D.E.L.T.A., Vol. 15, Nº 1, 1999 (209-216)

RESENHA/REVIEW

Biber, Douglas, Susan Conrad, and Randi Reppen. *Corpus linguistics - Investigating language structure and use.* Cambridge: Cambridge University Press, 1998. ISBN 0521496225.

Resenhado por A. P. BERBER SARDINHA *(Pontifícia Universidade Católica de São Paulo)*

PALAVRAS-CHAVE: Linguística do Corpus; Análise Multidimensional; Pesquisa assistida por computador.

KEY WORDS: Corpus Linguistics; Multi-Dimensional Analysis; Computer-assisted research.

## 1. Introdução

O lançamento do mais recente livro de Douglas Biber é recebido com satisfação pela comunidade de linguistas do corpus do mundo inteiro. O trabalho de Biber mais reconhecido é em análise multidimensional de registro, na qual há um interesse pelo estudo de aspectos relacionados à caracterização de variedades textuais. Embora a análise multidimensional não deixe de se situar dentro dos limites da Linguística do Corpus, ela não se definiu como parte da Linguística do Corpus desde seu início. Entretanto, Biber vem se definindo como linguista do corpus há algum tempo, e por isso este volume tem especial significação porque marca formalmente sua passagem para o grupo dos que se intitulam linguistas do corpus. Com a publicação deste volume seu nome será associado a outros pesquisadores mais reconhecidos dentro da Linguística do Corpus, como John Sinclair, Tony McEnery, e Jan Aarts.

A bem da verdade, o crescimento de seu nome como linguista do corpus está fortemente ligado ao desenvolvimento de uma escola de Linguística do Corpus particular. Poderia se situar esta escola do ponto de vista conceitual e geográfico. Conceitualmente, o tipo de Linguística do Corpus desenvolvida por ele tem como aspectos centrais o estudo de características lexicais e gramaticais com forte apelo estatístico. Geograficamente, os seguidores desta corrente estão concentrados primordialmente nos Estados Unidos. Pode-se

arriscar um palpite de que agora em diante vai se falar em uma 'escola norte-americana' de Linguística do Corpus localizada em Flagstaff, Arizona (sede da Universidade onde Biber trabalha), em contrapartida a outras mais antigas, como a britânica ou a escandinava, as quais refletem em maior ou menor grau as tendências emanantes de centros de pesquisa estabelecidos como Birmingham (Grã-Bretanha) e Bergen (Noruega).

As demais autoras, Susan Conrad e Randi Reppen foram alunas de Biber. Elas também participam como co-autoras de outros trabalhos (Biber et al. 1994, 1996, em preparação).

## 2. Conteúdo

A obra se divide em quatro partes principais: investigação do uso de características linguísticas, investigação de características de variedades, resumo e perspectivas, e quadros metodológicos. Há ainda uma introdução que precede a primeira parte.

A Introdução é o primeiro capítulo da obra, e nela os autores expõem os pontos principais que definem a Linguística do Corpus. Primeiramente, haveria dois tipos de investigação possíveis na linguística: o estudo da estrutura, e o estudo da linguagem em uso. Segundo, o foco na utilização da linguagem abre um leque de questões que podem ser investigadas, e este leque exige um tratamento diferente do tradicional (baseado em pequenas amostras e análises de poucos traços). Terceiro, na Linguística do Corpus há uma preocupação com a descoberta de padrões de associação no estudo do uso, e aqui entende-se padrão como um conjunto de traços típicos que co-ocorrem. Finalmente, o ser humano não é dotado da capacidade de perceber o que é típico, pelo contrário, ele é equipado para notar aquilo que se destaca, isto é, o atípico. A abordagem baseada em corpus permite buscar respostas à questão da tipicidade porque faz uso do computador, o qual é naturalmente programado para detectar ocorrências e co-ocorrências.

