os (sejam dicionários especializados ou comuns de língua). Nossa principal intenção foi tentar mostrar, especialmente para os que iniciam sua pós-graduação, algumas idéias fundamentais postas em textos igualmente fundamentais, apresentadas por dois autores preocupados em cativar pesquisadores, seja pela temática, seja pelas diferentes metodologias de estudo e de pesquisa que nos trazem. Esperamos que, assim como nós já nos empolgamos com os textos originais, nossos colegas também se empolguem com o que deles trazemos agora em português.

Porto Alegre, maio de 2012.

Maria José Bocorny Finatto
Leonardo Zilio
Fabiano Bruno Gonçalves
Organizadores

Sobre o autor e sobre texto
Representatividade em planejamento de corpus

Susana de Azeredo Gonçalves
Fabiano Bruno Gonçalves
Maria José Bocorny Finatto

Douglas Biber é um lingüista norte-americano, atualmente professor do Programa de Linguística Aplicada da *Northern Arizona University*. De sua ampla produção intelectual, podemos destacar, em nosso contexto, os livros *Corpus linguistics: Investigating language structure and use*, de 1998, em coautoria com Susan Conrad e Randi Reppen, e *University language: A corpus-based study of spoken and written registers*, de 2006. O nome desse autor, à semelhança do nome de John Sinclair, é amplamente conhecido e reconhecido entre todos aqueles que se dedicam à Linguística de Corpus. Douglas Biber é sinônimo de afinco de análise de dados de língua também entre aqueles que apenas adotam *corpora* eletrônicos e alguma metodologia de estudo da Linguística de Corpus, sem se filiarem à disciplina como um referencial teórico.

Segundo Biber (1998), os estudos de linguagem podem ser divididos em duas grandes áreas: *estudos de estrutura* e *estudos de uso*. Os *estudos de estrutura* são mais tradicionais (no sentido de serem mais comuns) e buscam identificar unidades estruturais e classes gramaticais. Segundo essa perspectiva, focaliza-se uma determinada característica linguística e investigam-se as formas em que estruturas similares ocorrem em diferentes contextos e como elas servem para diferentes funções. Por exemplo, uma análise estrutural vai descrever as similaridades gramaticais e as diferenças entre, por exemplo, as sentenças *I hope that I can go / I hope to go / I hope I can go.*

Por outro lado, os *estudos de uso* representam uma perspectiva diferente e não menos importante, cuja ênfase é o uso da linguagem. Em vez de tentar julgar a gramaticalidade, os especialistas que estudam o uso dão atenção a padrões típicos da linguagem. Dessa perspectiva, pode-se investigar como os falantes e escritores usam os recursos da linguagem. Uma análise do uso vai além da

descrição gramatical e procura perguntar por que uma língua X teria três estruturas que são tão similares em significado e em função gramatical, como, por exemplo, as sentenças do inglês citadas no final do parágrafo anterior.

Os estudos de estrutura e os estudos de uso são duas perspectivas para se "ver" a linguagem em sua dinamicidade. São maneiras diferentes de se observar um mesmo objeto e, portanto, não excludentes. Isso porque, para observarmos o uso da linguagem , é preciso saber em que estrutura um uso está inserido; e, para observar a estrutura, é preciso ver em que contextos de uso tal estrutura é usada. É nessa perspectiva de estudos de uso que se encontra a Linguística de Corpus, fortemente associada aos estudos da Linguística Aplicada.

A grande maioria dos trabalhos que se deixam nortear pela Linguística de Corpus faz uso de uma vertente sua que é centrada no estudo da frequência e da padronização. Tal vertente faz uso de ferramentas, tais como listas de frequência e de concordâncias (que são listas de contextos de uma dada palavra ou expressão gerados automaticamente a partir de um dado *corpus*), as quais auxiliam a observação da linguagem em uso. Tais trabalhos utilizam como base os estudos desenvolvidos por John Sinclair, nome eminente dos estudos com *corpora* que subsidiaram importantes dicionários da língua inglesa. No entanto, há outra vertente da Linguística de Corpus, a qual não segue, necessariamente, tal proposta, e que tem, justamente, como um de seus representantes Douglas Biber.

Biber apresenta um tipo de Linguística de Corpus que se caracteriza por apresentar um forte componente sociolinguístico, tendo como foco a descrição da padronização e da frequência. Indo mais além, no entanto, Biber destaca um tipo de metodologia conhecida como *Análise Multidimensional* (cuja sigla mais usual é AMD) e que se ocupa de descrever a variação ao longo de diferentes gêneros textuais ou gêneros de discurso ou, como Biber denomina, *registros.* Segundo Berber Sardinha (2004), a Análise Multidimensional "(...) é uma abordagem para análise de *corpus* que usa procedimentos estatísticos (principalmente análise fatorial) visando ao mapeamento das associações entre um conjunto variado de características linguísticas dentro de um *corpus* de estudo." (p. 300)

O *corpus* pode ser tanto uma seleção de textos específicos, quanto um conjunto de gêneros, quanto amostras de uma determinada língua. Para Berber Sardinha (2004), a Análise Multidimensional

> "(...) baseia-se exclusivamente em *corpora*, isto é, pretende descrever um grande número de textos autênticos; é essencialmente computacional, fazendo uso de ferramentas automáticas e semi-automáticas para rotulação das características de interesse nos textos; presta-se à descrição de conjuntos de textos ou registros, em vez de textos individuais; tem um caráter essencialmente comparativo, pois promove o contraste entre os textos ou registros; é multidimensional, ao reconhecer que a variação entre textos e registros pode ser mais adequadamente descrita por meio de múltiplos parâmetros; utiliza aparato quantitativo da descrição, que permite a especificação da concorrência dos traços linguísticos de modo objetivo." (Berber Sardinha, 2004, p. 301 e 302)

Há alguns termos e conceitos que são centrais na Análise Multidimensional. São eles: *traços*, *características*, *registro* ou *gênero*, *tipo de texto*, *fator* e *dimensão*.

