACY-VENTURA, 2007) e Parodi (2007). A de Biber et al. enfocou uma variedade ampla de registros (19), enquanto a de Parodi trabalhou com um espectro maior de variação (apenas três registros gerais: técnicos, orais e literários). Os resultados são, portanto, diferentes e aparecem nos dois quadros a seguir.

Quadro 5: Dimensões da língua espanhola (BIBER, DAVIES, JONES *et al*., 2006)

Dimensão		**Registros mais característicos**	
		Polo positivo	**Polo negativo**
1	Discurso oral *versus* letrado	Discurso oral: Conversas ao telefone e conversa coloquial face a face	Discurso letrado: Enciclopédia e prosa acadêmica
2	Discurso hipotético (‘irrealis’)	Mais hipotético: entrevistas políticas e debates políticos	Menos hipotético: Enciclopédia e prosa acadêmica
3	Discurso narrativo	Mais narrativo: Ficção e teatro	Menos narrativo: Enciclopédia e prosa acadêmica
4	Interação focada no interlocutor	Mais focada: Teatro e conversas telefônicas de negócios	Menos focada: telejornais e transmissões esportivas de TV
5	Relato informacional	Mais informacional: Enciclopédias e cartas comerciais	Menos informacional: Teatro e debate político
6	Estilo formal	Mais formal: Prosa acadêmica e editoriais	Menos formal: Conversas telefônicas de negócios e transmissões esportivas de TV

No quadro abaixo, apresentamos as dimensões do espanhol obtidas por Parodi (2007).

Quadro 6: Dimensões da língua espanhola (PARODI, 2007)

Dimensão		Registros mais característicos (a)	
		Polo positivo	Polo negativo
1	Foco na contextualização e interação	Foco mais contextual e interativo: registros orais	Foco menos contextual e interativo: registros técnicos
2	Foco na narração	Foco mais narrativo: registros literários	Foco menos narrativo: registros técnicos
3	Foco no comprometimento (‘commitment’)	Mais comprometimento: registros orais	Menos compromentimento: registros técnicos
4	Foco na modalização	Mais modalização: registros orais	Menos modalização: registros técnicos
5	Foco na informação	Mais informação: registros técnicos	Menos informação: registros orais e literários (b)

(a) Como há apenas três registros no corpus, é apontado apenas um deles para cada polo.
(b) Houve empate estatístico entre os escores dos registros nesse polo da dimensão.

Os resultados dos vários estudos resenhados aqui indicam que, embora línguas diferentes possuam dimensões diferentes, há certas dimensões que reaparecem, independente da língua e do tamanho do corpus. Segundo Biber e Conrad (2009 , p. 851), há duas oposições comuns a todas as línguas pesquisadas: a primeira, entre textos com foco informacional versus interativo/interpessoal, e outro entre textos com foco narrativo versus não narrativo.

Comentários finais

A Análise Multidimensional é uma metodologia que tem permitido enfocar a variação textual em corpora eletrônicos por meio de procedimentos estatísticos. Seus resultados fornecem uma visão sintética da variação de textos em corpora, em forma de escala, auxiliando no entendimento das variantes linguísticas estudadas, por meio de suas propriedades comunicativas, funcionais e discursivas.

Abstract

This article presents a particular methodology of Corpus Linguistics (SINCLAIR, 1991; MCENERY e WILSON, 1996; BIBER, CONRAD e REPPEN, 1998; BERBER SARDINHA, 2004; 2005; TEUBERT e KRISHNAMURTHY, 2007; BAKER, 2009; BERBER SARDINHA, 2009; LÜDELING e KYTÖ, 2009), namely Multi-feature, Multi-dimensional Analysis, or simply Multi-dimensional Analysis (AMD, BIBER, 1985 et seq.). MDA enables the study of large

scale variation in electronic corpora, leading to a classification of registers and genres along dimensions, which are patterns of co-occurrence of lexico-grammatical features underlying (oral and written) texts in a particular language or variety. As such, dimensions capture the space of variation among texts and depict the proximity or distance between texts.

