

Corpus Linguistics: Investigating Language Structure and Use. Douglas Biber, Susan Conrad, & Randi Reppen. Cambridge: Cambridge University Press, 1998. Pp. 300. ISBN 0-521-4995-7

Written with enthusiasm in a user-friendly style, *Corpus linguistics: Investigating language structure and use* has an attractive no-nonsense approach for its readers. The intended readership is broad, from undergraduate through to professional researcher, and Biber, Conrad and Reppen provide enough depth and breadth to keep different readers' attention, if not always their fascination.

As an introductory overview to corpus linguistics, the authors focus on explaining key methodological and technical issues in corpus linguistics within a research framework of example case studies in Parts I and II (pp. 21-132 and 135-229, respectively). The first part of the book takes up questions concerning individual linguistic features (lexicography, grammar, lexicogrammar and discourse), while the second part explores characteristics of varieties (register variation and ESP, language acquisition and development, as well as historical and stylistic questions). In addition to a closing overview, this 300-page volume also includes a set of short methodology boxes, a list of corpora and analytical tools, plus a concise bibliography.

One of the book's virtues is the consistency of organisation from chapter to chapter in Parts I and II. Beginning with a short two or three page summary of relevant wider research and trends in linguistics, the chapters quickly move to highlighting how corpus linguistics can refresh the parts that other linguistic approaches can't reach. Specific research questions are then raised, with the rest of each chapter explaining and discussing the example case studies. It's a fairly simple rhetorical strategy and it maintains a strong instructional edge.

A second strength of the book is that the authors use different corpora for different purposes and put forward a variety of comparisons between spoken and written English registers. Register is seen as one of the overarching influences on how language is used and structured. For example, the senses of the noun "deal" are contrasted through the noun collocates in two registers from the Longman-Lancaster Corpus, academic prose and fiction. The lessons learned from the corpus analysis are then juxtaposed with how five common dictionaries define the same word. Both are found wanting: The corpus analysis misses the sense of *deal* for card games, and the dictionaries downplay the use of "*good/great deal*" to express amount, which the corpus analysis suggests is the most common sense. Interestingly, in terms of word associations, it is the card-dealing association that speakers apparently tend to associate with the noun *deal*, rather than the more common sense of *amount* found in the corpus analysis (p. 41).

More than that, though, the authors underline how collocational patterning differs across registers and how register-insensitive generalisations about lexical use "are often not accurate for any variety, instead describing a kind of language that doesn't actually exist at all" (p.35). They illustrate this further by tracing noun and verb ratios in academic prose, fiction and speech (Chapter 3, *Grammar*), before usefully comparing what the corpus evidence shows about subject *that*-clauses as a feature of written expository prose with how different ESL textbooks present this grammatical pattern. As with the dictionaries mentioned earlier, the textbooks typically fail to include important functional guidelines that structure's function and use.

As we move through the book to the second part, these discrete insights are further expanded into a more comprehensive framework of language in use through multi-dimensional analysis of register variation. In brief, this involves tagging linguistic features across different corpora and then using factor analysis to identify the dimensions along which different registers may be meaningfully differentiated. In Chapter 6, *Register variation and English for Specific*

Purposes, the authors identify the following factors, or dimensions: involved versus informational production, narrative versus non-narrative discourse, elaborated versus situation-dependent reference, overt expression of argumentation and impersonal versus non-impersonal style. Yet, they do not group ESP as one register in itself; rather, they show how some of the dimensions vary between different academic disciplines. What this kind of corpus linguistic analysis can do is provide highly detailed linguistic evidence of how different discourse communities organise and encode their socioliterate practice.

However, a number of problems can be noted with the corpus linguistic approach. Chapter 7, *Language acquisition and development*, explores how corpus analysis can illuminate school children's development of writing proficiency, the characteristics of children's spoken and written registers, and the relationships between fifth graders' language and various dimensions of adult language. The first problem, as the authors point out, is the lack of publicly available "natural" texts to provide the data for the corpus. They base their analysis on the CHILDES database, which consists of 45,000 words of written text: 14,000 are student writing, and the rest children's readers and textbooks (31,000), with just 17,000 words of spoken text. This corpus of student writing is approximately 1/14th of the 200,000 word sub-corpora that Granger has organised each L2 variety by in the International Corpus of Learner English. Yet, even Granger's sample databases are dwarfed by the millions of words in major commercial databases such as COBUILD. Thus, the need to build up more substantial corpora of learner English across different academic disciplines and age groups is clear; however, it would seem that the constraints on doing this cannot be properly addressed unless educational institutions become more focussed on researching the development of L2 scholastic and academic literacy.

The second problem is that the authors used in-class writings only for inclusion in their sampling of student writing. They argue that this avoids "the confounding influence of teachers modifying student texts" (p.185), implying that this provides for natural language use. Such an argument is often raised in promotion of corpora as if the published texts used in commercially available corpora have never been edited and revised. If we think of the influence of editors and sub-editors in the production of newspapers, books and academic papers, as well as a writer's own constant revisions, it is difficult to accept the claim of "natural" (even if elaborated, careful and informational) language use for most written corpora. And in the specific case of the student writing in the CHILDES database, it is not surprising that the researchers find "many characteristics of on-line production" (p.185) such as frequent use of *And* to start sentences and unclear third-person pronominal reference. So, there remains a question of intrinsic bias in how example texts of English are selected and collected unless better account of different stages of discourse production is taken for both NS and NNS text-producers.

There exists a basic conundrum between corpus linguistics and second language vocabulary research: How lexis is used, and the frequency with which words come up in different registers and corpora, do not necessarily reflect how users organise and retrieve individual words from the mental lexicon, nor do corpora offer insights into how users process and combine lexical items for different discourse purposes. The one constantly contradicts the other. A sense of "Yes, but ..." persists as we look at the huge advances made by corpus linguistics and the wealth of often counter-intuitive insights that *Corpus linguistics: Investigating language structure and use* provides. Where, though, are the hypothesized connections to psycholinguistics? In the end, is it just horses for different courses, or should we expect a closer alignment between the twin imperatives of quantitative data and qualitative modelling?

References

Granger, S. (Ed.) (1998). *Learner English on Computer*. Harlow : Addison Wesley Longman.

Reviewed by Andy Barfield
University of Tsukuba