Tony McEnery, Richard Xiao and **Yukio Tono**. *Corpus-based language studies: An advanced resource book.* London and New York: Routledge, 2006. 408 pp. ISBN: 978-0-415-28623-7. Reviewed by **Bernard De Clerck**, University of Ghent.

Provided it works and provided you are not an Eskimo, a refrigerator is a great invention. The logic behind this simple (or simplistic) observation might also be applied when reviewing a book: first of all does it prove to be useful for the target audience and secondly, does it ‘work’? This review will revolve around answering these basic questions.

The book itself is part of the Routledge Applied Linguistics Series, whose target audience the series editors identify as “upper undergraduates and postgraduates on language, applied linguistics and communication studies programmes as well as teachers and researchers in professional development and distance-learning programmes” (p. xvi). The actual aim of the book is “to bring readers up to date with the latest developments in corpus-based language studies” by addressing both “how to” and “why” questions. The template that is used to realise this purpose is one that recurs throughout the series as a whole: an introductory part which explains key terms and concepts, an extension part which digs deeper by assessing and commenting on excerpts from selected key articles, and an exploratory section which puts theory into practice in student-oriented case studies and suggestions for further research. In the following paragraphs, I will first of all provide a concise summary of the material that is covered in each of the three parts. Secondly, and perhaps more importantly, I will provide personal comments on the content itself, the way it is conceptualised and its effectiveness in terms of the goals it wants to achieve.

In the first chapter of the introductory section the bare basics of corpus linguistics are covered by answering essential questions such as “what is a corpus?”, “why use a corpus to study language?” and whether corpus linguistics is actually a theory or a methodology. The answers to these questions are both concise and insightful. They also nicely sketch a range of debates that has taken place against the background of these central issues. The authors diplomatically take a stand as well (by favouring corpus-based approaches and treating corpus linguistics as a methodology) though not without pointing out overlap between and justification for both approaches. Next, a number of important key concepts are introduced and discussed in a pedagogically justified order which is very similar to the stages one goes through when building and/or using a corpus and the questions and issues that are raised during the process. First of all, the importance of crucial concepts such as representativeness, sampling and balance are given centre stage and practical instructions are given on how these can be achieved (as far as a corpus can of course be truly representative). Unit A3 provides an overview of kinds of information that can be added to the raw text material, such as mark-up, POS-tagging, pragmatic and stylistic annotation, actual parsing and alignment in the case of multilingual corpora. Students and teachers will definitely welcome the distinctions that are made between these different kinds of annotation and their relative importance in terms of the research questions one is asking. It will help them (and researchers in general) to make the right choices in selecting existing corpora or accurately tagging one's own collected text material. Furthermore, attention is paid to the importance of statistics in corpus linguistics and to the different kinds of possible corpora that can be used. The recurrent pedagogical concern about terminological confusion is also very much reflected in the book's active concern (one that is much appreciated) with defining and differentiating the different kinds of labels, terms and kinds of corpora from one another (e.g. the distinction that is made between parallel corpora, comparable and comparative corpora, development corpora and learner corpora, etc.). Unit A7 provides an overview of some of the major publicly available "off the peg" – mostly English – corpora. Reference is not only made to widely known available corpora such as the BNC and the diachronic Helsinki corpus, but also to little gems such as the SED (Survey of English dialects). While of course not all corpora could be covered, reference is made (on a number of occasions) to the authors' companion website for a more comprehensive survey of well-known and influential corpora for English and other languages. Units A8 and A9 are particularly interesting for people who want to build their own corpus. Advice is given on how to extract usable data from the Web with the right corpus-processing tools (e.g. Grab-a-site, HTTrack, WebGetter, MLCT) together with warnings about copyright issues and how to clear them. In Unit 10, the concluding section to the introductory part and in my view one of the most stimulating chapters of the book, we are presented with an overview of corpus linguistics being used – more or less convincingly – in a number of areas of linguistics, including obvious domains such as lexicographic and lexical studies (with the invaluable import of corpus data in the study of collocations, semantic prosody and preference), grammatical studies, studies on register variation and genre analysis, contrastive, translation and diachronic studies, and studies on language learning and teaching. In addition, reference is made to work being done in the field of semantics, pragmatics, sociolinguistics, discourse analysis and forensic linguistics (with the intriguing case of Derek Bentley found innocent on the basis of linguistic evidence after being wrongfully hanged in 1953). It is also worth mentioning that at the end of this chapter the authors do not shy away from pointing out the limitations in the use of corpora as well.