Keywords*: Corpus Linguistics, Multi-dimensional Analysis, Variation*

REFERÊNCIAS

ANTHONY, M.; GLADKOV, K. Rhetorical appeals in fundraising. In: BIBER, D. *et al* (Org.). *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam/Philadelphia: John Benjamins, 2007. p. 121-154.

ATKINSON, D. The evolution of medical research writing from 1735 to 1985: The case of the 'Edinburgh Medical Journal'. *Applied Linguistics*, v. 13, p. 337-374, 1992.

_____. Scientific discourse across history: A combined multidimensional/rhetorical analysis of the Philosophical Transactions of the Royal Society of London. In: CONRAD, S.; BIBER, D. (Org.). *Variation in English: Multi-dimensional studies*. Harlow: Longman, 2001. p. 45-65.

BAKER, P. (Org.) *Contemporary Corpus Linguistics*. London: Continuum, 2009.

BAKHTIN, M. M. *Speech genres and other late essays*. Austin, TX: University of Texas Press, 1986.

BERBER SARDINHA, T. Informatividade, interatividade e narratividade na reunião de negócios - Análise Multidimensional e palavras-chave. DIRECT Papers, 52. LAEL, PUC/SP, São Paulo / AELSU, University of Liverpool. Disponível em www2.lael.pucsp.br/direct.

_____. *Linguística de Corpus*. São Paulo: Manole, 2004.

_____ (Org.) *A língua portuguesa no computador*. Campinas / São Paulo: Mercado de Letras / FAPESP, p.295 p.ed. 2005.

_____. *Pesquisa em Lingüística de Corpus com WordSmith Tools*. Campinas: Mercado de Letras, 2009.

BÉRTOLI DUTRA, P. *Linguagem da música popular anglo-americana de 1940-2009.* (Tese de Doutoramento) - LAEL, PUCSP, São Paulo, SP, 2010.

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

BHATIA, V. K. *Analysing genre: language use in professional settings*. London: Longman, 1993.

_____. *Worlds of written discourse - A genre-based view*. London, New York: Continuum, 2004.

BIBER, D. Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics*, v. 23, p. 337-360, 1985.

_____. A comparison of British and American writing. *American Speech*, v. 62, p. 99-119, 1987.

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

_____. An analytical framework for register studies. In: BIBER, D.; FINEGAN, E. (Ed.). *Sociolinguistic perspectives on register*. Oxford: Oxford University Press, 1994. p. 31-56.

_____. Conversation text types: A multi-dimensional analysis. In PURNELLE, G.; FAIRON, C., DISTER, A. (Orgs.), *Le poids des mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*. Louvain: Presses universitaires de Louvain. 2004. p. 15-34.

_____. *University language: A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia: John Benjamins, 2006.

_____. Multi-dimensional approaches. In: LÜDELING, A.; KYTÖ, M. (Org.). *Corpus Linguistics -- An international handbook*. Berlin / New York: Walter de Gruyter, 2009.

BIBER, D.; CONRAD, S. *Register, genre, and style*. Cambridge ; New York: Cambridge University Press, 2009.

BIBER, D. *et al. Corpus Linguistics - Investigating language structure and use*. Cambridge: Cambridge University Press, 1998.

_____. Speaking and writing in the University: A Multi-Dimensional comparison. *TESOL Quarterly*, v. 36, n. 1, p. 9-48, 2002.

_____. Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora*, v. 1, n. 1, p. 1-37, 2006.

BIBER, D.; FINEGAN, E. Drift and evolution of English style: A history of three genres. *Language*, v. 65, p. 487-517, 1989.

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

BIBER, D.; TRACY-VENTURA, N. Dimensions of register variation in Spanish. In: PARODI, G. (Org.). *Working with Spanish Corpora*. London: Continuum, 2007. p. 54-89.

