ANDERSON, W.; CORBETT, J. *Exploring English with online corpora: an introduction*. London: Palgrave Macmillan, 2009. 205 p.

Vander Viana*
Queen's University Belfast - Belfast / Irlanda

O público leitor interessado na investigação da língua inglesa tem uma nova publicação à sua disposição: *Exploring English with online corpora: an introduction*. Como o próprio título indica, o volume é destinado àquelas pessoas que não têm conhecimento especializado a respeito do funcionamento da língua inglesa. De forma semelhante, também não se pressupõe que os leitores sejam versados em linguística de *corpus*.

O volume se estrutura em oito capítulos, um apêndice e glossário. Cada capítulo inicia com uma breve introdução, esclarecendo os seus objetivos para o leitor e / ou apresentando perguntas a serem exploradas. No fim, há um pequeno resumo do texto e a indicação de uma lista de leitura adicional, que cita algumas referências-chave sobre o assunto abordado.

O primeiro capítulo apresenta uma introdução aos tópicos mais importantes para a compreensão do livro como um todo. Dessa forma, há uma discussão, por exemplo, sobre os conceitos de *corpus*, representatividade, tipologia e anotação. Posteriormente, são apresentados cinco *corpora on-line* que serão tratados no decorrer do volume.

Mais pontualmente, o Capítulo 2 apresenta conceitos técnicos da linguística de *corpus*, especialmente os que envolvem a compreensão de procedimentos estatísticos. Após uma breve introdução às questões envolvidas em análises qualitativas e quantitativas, os autores se voltam para as noções de média, desvio padrão, frequência bruta e normalizada, informação mútua, palavras-chave, dispersão, análise de variância, entre outros.

É no Capítulo 3 que a investigação da língua inglesa propriamente dita se inicia com o estudo do léxico, já que os autores acreditam ser essa área a mais

* vander.viana@gmail.com

transparente para o novato. Após discorrer sobre o conceito de 'palavra', o capítulo se concentra em como obter informações lexicais a partir da análise de *corpus*, o que é ilustrado com o estudo de '*caravan*' no inglês escocês. A padronização lexical também encontra espaço no volume à medida que são comentadas noções de colocação, coligação, preferência semântica e prosódia semântica. Por fim, são apresentadas formas por meio das quais se têm acesso a dados macrocontextuais, como os de cunho temporal e geográfico.

O conhecimento gramatical é apresentado no Capítulo 4, que se inicia com uma discussão a respeito da natureza desse tipo de conhecimento. Grande parte do texto é dedicada a noções básicas acerca da língua inglesa, como a divisão em itens lexicais e gramaticais, tipos de sintagmas e de verbos, funções sintáticas, entre outras. Os conceitos de delexicalização verbal, coligação e sistema verbal também são discutidos na parte final.

Se os Capítulos 3 e 4 tratam de unidades em nível do período, o quinto capítulo oferece uma investigação discursiva com o uso de *corpora*. Após uma explicação do que pode ser entendido pelo conceito de 'discurso', os autores traçam as diferenças entre a oralidade e a escrita principalmente por meio de marcadores discursivos. A vantagem de se trabalhar com tais itens linguísticos é que eles podem ser facilmente recuperados por meio de uma lista de concordância uma vez que sejam definidas as palavras de busca.

No Capítulo 6, os leitores são convidados a investigar um novo aspecto micro: pronúncia. Aqui são detalhados os conceitos de consoante e vogal (tônica e átona) assim como suas respectivas transcrições de acordo com o Alfabeto Fonético Internacional (ou, em inglês, *International Phonetic Alphabet*, também conhecido pela sigla IPA). Há também a apresentação de conceitos de outra ordem como sotaque, variação e entoação, para citar apenas alguns exemplos.

O Capítulo 7 retoma a preocupação macro, indicando que é possível recuperar informações acerca do contexto de uso em *corpora* disponíveis *on-line*. São apresentados os meios pelos quais essas informações podem ser obtidas para cada uma das linhas de concordância a ser investigada. Assim sendo, pode-se ter acesso, por exemplo, ao sexo, idade, profissão, língua materna, nível de instrução, religião, lugar de residência dos falantes / escritores representados nos *corpora* a depender do recurso utilizado. Pode-se igualmente obter dados a respeito do contexto de produção do discurso oral ou escrito como o tipo de evento, nível de interação entre os participantes e divisão acadêmica.