A introdução toca ainda em outros pontos importantes. Pode-se citar dois. O primeiro ponto seria um resumo das características principais da abordagem baseada em corpus: base empirista, utilização de corpora, emprego de computador, e uso de técnicas quantitativas e qualitativas de interpretação. E o segundo seria a própria definição de corpus: uma coletânea grande e

criteriosa ('principled') de textos de linguagem natural (isto é, não artificiais como linguagem de programação de computador ou matemática).

Ao capítulo introdutório se segue a primeira parte, a qual se intitula 'Investigação do uso de características da linguagem. Esta parte é composta de três capítulos: Lexicografia, gramática, e o estudo de características discursivas.

Nesta parte, a discussão se desenrola em torno da relação entre as áreas tradicionais da investigação linguística e a Linguística do Corpus. Os autores mostram como questões pertinentes à lexicografia, sintaxe, pragmática, e análise do discurso podem ser investigadas por meio de uma abordagem baseada em corpus. Os autores se centram na descrição a partir de unidades abaixo do nível da oração, isto é, o léxico e estruturas gramaticais. A descrição no nível do texto é reservada para a próxima parte do livro.

A segunda parte leva o título de 'Investigação das características de variedades'. Como na primeira parte, esta também é composta de três capítulos. Eles se intitulam, respectivamente, Variação de registro e Inglês para Fins Específicos, Aquisição e desenvolvimento da linguagem, e Investigações históricas e estilísticas.

Nesta parte, os autores apresentam uma discussão a respeito da aplicação da abordagem baseada em corpus para o estudo de variedades específicas de linguagem. Seguindo a nomenclatura adotada por Biber em outros trabalhos, os autores utilizam o termo 'registro' ('register') para se referir a estas variedades específicas. Um registro seria um tipo de linguagem definida por meio de características situacionais, isto é, não-linguísticas (domínio, sexo, formalidade, etc). A abordagem é aplicada basicamente à descrição de corpora de tipos de texto específicos.

A terceira parte é a menos extensa de todas, e se intitula e 'Resumo e prospectos futuros'. Ela comporta apenas um capítulo, chamado de conclusão. Os autores fazem um sumário das idéias principais e fazem sugestões de como os leitores podem se aprofundar em vários temas relativos ao conteúdo da obra.

A quarta parte tem o título de 'Notas metodológicas' (Methodology boxes), e contém dez pequenos textos referentes a questões de cunho prático que permeiam a visão de Linguística do Corpus contemplada no livro. As dez notas são: Questões de desenho de corpus, Questões de desenho de corpus diacrônico, Programas de concordância versus programação para análise de corpus, Características de corpora etiquetados, O processo de etiquetagem, Normalização de contagens, Medidas estatísticas de associação lexical, A unidade de análise em estudos baseados em corpus, Testes de significância e o relato de estatísticas, e Cargas fatoriais e escores dimensionais.

Há um apêndice que inclui corpora disponíveis comercialmente e ferramentas analíticas. A obra contém ainda uma lista única referências e um índice analítico, o qual fecha o volume.

3. Avaliação

'Corpus linguistics' é uma obra indispensável para se conhecer uma parte significativa da Linguística do Corpus atual. A obra cobre com sucesso vários tópicos centrais da área, e apresenta estes tópicos dentro de uma visão particular da Linguística do Corpus, visão esta que está fortemente ligada aos tipos de análise levados a cabo por Douglas Biber ao longo dos anos.

É preciso fazer uma observação em relação ao título. Embora apareça no título, a expressão 'corpus linguistics' figura somente na capa e no prefácio, desaparecendo posterioremente (nem mesmo a título de referência no índice ela é incluída). Ela é substituída por 'corpus-based approach' (abordagem baseada em corpus). As duas expressões não são exatamente sinônimas, visto que 'Linguística do Corpus' pressupõe uma área de investigação ou disciplina mais ou menos delimitada, ao passo que 'abordagem baseada em corpus' é mais abrangente e permite a inclusão mais confortável de várias outras áreas já estabelecidas.

