

Review in ICAME Journal, Volume 40, 2016, DOI: 10.1515/icame-2016-0009

Douglas Biber and **Randi Reppen** (eds.). *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press, 2015. 639 pp. ISBN 978-1-107-03738-0 (hardback). Reviewed by **Adam Smith**, Macquarie University.

This addition to the *Cambridge Handbook* series presents an expansive coverage of the achievements and potential of corpus linguistics as a research approach over a wide range of subdisciplines. In their Introduction, **Biber** and **Reppen** write that it is, unlike other textbooks and handbooks available in the area, “a critical discussion of the ‘state of the art’ rather than an introductory overview of the field in general” (p. 4). It is designed for practising scholars and advanced students, focussing on descriptions of landmark studies in each area, as well as providing an empirical case study in each chapter, to demonstrate the application of corpus techniques.

The book is organized in four parts. The first presents a set of methodological considerations, in the form of an introduction to corpora, a survey of computational tools for corpus compilation and analysis, and a summary of statistical techniques available for the interrogation of corpus data. Part II looks at corpus analysis of linguistic characteristics, from discourse intonation to collocation/phraseology, grammar, discourse analysis and pragmatics. The third section is concerned with the analysis of varieties, of register and genre as well as across dialects, world Englishes and learner language. The final section is a miscellany, covering applications of corpus use as diverse as lexicography, pedagogy and translation.

Davies’s introductory chapter surveys the kinds of research that can be carried out in different classes of corpora, going from the smaller Brown family to larger genre-balanced corpora like the BNC and his own Corpus of Contemporary American English (COCA), to massive text archives like Lexis-Nexis and Google Books, and the Web as a corpus. He offers the intriguing prospect that, with access to data and biodata from social media, and improved processing speeds, researchers may soon be able to “examine the use of a particular word, or phrase, or syntactic construction virtually in real time” (p. 30). **Rayson’s**

chapter on corpus compilation and analysis takes a similarly positive approach to the technological developments that have occurred over corpus linguistics' brief history, charting the evolution of tools for compilation, annotation and retrieval. Rayson sees the need for improved visualization techniques, given the amount of data now available to be analysed, but predicts that methods developed in corpus linguistics will have an impact on other text-based disciplines in the social sciences. **Gries**, on the other hand, sees corpus linguistics lagging behind disciplines such as psychology or sociology when it comes to statistical methods, and is openly critical of the approach of some previous studies. It is hard to argue with his assertion that a knowledge of important concepts such as dispersion, and informed application of statistical techniques, is essential for such a data-driven area of linguistics. However, given the range of academic areas where corpora are used, as attested by this volume, a rather friendlier presentation of the tools available would have assisted Gries' promotion of statistics. For example, the formula for calculating relative entropy is presented early in the chapter, with only a cursory explanation of how it works. Compare this with **Xiao's** careful unpacking of formulae used to calculate statistically significant collocations in a later chapter.

Part II of the handbook is organized as a progression of lexical levels, from prosody, through lexical characteristics (keywords, collocations and phraseology), followed by four chapters on aspects of grammar, then concluding with corpus-based studies of discourse functions and pragmatics. This structure means that the opening chapter, on discourse intonation, addresses an area that is relatively poorly served by existing corpora – although **Cheng's** case study on the use of prominence on pronouns, using the Hong Kong Corpus of Spoken English, demonstrates the rich possibilities that prosodically transcribed corpora can offer. In contrast, the lexically-based chapters (by **Culpeper** and **Demmen**; **Xiao**; **Gray** and **Biber**) can point to large existing fields of research, and great progress in the development of analytical tools, particularly to identify units of meaning. Corpus research has had a profound influence on many linguists' approach to grammar, in reassessing its relationship to the lexis and showing how it can vary across regions and genres, and change over time. The chapters here on descriptive grammar (**Leech**), grammatical variation (**Kolbe-Hanna** and **Szmrecsanyi**), grammatical change (**Hilpert** and **Mair**) and lexical grammar (**Hunston**) give a very thorough overview of this influence, and present a range of corpus approaches, including the use of the *Oxford English dictionary's* quotation database as a means of charting diachronic change. The final three chapters in this section apply corpus methodologies to areas of linguistics that have traditionally been areas of qualitative rather than quantitative study – dis-

course analysis and pragmatics. The authors (**Partington** and **Marchi**; **Clancy** and **O’Keeffe**; and **Taavitsainen**) look at the ways in which corpus linguistics can and has added value to previous approaches.

Part III is concerned with the analysis of varieties, looking at the area of register variation in general (**Conrad**), diachronic change (**Kytö** and **Smitherberg**), as well as particular registers such as spoken (**Staples**), written academic (**Hyland**) and literary (**Mahlberg**). Biber’s multidimensional analysis model is a feature of several of these studies, as a means of differentiating register-specific characteristics. As Conrad notes, corpus-based research is particularly well suited to the study of register variation, with most well-designed corpora providing representative samples of registers. The question of regional variation is then covered by **Grieve** (on dialect variation), **Hundt** (on World Englishes), **Mauranen**, **Carey** and **Ranta** (on English as a lingua franca), and finally the question of learner language as a variety is addressed by **Gilquin** and **Granger**. Again, these studies are well-served by a wide range of existing corpora, from the Freiburg English Dialect Corpus (FRED), to the many varieties of English represented in the International Corpus of English (ICE), and then the ICLE, focussing on learner English. It is a slight disappointment that Hundt’s chapter concentrating, as it does, on ICE – because of the structural comparability of its corpora – does not investigate the possibilities (and drawbacks) to the study of World Englishes presented by the recently published Corpus of Global Web-based English (GloWbE).