O Capítulo 8 apresenta conclusões a respeito do uso de *corpora* no ensino e pesquisa em língua inglesa (seja como primeira ou segunda língua, ou língua estrangeira). No primeiro caso, algumas aplicações pedagógicas são apresentadas, considerando a distinção entre ensino focado no produto e no processo. Com relação à atividade de pesquisa, são traçadas algumas diretrizes para o futuro da linguística de *corpus*.

Quanto ao apêndice, trata-se de uma contribuição mais do que bem-vinda ao volume, já que enumera alguns *corpora* disponibilizados na grande rede de computadores. Esses recursos, listados em ordem alfabética, são acompanhados de uma descrição sucinta.

A preocupação com o leitor não-iniciado na área também fica transparente na compilação de um glossário com os principais termos empregados no livro. Esse é de grande auxílio para assegurar a clara compreensão dos tópicos que são discutidos na publicação.

A tarefa de redigir uma obra que possa ser útil a um público-alvo não-experiente não é tão simples quanto possa parecer. De um macro prisma, a organização do livro talvez merecesse uma revisão. Com relação aos tópicos, em alguns momentos, os leitores são apresentados a métodos específicos de análise de um determinado padrão sem que o isso tenha sido apresentado: um exemplo é a explicação do cálculo relativo à medida de colocação quando tal padrão lexical só é explicado posteriormente. Em outras partes, algumas informações aparecem tardiamente, como no caso da distinção entre cotexto e contexto, que é discutida no Capítulo 7 apesar de os termos terem sido anteriormente empregados.

Quanto à redação propriamente dita, o texto parece adotar um tom reducionista em algumas partes. Nas páginas iniciais, por exemplo, lê-se sobre: "a acumulação de grande arquivos eletrônicos, ou *corpora*, de textos escritos e orais, textos armazenados em computadores e manipulados fácil e rapidamente por programas de busca"[1] (p. 1-2). Esse trecho pode dar margem à interpretação de que qualquer coleção de textos disponível em um computador pode ser considerada um *corpus* na acepção que esse termo assume na linguística de *corpus*. Assim, os arquivos de texto armazenados na pasta intitulada 'Meus documentos' no Windows podem ser facilmente considerados como um

[1] Tradução minha do seguinte fragmento: "the accumulation of vast electronic archives, or corpora, of written and spoken texts, texts stored on computers and manipulated easily and quickly by search programs".

corpus pelo leitor inexperiente. Uma definição mais precisa só é oferecida na página 4, quando aspectos como 'seleção' e 'representatividade' são discutidos.

Uma ressalva que pode ser apontada diz respeito a algumas omissões (explícitas ou implícitas). Por exemplo, no Capítulo 2, menciona-se que o trabalho de investigação qualitativo é central à investigação baseada em *corpus*: "a evidência sempre requer interpretação, e as habilidades interpretativas e críticas do pesquisador da área de Humanas são ainda muito valorizadas em discussões acerca do que os dados advindos do corpus *significam*."[2] (p. 22). Por essa explícita menção ao trabalho qualitativo, era de se esperar que o procedimento analítico envolvido fosse delineado no mesmo capítulo, o que não ocorre. Entre as omissões implícitas, por exemplo, está a não-referência ao escore T como uma medida de colocação apesar de a informação mútua ser apresentada no livro.

Outra questão incompreendida relaciona-se à inclusão de certos tópicos em um livro que se dedica à exploração de *corpora on-line*. Nesse sentido, a discussão de dispersão ou palavra-chave pode estar deslocada já que requer o uso de um programa computacional, a ser comprado pelo leitor, e a disponibilização do *corpus* para *download*, o que só ocorre em um dos cinco *corpora* citados no livro.

Da mesma forma, a referência a outros *corpora* não apresentados no início, como ocorre com o *English Language Interview Corpus as a Second-Language Application* (ELISA), pode confundir o leitor não especializado. No entanto, o mais importante a ser ressaltado é a inclusão de recursos que não são *corpora on-line*. Parece que, em alguns momentos, o foco na linguística de *corpus* fica em um plano secundário (como ocorre no capítulo sobre gramática em que os *corpora* aparecem quase como um banco de dados de exemplos) ou inexistente (por exemplo, no capítulo sobre pronúncia, há algumas referências ao *Speech Accent Archive*).

Ainda a respeito dos *corpora* utilizados no livro, nota-se um desequilíbrio nas menções feitas ao *Scottish Corpus of Texts & Speech* (SCOTS), de inglês escocês. Na maior parte das vezes em que a questão da variação nacional é mencionada, o foco recai nessa variedade da língua inglesa. Talvez essa referência fique ainda mais explícita no capítulo a respeito de pronúncia, no

[2] Tradução minha do seguinte fragmento: "evidence always requires interpretation, and the interpretative, critical skills of the humanities researcher are still highly prized in the discussions about what corpus data actually *means*."

qual sete das onze tarefas que envolvem a utilização de *corpora* fazem referência ao SCOTS.