Além disso, outros autores assumem uma diferenciação entre 'corpus-based' e 'corpus-driven linguistics' (Tognini-Bonelli 1993 *apud.* Pearson 1998), o que poderia causar conflito com a posição de Biber: na primeira acepção, o corpus seria usado apenas para provar uma teoria ou posição *a priori*, enquanto na segunda o corpus teria o papel de permitir contraprova a posições iniciais assumidas pelos pesquisadores ou pela comunidade em geral. A posição defendida por Biber seria, segundo esta dicotomia, definida como 'corpus-driven' e não como 'corpus-based'.

De qualquer modo, a dicotomia 'based / driven' não é universalmente aceita, e portanto a substituição de Linguística do Corpus por 'corpus-based approach' não acarreta maiores problemas. De fato, há uma propagação do uso de 'corpus-based approach' (e.g. Ljung 1997). Por esta ótica, a decisão de se ater à abordagem é acertada, visto que os autores defendem a exploração de corpora em várias disciplinas do estudo da linguagem. Mas mesmo assim falta uma justificativa para o desaparecimento brusco do termo Linguística do Corpus.

O leitor não deve esperar obter uma visão ampla e variada das várias correntes analíticas atuantes na Linguística do Corpus. Algumas questões tratadas são pertinentes a qualquer tipo de análise baseada em corpus, seja de que orientação for. Entretanto, a obra se destaca por apresentar uma visão sólida e criteriosa daquilo que caracteriza mais o trabalho de Douglas Biber e da corrente de Linguística do Corpus que ele representa.

Muitos tópicos não são discutidos, mas visto que a abrangência não era o critério norteador dos autores, não chega a ser um problema. Isto é deixado bem claro no prefácio, no qual os autores declaram que o objetivo do volume é relatar aquilo que os autores acham de mais apaixonante na Linguística do Corpus: a investigação empirista da linguagem em uso, e não a apresentação do estado da arte ou de detalhes técnicos de análise. É uma decisão acertada, visto que dado o crescimento vertiginoso da área, torna-se cada vez mais difícil combinar-se abrangência com profundidade. Vide, por exemplo, outras obras, como McEnery e Wilson (1996), a qual consegue ser um pouco mais abrangente mas deixa a desejar em profundidade, e Kennedy (1998), a qual é extremamente abrangente mas pouco detalhada.

Os capítulos estão redigidos de forma clara e organizada. Em sua maioria, a leitura é fácil e assume pouco conhecimento prévio, embora o teor do texto não seja introdutório. A leitura é mais difícil nas partes onde se exige um conhecimento e uma familiaridade com a quantificação. As partes mais herméticas são aquelas em que se aplica a abordagem multidimensional e se faz referência à análise fatorial.

O leitor mais privilegiado será aquele que tiver algum conhecimento da área, ao qual poderíamos nos referir como 'falso iniciante'. O leitor ideal é aquele que já tem algum conhecimento do trabalho prévio de Douglas Biber

na análise multidimensional. Para os mais avançados, ou seja, os que já atuam na Linguística do Corpus, a leitura fica um pouco prejudicada porque muitas das análises podem ser questionadas quanto a seus pressupostos ou sua apresentação (mas não quanto à sua condução).

Em geral, não é aconselhável a leitura linear, capítulo por capítulo. Para o leitor mais iniciante, recomendo o primeiro capítulo (a introdução), a seção 3 do capítulo 4 (comparação do uso de 'begin' e 'start'), e a seção 3 do capítulo 5 (mapeamento da ocorrência de características verbais em artigos acadêmicos), nesta ordem. O leitor mais avançado pode incluir a parte 3 do capítulo 6 (resumo da metodologia da análise multidimensional) e o capítulo 8 inteiro (comparação histórica de várias características textuais).