BRONCKART, J. P. *Le Fonctionnement des discours - Un modèle psychologique et un méthode d'analyse*. Neuchatel, Paris: Delachaux & Niestlé, 1985.

_____. *Atividades de linguagem, discursos e textos*. São Paulo: EDUC, 1999.

CONDE, H. M. D. A. *Escolhas léxico-gramaticais em composições de alunos avançados de inglês originários de instituições de ensino bilíngües e monolíngües - um estudo multidimensional baseado em corpus*. (Dissertação de Mestrado) - LAEL, PUCSP, São Paulo, SP, 2002.

CONNOR, U.; UPTON, T. A. Linguistic dimensions of direct mail letters. In: LEISTYNA, P.; MEYER, C. F. (Org.). *Corpus analysis: language structure and language use*. Amsterdam ; New York: Rodopi, 2003. p. 71-86.

CONRAD, S. Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and Education*, v. 8, n. 299-326, 1996.

CROSSLEY, S.; LOUWERSE, M. M. Multi-dimensional register classification using bi-grams. *International Journal of Corpus Linguistics*, v. 12, n. 4, p. 453-478, 2007.

DE MÖNNINK, I. M. *et al.* Using the MF/MD method for automatic text classification. In: GRANGER, S.; PETCH TYSON, S. (Org.). *Extending the scope of corpus based research: new applications new challenges*. Amsterdam: Rodopi, 2003. p. 15-25.

EGGINS, S. *An introduction to Systemic Functional Linguistics*. London: Pinter, 1994.

EGGINS, S.; MARTIN, J. R. Genres and registers of discourse. In: VAN DIJK, T. A. (Org.). *Discourse as structure and process*. London: Sage, 1997. p. 230-256.

FERGUSON, C. Dialect, Register, and Genre: Working Assumptions About Conventionalization. In: BIBER, D.; FINEGAN, E. (Org.). *Sociolinguistic perspectives on register*. Oxford: Oxford University Press, 1994. p. 15-30.

HALLIDAY, M. A. K.; MATTHIESSEN, C. M. I. M. *An Introduction to Functional Grammar*. 3rd. ed. London, New York: Arnold, 2004.

HELT, M. A multi-dimensional comparison of British and American spoken English. In: CONRAD, S.; BIBER, D. (Org.). *Variation in English: Multi-Dimensional Studies*. Harlow: Longman, 2001. p. 171-184.

JANG, S.-C. *Dimensions of spoken and written Taiwanese: A corpus-based register study*. (Tese de doutoramento), University of Hawaii, Honolulu, 1998.

KANOKSILAPATHAN, B. Rhetorical moves in biochemistry research articles. In: BIBER, D. *et al* (Org.). *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam/ Philadelphia: John Benjamins, 2007. p. 73-120.

KAUFFMANN, C. *O corpus do jornal: variação lingüística, gêneros e dimensões da imprensa diária escrita*. (Dissertação de Mestrado) - LAEL, PUCSP, São Paulo, SP, 2005. Disponível em:<http://www.pucsp.br/pos/lael/lael-inf/teses/carlos_kauffmann.zip>.

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

LAMB, W. *Speech and writing in Scottish Gaelic: A study of register variation in an endangered language*. (Tese de doutoramento), University of Edinburgh, Edinburgh, 2002.

_____. *Scottish Gaelic speech and writing : register variation in an endangered language*. Belfast: Cló Ollscoil na Banríona, 2008.

LEE, D. Y. W. *Modelling variation in spoken and written language: the Multi-Dimensional Approach revisited*. (Tese de doutoramento), Department of Linguistics and Modern English Language, Lancaster University, Reino Unido, 1999.

LÜDELING, A.; KYTÖ, M. *Corpus Linguistics -- An international handbook*. Berlin / New York: Walter de Gruyter, 2009.