Part IV, on “other applications of corpus analysis” presents corpus research into a variety of areas that do not quite fit under a common heading, but are nevertheless central to the development of corpus linguistics. **Martinez** and **Schmitt** look at vocabulary lists such as West’s General Service List, and Coxhead’s Academic Word List, and discuss how newer versions of these lists have applied more refined corpus techniques in an attempt to enhance their usefulness. Corpora have long informed the making of dictionaries, and **Paquot** concentrates on an area where dictionaries still have a way to go – in their coverage of phraseology. Paquot suggests future developments where corpora will be more closely integrated into dictionaries to present patterns of word usage relevant to the user’s interest. Two studies of pedagogical use of corpora follow – classroom applications of corpus analysis (**Cobb** and **Boulton**) and the effect that corpora have on the presentation of grammar in pedagogical materials (**Meunier** and **Reppen**). The final chapter takes us beyond the book’s central focus on English to look at issues of translation such as simplification, normalization and conservatism of language (**Bernadini**).

The most innovative feature of this volume, the inclusion of an empirical case study demonstrating corpus methodology, is addressed quite differently by different authors. A few chapters do not include one, for instance Leech's on descriptive grammar, which restricts itself to an overview to the contributions of corpus linguistics to the description of grammar, and acts as an introduction to the following chapters on grammatical variation, grammatical change and lexical grammar. Those that do, vary in their approaches, with some presenting brand new studies, such as those on forced primings in a discourse analysis of White House briefings by **Partington** and **Marchi**, and another by **Hyland** looking at author identity in academic writing. There are several chapters where previous work is summarized or built on, as in **Culpeper** and **Demmen**'s demonstration of the use of keywords in a semantic domain analysis of *Romeo and Juliet*, and **Kytö** and **Smitterberg**'s discussion of a study of the use of *thou* vs. *you* across different registers in Early Modern English. Some of the empirical studies are not based on the analysis of corpus data, notably the chapters on pedagogical applications of corpora, with one providing a meta-analysis of existing studies on the effect on different modes of learning of the use of corpora (**Cobb** and **Boulton**), another comparing the treatment of the passive between non-corpus and corpus-informed grammars (**Meunier** and **Reppen**). This variety of approaches serves to enhance an appreciation of the range of studies possible within the field of corpus linguistics, although sometimes – particularly in pieces of original research – we get a sense that the author is too constricted by space to provide a full account of their material. This is evident in **Kolbe-Hanna** and **Szmrecsanyi**'s case study on the variation in the use of the complementizer *that* across different dialect areas in Britain, where a lot of the detail of the analysis has to be condensed into an extended footnote. Also, in the spoken discourse chapter, **Staples** gives us a whole range of interesting data on stance features found at different phases of nurse-patient interactions, but no summary conclusions on the findings are presented.

Another element of each chapter is the survey of existing literature, and the state of the art for each field. The promise of the Introduction is that each chapter will present a discussion of “the most important studies” (p. 6) in each area, which is again a challenge within the limited space available. One approach that comes across as being particularly useful, especially to researchers new to a field, is the presentation of summary tables relating researchers/papers to areas of interest within a field. These are provided in several chapters, including **Gray** and **Biber**'s on phraseology, and **Conrad**'s on register variation.

There are a few additions/improvements that could have been considered to make this work a little more user-friendly. One would be a glossary. As previ-

ously noted, there are issues over terminology as fundamental to the field as *register* and *genre*. Add to this various clearly technical statistical terms that may be unfamiliar to even experienced corpus linguists, and potential confusion over the use of phrases such as *corpus-driven/corpus-based* versus *top-down/bottom-up*, and the case for a consolidated glossary is quite strong. Also, a more comprehensive list of corpora than the short summary of “major corpora cited in the handbook” would be a valuable tool. It is not clear what quality allows a corpus to be classified as “major” (size, frequency of use/mention?), but there are many others cited that are the basis for important research, and most users of this handbook would be glad of a ready reference to find those most relevant to them. Finally, while this is a very well-produced book, with the numerous tables and figures usually clearly presented to convey complex information, occasionally some reproductions of search outputs that are designed for a screen interface do not render so clearly in print.

But these are minor quibbles. This volume is dedicated to the memory of Geoffrey Leech, and its comprehensiveness and quality are fitting testament to his innovation and versatility as a linguist. It is a handbook in a very practical sense, demonstrating the many applications of corpus linguistic techniques. This relatively new approach to research has already made huge strides in providing a set of ever more sophisticated tools for enhancing our understanding of language. The *Cambridge handbook of English corpus linguistics* not only charts the important achievements already made, but sets a template and highlights the potential for future developments in the field that will allow a still greater range of work to be done.