O capítulo mais questionável é o de número 5, o qual é dedicado ao estudo do discurso. Os autores acertam ao afirmar que a investigação de aspectos discursivos é dificultada devido à problemática da definição da unidade de análise e da inadequação de programas de computador para identificar tais unidades. Mas o capítulo deixa de fazer menção a um crescente número de trabalhos assistidos por computador e baseados em corpora que investigam uma ampla gama de aspectos discursivos, como a identificação de fronteiras internas do texto (conhecida por segmentação discursiva), a localização de tópicos por meio da ocorrência lexical, a análise da ideologia, da imagem, bem como de vários elementos relacionados ao que se convém chamar de uma visão crítica do discurso. Não está se sugerindo aqui que o capítulo contivesse uma análise detalhada destas áreas, mas sim que pelo menos se fizesse menção a elas. O capítulo inclui apenas uma análise de referência pronominal e uma de distribuição de tempos verbais ao longo de seções de artigos acadêmicos.

As notas metodológicas finais são as mais discutíveis. Pode-se inferir que a seleção dos tópicos para inclusão seguiu basicamente dois critérios: o primeiro, de relevância para o entendimento do texto da obra, e o segundo, de aprofundamento de questões práticas que, embora relevantes, eram específicas demais para discussão no corpo do trabalho. Por isso, há um desequilíbrio visível no elenco de temas. Alguns temas são claramente fundamentais e aplicáveis à maioria das abordagens, independente da escola, como por exemplo, questões de desenho de corpus. Outras, porém, são extremamente específicas a uma corrente particular de análise, como 'Cargas

fatoriais e escores dimensionais', a qual se refere ao tipo de análise criada e desenvolvida (brilhantemente, diga-se de passagem) pelo autor principal, Douglas Biber. No conjunto, apesar de problemas de representatividade, a seleção é satisfatória. Deve-se notar que três dos temas incluídos, viz. desenho de corpus, programação para linguistas do corpus, e etiquetagem, foram foco de discussão recente na principal lista de email da disciplina (CORPORA).

O apêndice é útil pois apresenta não somente vários programas e corpora disponíveis mas também os endereços para contato com os responsáveis pelos mesmos. Uma falta gravíssima no apêndice é o fato de ter sido deixado de fora o programa WordSmith Tools (Scott 1997), que é padrão de referência para muitos usuários, principalmente no ambiente operacional Windows.

A bibliografia é adequada para o conteúdo da obra mas não serve como referência para a Linguística do Corpus. É característico do volume o fato de uma das sete páginas da bibliografia ser dedicada às referências a Douglas Biber.

Em conclusão, mesmo com as ressalvas elencadas acima, 'Corpus Linguistics – Investigating Language Structure and Use' é uma obra valiosa e indispensável para linguistas do corpus, estudiosos da linguagem em uso, e curiosos a respeito desta recente e excitante área de investigação, a Linguística do Corpus.

REFERÊNCIAS BIBLIOGRÁFICAS

BIBER, D., S. Conrad and R. Reppen. "Corpus-based approaches to issues in Applied Linguistics." *Applied Linguistics* 15 (1994): 169-223.

BIBER, D., S. Conrad, R. Reppen and S. Rilling. Corpus linguistics and language teaching: Concordancing and beyond. Colóquio apresentado na 30ª Convenção TESOL, 28 de março de 1996, Chicago, Ill, EUA.

BIBER, D., Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *The Longman Grammar of Spoken and Written English*. London: Longman, em preparação.

KENNEDY, G. (1998) *An introduction to Corpus Linguistics*. New York: Longman.

LJUNG, M. (org.) (1997) *Corpus-based Studies in English - Papers from the Seventeenth International Conference on English Language Research Research on Computerized Corpora (ICAME 17)*. Amsterdam/ Atlanta,GA: Rodopi.

McEnery, T. and A. Wilson. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Pearson, J. (1998) *Terms in context*. Studies in Corpus Linguistics, 1. Amsterdam: Benjamins.

Scott, M. (1997) *WordSmith Tools Version 2*. Oxford: Oxford University Press.

Tognini-Bonelli, E. Rationale and aims of Corpus Linguistics. Internal memorandum. Corpus Linguistics Group, Birmingham University, UK.

(Recebido em outubro de 1998; Aceito em dezembro de 1998)