MACHADO, A. R. A perspectiva interacionista sociodiscursiva de Bronckart. In: MEURER, J. L. *et al* (Org.). *Gêneros - teorias, métodos, debates*. São Paulo: Parábola, 2005. p. 237-259.

MAINGUENEAU, D. Analysis of an academic genre. *Discourse Studies*, v. 4, n. 3, p. 319-342, 2002.

_____. *Análise de textos de comunicação*. São Paulo, SP: Cortez, 2005.

MARTIN, J. R.; ROSE, D. *Genre relations: mapping culture*. London ; Oakville, CT: Equinox Pub., 2008.

MCENERY, T.; WILSON, A. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

MEURER, J. L. *et al.* (Orgs.) *Gêneros - teorias, métodos, debates*. São Paulo: Parábola. 2005.

OLIVEIRA, L. P. *Variação intercultural na escrita: Contrastes multidimensionais em inglês e português*. (Tese de doutoramento), PUC-SP, São Paulo, 1997.

OLIVEIRA, L. P. D. Compilação de um corpus representativo do português do Brasil e análise multidimensional da variação entre gêneros discursivos. *Projeto de Pesquisa*. UERJ, Rio de Janeiro, RJ: PUC-Rio, 2007.

OLIVEIRA, L. P. D. *et al.* Corpobras: um corpus representativo do português do Brasil. Comunicação, VIII ELC (Encontro de Linguística de Corpus). UERJ, Rio de Janeiro, RJ. 2009.

PARODI, G. Variation across registers in Spanish: Exploring the El-Grial PUCV Corpus. In: PARODI, G. (Org.). *Working with spanish corpora*. London: Continuum, 2007. p. 11-53.

QUAGLIO, P. *Television dialogue: The sitcom Friends vs. natural conversation*. Amsterdam: John Benjamins, 2009.

REPPEN, R. *Variation in elementary student language: A multi-dimensional perspective*. (Tese de Doutoramento) - English Department, Northern Arizona University, Flagstaff, 1994.

REY, J. Changing gender roles in popular culture: Dialogue in Star Trek episodes from 1966 to 1993. In: CONRAD, S.; BIBER, D. (Org.). *Variation in English: Multi-Dimensional studies*. Harlow: Longman, 2001. p. 138-156.

SANTOS, V. B. M. P. D. As características léxico-gramaticais de um manual de gestão a partir da análise multidimensional de Biber. *The ESPecialist*, v. 24, n. 2, p. 201-227, 2003.

SCOTT, M. Focusing on the text and its key words. In: BURNARD, L.; MCENERY, T. (Org.). *Rethinking language pedagogy from a corpus perspective - Papers from the third International Conference on Teaching and Language Corpora*. Frankfurt am Main: Peter Lang, 2000. p. 103-122.

SHERGUE, O. *Dimensão de variação no discurso médico-acadêmico: O artigo de pesquisa e a apresentação de trabalhos científicos em congressos*. (Dissertação de Mestrado) - LAEL, PUCSP, São Paulo, SP, 2003.

SHIMAZUMI, M. Investigating EFL writing: A multidimensional analysis. Comunicação apresentada na 6th Braz-TESOL Convention, Recife, PE, 1998.

SILVEIRA, M. S. D. Contrastes interculturais na linguagem de negócios: Análise Multidimensional em L1 e L2. *Projeto de Iniciação Científica.* Rio de Janeiro: PUC-Rio, 1997.

SINCLAIR, J. M. *Corpus, concordance, collocation*. Oxford, New York: Oxford University Press, 1991.

SWALES, J. M. *Genre analysis - English in academic and research settings*. Cambridge: Cambridge University Press, 1990.

TEUBERT, W.; KRISHNAMURTHY, R. (Orgs.) *Corpus Linguistics (Critical concepts in linguistics)*. London: Routledge, 2007.