Traços são elementos linguísticos essenciais para a análise, os quais podem ser quantificados. São escolhidos por meio de pesquisa na literatura disponível. As *características* dividem-se em características linguísticas e situacionais: as linguísticas são os traços escolhidos para serem quantificados; as situacionais descrevem as características de uso de uma variedade. *Registro* ou *gênero* (sentido idêntico) são uma variedade definida por variáveis situacionais, cujos rótulos são empregados por falantes nativos no dia a dia. São exemplos de registro ou gênero: prosa acadêmica, editoriais jornalísticos, conversação espontânea. *Tipo de texto* é um conjunto de textos formado exclusivamente com base em critérios linguísticos. O tipo de texto é definido em estágios avançados da Análise Multidimensional, quando já foram descritas as dimensões e mapeados os registros. *Fator* é um grupo de variáveis que coocorrem de maneira significativa do ponto de vista estatístico. *Dimensão* é o estatuto que um fator assume assim que é interpretado do ponto de vista de sua função comunicativa. (Berber Sardinha, 2004, p. 303 e 304).

Uma pesquisa que utiliza Análise Multidimensional envolve várias etapas que, segundo Berber Sardinha (2004, p. 306), compreendem "análises macroscópicas e microscópicas". As macroscópicas ocorrem quando há a computação dos fatores, e as microscópicas quando há a interpretação dos fatores de modo funcional.

Dos mais de 150 artigos de Douglas Biber, o texto aqui apresentado é *representativo* e básico para aqueles que forem empreender um estudo que envolva *corpus*, esteja esse ou não inserido do que se chama de Linguística de Corpus, disciplina na qual o artigo é centrado.

Inicialmente, pode-se afirmar que seu ponto central é a construção de um *corpus* que seja representativo em termos de idioma, população, texto e tamanho de amostra. Segundo Biber, necessita-se de uma pesquisa teórica bem embasada antes de se construir um *corpus*, de modo a identificar a problemática de que se queira tratar. A prioridade é haver uma definição completa em relação à população-alvo e às decisões a respeito dos métodos de amostragem. Vemos que a representatividade se refere ao quanto uma amostra inclui toda a gama de variabilidade de uma população, sendo que se entende por "população" o objeto do estudo a ser realizado; sendo essa população a língua como um todo, são feitas propostas de cálculos por meio de fórmulas para que se atinja uma representatividade para o estudo que se vai realizar. Por meio da apresentação de um método estatístico para se chegar à representatividade, baseado no cálculo da distribuição, ou seja, média, desvio padrão e erro da amostra, chega-se à

representatividade usando-se uma fórmula que determina o tamanho mínimo necessário de uma amostra para representar a população de determinada categoria linguística ou de determinado objeto de estudo. Faz-se necessário determinar uma variedade X de tipos de texto para uma dada população; com base nessa determinação, vai-se aproximando do estabelecimento de sua representatividade. O texto seria, em suma, um conjunto de princípios para se atingir a representatividade de um *corpus*.

Embora sua publicação original tenha sido no *Literary and Linguistic Computing*, vol. 8, nº 4, em 1993, o texto que ora apresentamos continua atual e oferece um excelente embasamento em termos de metodologia de pesquisa e *representatividade*. De agora em diante, acreditamos que o texto possa circular mais ainda, visto que até agora, pelo que apuramos, esta é a sua primeira tradução acadêmica para o português do Brasil.

Referências:

BERBER SARDINHA, Tony (2004). *Lingüística de Corpus*. Barueri, SP: Manole.
BIBER, Douglas. (1998) *Corpus Linguistics*: Investigating language Structure and Use.Cambridge, UK: Cambridge University Press. 301 p.

Representatividade em planejamento de *corpus*[1]

Douglas Biber[2]

Tradução de Paula Marcolin
Revisão de Fabiano Bruno Gonçalves
e Susana de Azeredo Gonçalves
Revisão técnica de Maria José B. Finatto

Resumo

O presente artigo aborda uma série de questões relacionadas à obtenção da "representatividade" no planejamento de um *corpus* linguístico; essas questões englobam: a discussão sobre o que significa "representar" uma língua; a definição de população; a estratificação *versus* a amostra proporcional de uma língua; amostras dentro de textos; e questões relacionadas ao tamanho necessário da amostra (número de textos) de um *corpus*. O artigo destaca, dentre diversas maneiras, que características linguísticas podem ser distribuídas dentro de textos e entre textos diferentes; são analisadas as distribuições de diversas características particulares e são discutidas as implicações dessas distribuições para o planejamento do *corpus*.

Este artigo defende que a pesquisa teórica deveria ser anterior ao planejamento do *corpus*, identificando, então, os parâmetros situacionais que variam entre os textos de uma comunidade discursiva e também os tipos de características linguísticas que serão examinadas no *corpus*. Estas considerações teóricas devem ser complementadas por investigações empíricas da variação linguística de um *corpus*-piloto como base para decisões específicas de amostras. A construção efetiva de um *corpus* ocorreria, então, em ciclos: o planejamento original baseado em análises teóricas e de estudo-piloto seguido de uma coleta de textos, por investigações empíricas mais detalhadas da variação linguística, e por uma revisão do planejamento.

1 Original: **Representativeness in Corpus Design.** Literary and Linguistic Computing, Vol. 8, No. 4, 1993, p. 243- 257. Tradução para o Brasil com especial permissão do autor para esta publicação.

2 Departamento de Inglês, Northern Arizona University.