RESENHA/REVIEW

BAKER, Paul, HARDIE, Andrew & MCENERY, Tony. 2006. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Resenhado por Vander VIANA
(PUC-Rio)

A área de lingüística de *corpus* conta com uma recente publicação: um glossário escrito por Paul Baker, Andrew Hardie e Tony McEnery. Todos os três autores desenvolvem pesquisas baseadas em *corpus* no Departamento de Lingüística e Língua Inglesa da Universidade de Lancaster. Paul Baker é o autor de *Using Corpora in Discourse Analysis* (Baker 2006) e Tony McEnery, juntamente com Andrew Wilson, assina o livro *Corpus Linguistics* (McEnery e Wilson 1996).

Nas notas introdutórias, os autores esclarecem que em face à constante mudança de endereços de algumas páginas da internet, optam por não incluir os mesmos na obra, deixando ao leitor o trabalho de encontrar tais referências. Ainda segundo eles, são incluídos somente os endereços das páginas que dificilmente serão alterados. Apesar disto, o endereço eletrônico do *British National Corpus*, por exemplo, não consta na entrada referente a este *corpus*. É também nesta parte inicial que são listadas aproximadamente 150 abreviações e acrônimos comuns em lingüística de *corpus*.

A organização de entradas no glossário segue a de um dicionário, ou seja, os termos são listados por meio de seus nomes completos e em ordem alfabética. Tal tipo de disposição é útil caso o leitor deseje consultar um termo específico como '*segmentation*' ou '*semantic prosody*'. No entanto, a mesma não facilita a leitura linear das entradas nem mesmo a identificação de todas as entradas relacionadas a um tópico específico. Neste sentido, teria sido mais produtivo se os autores oferecessem listas de entradas agrupadas tematicamente. Desta forma, os leitores poderiam identificar, por exemplo, todos os etiquetadores ou testes estatísticos que são explicados

no glossário. Tal ausência não invalida ou diminui a importância do glossário, mas deve ser considerada em futuras edições.

As entradas podem ser agrupadas em seis categorias distintas. A primeira diz respeito aos principais conceitos em lingüística de *corpus* como '*keyword*', '*lexical density*', '*token*' e '*type*'. Contudo, também é possível encontrar conceitos ligados a outras áreas. Mais especificamente, há uma entrada que se refere à oposição Chomskyana entre 'competência' e 'performance'. Outros conceitos, que aparentemente não estariam relacionados à lingüística de *corpus*, também podem ser encontrados como '*clitic*', '*compliant*', '*ethics*' e '*function words*'. Nestes casos, no entanto, são comentadas as relações que tais conceitos mantêm com a área em questão.

A segunda categoria se refere aos *corpora* listados na obra. Há referências aos mais utilizados e citados na literatura específica como *American National Corpus*, *British National Corpus*, *International Corpus of English* e *London-Lund Corpus*. É também possível encontrar referências a *corpora* mantidos por grandes editoras como o *Cambridge International Corpus* e o *Longman Learners' Corpus*. A abrangência do glossário é relativamente atualizada com a inclusão do *Corpus del Español*, compilado em 2001/2002 por Mark Davies. O grande diferencial do glossário neste grupo refere-se à inclusão de *corpora* que são provavelmente desconhecidos por grande parte dos pesquisadores da área como um *corpus* de textos da área petrolífera em língua inglesa de origem britânica e americana – *Guangzhou Petroleum English Corpus* (p. 80) e um *corpus* de cartas pessoais produzidas por estudantes japoneses de inglês como língua estrangeira – *Personal Letters Corpus* (p. 130). Apesar da menção a diversos *corpora*, o glossário não faz nenhuma referência à interface criada por Mark Davies para exploração do *British National Corpus* – anteriormente conhecida como *Variation in English Words and Phrases (VIEW)* – e atualmente disponível em http://corpus.byu.edu/bnc. Não há também a referência ao *Corpus* do Português – http://www.corpusdoportugues.org/ – compilado por Mark Davies e Michael J. Ferreira.

Um outro grupo de entradas se refere a programas computacionais empregados em investigações baseadas e/ou dirigidas por *corpus*. Há, portanto, a definição de termos como '*concordancer*', '*parser*' e '*tagger*'. Além disto, faz referência também a conceitos relacionados ('*concordance*', '*part-of-speech tagging*', '*skeleton parsing*'), programas (*ConcApp*, *Concordancer / Le Concordanceur*, *Constraint Grammar Parser of English*, *Link Grammar Parser*, *TAGGIT* e *Trigrams'n'Tags*) e *corpora* etiquetados (*CHRISTINE Corpus* e *Gothenburg Corpus*).