Outra ressalva ao livro se volta para as tarefas, que, em sua grande maioria, não são acompanhadas de respostas, o que seria de grande relevância para o público-alvo da publicação; e nem sempre se relacionam com o uso de *corpora*. Há tarefas que objetivam a conscientização linguística do leitor (Tarefa 4.6 sobre a concordância intuitiva que o leitor faria entre o uso de '*everyone*' e de um pronome pessoal – '*he*', '*she*', *he/she*' ou '*they*'), e a prática mecânica de atividade computacional (Tarefa 5.1 a respeito de como baixar um arquivo do SCOTS) ou de análise linguística (Tarefa 6.1, que lida com a identificação do número de vogais e consoantes em palavras como '*through*' e '*hopeless*'). Outras atividades podem demandar muito do leitor, como o que ocorre com a Tarefa 4.21. Após analisar 15 linhas de concordância para a palavra '*eco-friendly*', sugere-se que o leitor deva investigar, por conta própria, 100 linhas das 8.000 ocorrências de '*baby*' no *British National Corpus* (BNC). Talvez tivesse sido mais interessante inverter as tarefas, apresentando a análise da última, que é mais trabalhosa, e sugerindo a prática com a outra, que é mais concisa.

Ainda em relação às tarefas, pergunta-se se as expressões de busca são realmente as mais adequadas em todos os casos. Na Tarefa 1.3 (p. 11), que objetiva investigar a voz passiva em inglês, as instruções indicam que o levantamento deve ser feito a partir de '*was *ed by*'. As consequências desse uso são várias: todos os sujeitos no plural não serão incluídos nos resultados, nenhum verbo que tenha uma forma irregular no passado aparecerá nas linhas de concordância, e se assume que toda a construção passiva necessariamente inclui um sintagma preposicional por meio do qual o agente é especificado. Nenhuma dessas considerações é discutida na tarefa. Além disso, um dos exemplos dado para a voz passiva ("*the book was proofread twice*") nunca apareceria nos resultados pelo fato de o verbo principal não assumir uma forma regular com a terminação *–ed* e pela ausência da indicação do agente. Pode-se também apontar que a tarefa deixa à margem da análise os casos de voz passiva no qual o verbo auxiliar não corresponde ao '*to be*'. Nesse sentido, todas as construções passivas com, por exemplo, o verbo '*to get*' não seriam consideradas. Finalmente, vale ressaltar que os dois exemplos dados ("*he was followed by a big dog*" e "*the book was proofread twice*") não ocorrem no BNC, *corpus* no qual a tarefa deve ser realizada.

Algumas das questões aqui levantadas poderiam ser resolvidas com uma introdução ao volume como um todo. Nesse texto, poder-se-ia oferecer uma explicação mais clara de como o livro foi organizado, além da apresentação do conteúdo. Da forma como o volume se apresenta, é possível, por exemplo, que o leitor não perceba a existência de um glossário entre as referências bibliográficas e o índice, uma vez que não há referências textuais a ele.

Feitas as ressalvas, cabe dizer que a avaliação do livro deve ser realizada de acordo com a sua natureza. Ele não se configura como um manual de uso de *corpora* como a publicação de Aston e Burnard (1998) em relação ao BNC. Ao mesmo tempo, o volume não pretende explorar as possibilidades de análise linguística que surgem exclusiva ou prioritariamente a partir da investigação de *corpora*, como o faz Sinclair (2003). O foco do livro, como o próprio título indica, é explorar a língua inglesa, não chegando a ser uma abordagem abrangente. Ao contrário, o objetivo é oferecer um panorama geral e inicial dessa língua com referências a alguns *corpora on-line*. É essa abordagem simples e o emprego de uma linguagem clara que faz com que o livro seja acessível a pessoas com pouco ou nenhum conhecimento prévio. Assim sendo, pode ser útil para aqueles que ainda não iniciaram seus estudos ou que estão nos estágios iniciais. A publicação pode, portanto, sensibilizar os leitores iniciantes para a investigação da língua inglesa com o uso de recursos disponíveis na *Internet*.

Referências

ASTON, G.; BURNARD, L. *The BNC handbook*: exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press, 1998.

SINCLAIR, J. *Reading concordances*: an introduction. London: Longman, 2003.

Recebido em 28 de dezembro de 2009. Aprovado em 22 de março de 2010.