BOOK REVIEWS

A non-native user perspective of corpus - based dictionaries of English and French.

by Estela Carvalho

Corpus-based dictionaries are intended to any native or non-native speaker of a particular language. Nonetheless, terminology researchers insist that professional translators need to be aware of the array of resources made available to them by modern dictionaries.

Referring to dictionaries is an essential and time-consuming activity of a translator's daily work. Even though bilingual dictionaries are useful, they cannot take the place of monolingual dictionaries, particularly after the end of the nineties, when the most important English learners' dictionaries started to be reformulated owing to large quantities of text of varied sources in machine-readable form.

Combinations of words in a language happen more frequently than would happen by chance. The so-called collocation is the primary tool offered by corpus-based dictionaries.

As pointed out on its back cover, the Oxford Advanced Learners' Dictionary [OALD] (2002 edition, with Genie CD-Rom) is based on the Oxford Corpus Collection and the British National Corpus. It is filled up with numerous examples based on real context of use.

Sixteen highlighted study pages offer important information to the user on how to consult the dictionary. Number B3, which refers to Collocation, is particularly useful; it makes clear that the examples shown in the dictionary are based on natural occurrences.

The noun *Wine* is worthy of note. To start with, a question is raised: *Can you say "**pink wine**"?* Afterward, the text of the entry of that noun is reproduced. The non-native speaker will therefore be able to realize that he can say *dry / sweet / red / rosé or white wine*. The same sort of explanation is given with reference to which verbs can be used with a specific noun and which adverbs are normally employed with particular adjectives. Important collocations are also printed in bold type, as in the example: *She writes **under a pseudonym***.

The grammar points are particularly useful for a non-native speaker as well. Slight differences in the use of, for example, *can / be able to / could / manage to*, are found in note boxes placed close to one of the entries, and they are also referred to in the mentioned entries.

The Genie CD-Rom can be used either by inserting the CD into the CD-tray or by installing the software on the hard disk. By clicking twice on the corresponding icon shown on the Desktop of a PC,

an image similar to a palmtop is shown. One should only have to type a word on the dialog box and then type < enter> in order to know its meaning. When using a word processor or the Internet, it is only necessary to point the word with the cursor of the mouse and the dictionary entry will appear on the Genie screen.

Grammar points are also featured on the Genie screen, a useful feature for production purposes.

Conversely, if the word in question is mistyped, the Genie will not display any possible option. One must first know the correct spelling of a word to look it up on the Genie CD-Rom.

The study pages are not illustrated on the CD-Rom, neither the illustrations. For example, considering the hypothesis of one trying to comprehend the difference between pie / pudding / pancake, and if the entries on the screen are not enough for someone to deduce it, the printed version of the dictionary should be referred to in order to find it out. Therefore, when traveling or working outside the office, a translator using his laptop will also have to carry the printed version of the dictionary.

The New Cambridge Advanced Learner's Dictionary [CALD] CD-ROM is much superior to Oxford's Genie CD-ROM.

To begin with, it is very understandable: one should just have to start typing a word to display automatically a list of similar entries on the left side of the window screen. Different information comes together with the description of the meaning: spelling, word building, and smart thesaurus.

Word Building is displayed through a pop-up window that reveals related words formed with the same root, that is, nouns, adjectives, verbs, and adverbs.

The Smart Thesaurus illustrates the words related to a particular entry. All information can be copied or printed. The option "copy all entries" makes possible to paste the content of all related words to a new Word document, including their meaning and examples.

In the Pictures section, there is an onomasiologic term-bank, varying from animals, to car & trucks, kitchen equipment and cooking. This represents an enormous economy of luggage if working outside the office is needed. A portable computer and the CALD's CD-Rom would be as much as necessary.

The 2004 edition of the *Nouveau Petit Robert* [PR] is illustrated by examples and citations derived from a corpus formed by literature, scientific texts, extracts from films and music.

It is the first time it benefits from information technology techniques of corpus compilation and research. A commentary found in one of the fourteen pages of introduction justifies that as press goes faster than literature, many citations are taken from newspapers, and not only from literature, especially in what concerns neologisms.

The editors present a new way of explaining how a lexicographer can use his intuition or corpus information to form examples to illustrate each entry: the circumstances in which a lexicographer produces examples are as natural (or as artificial) as those of a writer while writing fiction texts.