O quarto grupo de entradas se relaciona a testes estatísticos. Explicações sucintas são oferecidas para a diferença existente entre testes paramétricos e não-paramétricos. Há também menção a alguns dos testes mais comumente empregados em lingüística de *corpus* como o qui-quadrado.

São igualmente listados no glossário periódicos como o *International Journal of Corpus Linguistics* e o *Journal of Quantitative Linguistics*. Em relação às associações, o glossário faz referência principalmente às localizadas na Europa como *European Association for Lexicography* e *European Language Resources Association*.

Por fim, são citados projetos, banco de dados e universidades. Há referência, por exemplo, ao *Alex Catalogue of Electronic Texts*, um arquivo de textos livres de direitos autorais, disponível gratuitamente on-line (p. 8-9); e ao *Open Language Archives Community*, uma biblioteca virtual de recursos lingüísticos (p. 125). O único centro de pesquisa com uma entrada no glossário é o *University Centre for Computer Corpus Research on Language (UCREL)*. Neste sentido, parecem faltar entradas dedicadas a outros centros de pesquisa dedicados à investigação na área de lingüística de *corpus*.

As entradas e suas respectivas definições existentes no glossário são, em termos gerais, adequadamente explicadas e exemplificadas. Cada página do glossário contém em média três definições, sendo que algumas poucas chegam a se estender por duas páginas, como no caso de '*collocation*' e '*header*'. Um aspecto útil do livro é a referência cruzada seja na própria definição ou após a mesma. No primeiro caso, a referência é feita pela utilização de negrito nos termos que são definidos no glossário. Desta forma, facilita-se o trabalho do leitor, indicando ao mesmo que termos utilizados na definição também se encontram explicados no livro.

Em relação às referências bibliográficas, há quase 200 trabalhos listados ao final do livro. Freqüentemente nas próprias definições são indicados os textos que devem ser consultados para que se possa aprofundar a compreensão de um determinado conceito.

Algumas poucas definições, no entanto, podem causar dúvidas aos leitores. Ao definir o termo 'probabilidade', os autores afirmam que a lingüística de *corpus* busca verificar o possível e/ou provável na linguagem. Parece que o primeiro adjetivo não deveria ter sido empregado na definição. Como coloca Hunston (2006:240), por exemplo, a análise de corpus

permite determinar a probabilidade de cada escolha lingüística. Em outras palavras, a possibilidade de ocorrência não é o foco da lingüística de *corpus*, que se ocupa da probabilidade de uso de traços lingüísticos. No entanto, ressalta-se que na mesma entrada, os autores afirmam (de forma acertada, mas contraditória com o exposto anteriormente) que com o uso a lingüística de *corpus* pode-se ir além da medida de possibilidade de uma característica lingüística individual para verificar o que é provável no uso real da linguagem.

É fato que toda a obra precisa fazer um corte, uma seleção. Porém, mesmo com aproximadamente 500 entradas, algumas ausências são notadas no glossário. Há no glossário uma entrada para 'empirismo', mas não se explica o que é entendido por 'racionalismo'. Este conceito é apenas mencionado de forma sucinta dentro do primeiro. O glossário também não corresponde às necessidades específicas de leitores brasileiros uma vez que não inclui nenhuma referência a *corpora* de língua portuguesa. A única exceção neste caso concerne a entrada 'Floresta Sinta(c)tica', um projeto conjunto entre portugueses e dinamarqueses.

Apesar das pequenas inadequações e ausências apontadas, o glossário é uma iniciativa importante na área de lingüística de *corpus.* O livro interessa a pesquisadores em qualquer estágio de desenvolvimento. Para os iniciantes, o glossário serve como guia de referência para os diferentes conceitos da área assim como as inúmeras abreviações e acrônimos. Para os mais avançados, o livro traz várias informações a respeito de *corpora* (geralmente especializados) possivelmente desconhecidos da maior parte da comunidade acadêmica. Ressalta-se, por fim, o caráter inovador do glossário, que preenche uma lacuna nesta área do conhecimento.

Recebido em novembro de 2007
Aprovado em março de 2008
E-mail: vander.viana@terra.com.br

Referências

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Hunston, Susan. 2006. Corpus Linguistics. In: Keith Brown. Ed. *Encyclopedia of language and linguistics*. 2nd edition. Oxford: Elsevier.

McEnery, Tony & Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.