Five pages describe, in alphabetical order, the names of the authors quoted in the citations. Furthermore, forty-seven periodicals are listed as contributing to the compilation of the corpus, from the legendary *Le Monde* to *Cosmopolitan* and *Le Québec tel quel*. Moreover, 28 films, from *Les Dames du Bois de Boulogne* to *Et Dieu Créa la Femme* are quoted, and the scriptwriters' names are also indicated.

Most entries begin with their first year of appearance [i.e., *cité* (1080), *baladeur* (1985), *logiciel* (1970), *souris* (1983), cédérom / cd-rom (1989), courriel (1990)][1].

Finally, some entries are illustrated either by citations or by invented examples. When it refers to a citation, the author or periodical is shown just after the sentence: "*L'oursin, dont la bouche s'appelle, on ne sait pourquoi, lanterne d'Aristote, creuse le granit.*" (Hugo). "*Les raves, ces nuits secrètes [...] dans des vieux hangars désaffectés.*" (Le Nouvel Observateur, 1995).

To a certain extent, the PR is a huge investment for a Brazilian translator. Additionally, if a CD-Rom with the same contents is required, one should have to disburse the double amount.

Hence, if working outside the office is needed, the 2949-page book should be also taken. Conversely, the OALD or the CALD are worth half that sum, with a CD-Rom included.

Note

1. Respectively, city / walkman / software / mouse / CD-Rom / e-mail.

Lingüística de Corpus

Por Iria Sponholz

Tony Berber Sardinha, um dos maiores pesquisadores da lingüística de corpus no Brasil, atuante na PUC/SP como professor associado do Departamento de Lingüística e do Programa de Estudos Pós-graduados em Lingüística Aplicada e Estudos da Linguagem, lança, em homenagem ao aniversário dos 40 anos do corpus Brown (1964), o livro *Lingüística de Corpus*. A produção dessa obra é um excelente passo nesta área de pesquisa que está ganhando cada vez mais espaço no Brasil. Trata-se de uma obra de iniciação da área e, neste sentido, cumpre de forma adequada com a sua proposta. O autor esclarece no prefácio que alguns artigos que constituem a obra já foram publicados em revistas, porém, com a devida autorização das editoras, conseguiu reunir suas pesquisas em um só lugar.

Sardinha apresenta, então, ao seu público endereçado, os universitários, essa obra preparada de forma didática, que engloba aspectos teóricos e técnicos em

409 páginas, publicada em 2004 pela editora Manole, de Barueri – SP. A obra é composta de 11 capítulos, assim distribuídos: Visão geral da Lingüística de Corpus; Coleta, armazenamento e pré-processamento de corpora; Listas de palavras, concordâncias, palavras-chave: o programa WordSmith Tolls; Etiquetagem morfossintática (POS tagging); Exemplo de desenho de corpus: um corpus de português especializado na linguagem profissional; Freqüência de palavras da língua portuguesa segundo o Banco Português; Análise de palavras em contexto: concordâncias; Padronização na língua portuguesa segundo a Lingüística de Corpus: a partícula só; Lingüística de Corpus e tradução: prosódia semântica; Lingüística de Corpus e Lingüística Aplicada: o ensino de língua estrangeira e Estudo da variação com Lingüística de Corpus: a Análise Multidimensional. Além disso, é importante ressaltar que, no final do volume, o autor oferece ao leitor quatro anexos com informações muito úteis. No primeiro, constam as três mil palavras mais freqüentes do Banco de Português; no segundo, estão disponíveis endereços de Corpora na Internet; no terceiro, são citados endereços de programas e *sites* relacionados, os quais já haviam sido mencionados no decorrer do livro; por fim, no quarto, estão algumas amostras de concordâncias do Banco Português no qual o pesquisador tem uma idéia do que é possível ser explorar em um corpus.

O livro proporciona uma visão geral dos principais marcos da Lingüística de Corpus antes de a tecnologia a revolucionar como também a re-emergência desse campo de pesquisa a partir da década de 80. Sardinha discute as principais posições filosóficas na lingüística: o empirismo e o racionalismo, citando alguns dos seus principais representantes, e postula que a lingüística chomskiana, que predominou nas academias norte americanas, postergou o desenvolvimento da lingüística de corpus naquele país. Contudo, estudos estão sendo realizados, cada vez em maior número, e o pesquisador de maior destaque é Douglas Biber que se dedica à Análise Multidimensional, fundamentada na sociolingüística e baseada em corpus. Na Europa, do contrário, se encontram os grandes centros de pesquisa, sendo John Sinclair o lingüista britânico que mais vem influenciando pesquisadores de outros países. Além disso, esses centros recebem também a atenção da iniciativa privada que tem interesse em que sejam produzidos resultados possíveis de serem aplicados, por exemplo, em sistemas de reconhecimento de voz, processamento de textos e tradução.

A discussão sobre algumas definições – como a de corpus – foram aprofundadas pelo autor e, no desenvolvimento da argumentação, ele mostra os pontos frágeis, passíveis de críticas dessas concepções, e apresenta, no final da construção dos argumentos, a definição de corpus que ele considera ser a mais completa. Sardinha também questiona a representatividade que se atribui a um

corpus, advertindo que não se deveria fazer generalizações sobre os resultados obtidos a população inteira, já que a representatividade é de uma amostra. O autor afirma que uma das grandes dificuldades de tratar dessa área de pesquisa ainda se deve ao fato de os pesquisadores não terem focado a questão central, isto é, "discutir o porquê da linguagem ser usada de tal modo que ela exiba os padrões e fenômenos" e por isso vem sendo acusada de simplesmente estar fazendo um "tipo de contabilidade lingüística", ou seja, apenas registrando ocorrências lexicais e estruturais.

Outra polêmica abordada na obra é sobre o status da lingüística de corpus - se é disciplina ou metodologia. No decorrer da argüição, o autor propõe que ela não é nenhuma das duas, mas sim, uma abordagem por ser mais do que um instrumento metodológico. Essa declaração encontra respaldo na literatura da área, pois pesquisadores, como Hoey e Biber, defendem essa perspectiva. Contudo, ainda não há um consenso na área e a divergência quanto ao status permanece, mas com o crescente número de pesquisas, essa discussão deve ganhar mais espaço e ser definida.

No capítulo três, o livro adquire um caráter de um manual do programa *WordSmith* de autoria de Mike Scott, usado para ajudar na análise de corpus. Além de ser oferecido o endereço eletrônico para fazer o *download* do programa, disponível somente na Internet, dispõe-se também de toda a instrução para operar as ferramentas. Essas instruções são dadas passo a passo, inclusive com ilustrações para ajudar o leitor a compreender como executar o comando que está sendo descrito. Apesar de existirem outros programas, o autor justifica o destaque dado ao programa *WordSmith Tools* por rodar em Windows, ser de fácil operação e ao qual a maioria dos estudantes/professores/pesquisadores tem acesso. Contudo, o programa não consegue fazer a etiquetagem. Sendo assim, Sardinha dedica o capítulo 4 a programas etiquetadores. Existem mais opções que operam em Unix e Linux sendo que as opções para Windows são mais restritas. Além disso, apresenta as etiquetas usadas pelos respectivos programas e disponibiliza os endereços para a obtenção tanto dos programas como do conjunto de etiquetas. Essa parte técnica é útil principalmente para quem está começando a se familiarizar com esses recursos como também para pesquisadores como material de consulta.

Ao apresentar freqüências de palavras da língua portuguesa do Banco de português, um corpus de linguagem geral, o autor mostra a importância desse tipo de estudo para o entendimento da linguagem. Esses dados revelam aspectos que não estariam acessíveis de outra forma, como por exemplo, ao explorar a presença de estrangeirismos no português do Brasil, mostrando que o projeto de lei n. 1.676, de 1999, do deputado Aldo Rebelo, não encontra respaldo nos dados empíricos do corpus, pois, depois de o autor exemplificar

várias análises, chega à conclusão que as palavras exploradas, como li*ght, payoffs, delivery, sale, valet* e outras, possuem freqüência baixa.

Em suma, o livro é escrito de forma acessível, aborda vários aspectos da pesquisa em lingüística de corpus e apresenta alguns trabalhos e resultados que mostram o que é possível descobrir ao se analisar, com os recursos descritos, um corpus.