Section B, as noted above, is basically composed of excerpts from published material, which elaborate on and provide further background to the key concepts provided in Section A and related points of debate. Part 1 “Important and controversial issues” gives further support to the claim earlier made that external (or situational, social or extra-linguistic) criteria rather than internal (or linguistic) criteria should be used in initial corpus design by drawing upon two highly relevant works, namely Biber’s (1993) “Representativeness in corpus design” and Atkins *et al*. (1992) “Corpus design criteria”. These articles also foreground the related importance of stratified sampling both in terms of language production and perception. In addition, the reader can enjoy part of a very lively debate on the controversial issue regarding the role of corpora in linguistic analysis, language teaching and learning in excerpts taken from Henry Widdowson, Michael Stubbs and John Sinclair. As the excerpts point out, their viewpoints were or are in fact not that diametrically opposed as one (especially the authors themselves) expected or suspected them to be.

Units B3 to B6 present and illustrate some of the studies in the different fields of linguistics that have been introduced and illustrated in A10. More specifically, the use of corpora and corpus analysis is illustrated in lexical studies on the basis of excerpts taken from Krishnamurthy and Partington on collocation and semantic prosody respectively, which provide background knowledge for Case Study 1 in Section C. Grammatical studies such as Carter and McCarthy’s account of the English *get*-passives in spoken discourse and Kreyer’s study of genitive and *of*-construction in written English pave the way for Case Study 2 on the syntactic conditions which influence the choice between a *to*-infinitive and a bare infinitive following *help*. On the topic of language variation, studies are presented by Hyland and Kachru, who focus on metadiscourse in different scientific disciplines and definite reference in world Englishes respectively, and by Lehmann, who presents an analysis of subject relatives with a zero relativiser in American and British English. A more challenging and fairly complex study on register and genre variation is presented in Biber’s multifeature/multidimensional (MF/MD) analysis, which is taken up again in the Exploration section as one of the most labour-intensive corpus-based studies. Contrastive and diachronic studies are represented by McEnery, Xiao and Mo’s cross-linguistic study of aspect markers and by Kilpiö who traces the developments in the functions of the verb *be* from Old English to Early Modern English. Mair, Hundt, Leech and Smith in their turn report on shifts in part-of-speech based on the frequencies in the matching LOB and FLOB-reference corpora. Contributions of corpus-based language studies to the field of language learning and teaching are presented in extracts from Gavioli & Aston,Thurstun & Candlin, and Conrad. These studies show the possibilities and limitations of real language data for language learning purposes and make clear that while corpora do not automatically guide us in deciding what should be taught, they can help us to make better-informed decisions and oblige us to motivate those decisions more carefully.

In the last section of the book, Section C "Exploration", McEnery *et al*. offer the reader the chance to carry out corpus-based analysis in case studies which are thematically linked to the A and B sections of the book. Not only do the authors present a step-by-step manual on how to carry out the searches themselves in view of the particular research questions, they also nicely foreground possible pitfalls in analysing results and doing statistics. In this way, the reader is taught the basic steps in operating the Concord and Keyword functions of the corpus-processing tool WordSmith, practical uses of the BNCWeb, as well as MonoConc Pro and ParaConc and the commonly used statistics package SPSS. At the end of each case study, readers are given further tasks to gain first-hand experience in using the tools and techniques just learned to solve language problems.

I greatly appreciated the book's fusion of theory, practice, technical knowledge and background reading. These, to me, are the most important ingredients for stimulating corpus linguistics research and having it carried out in a correct way by the target audience. Even if some of the key issues covered in the book may be common knowledge to the die-hard corpus linguist (who may perhaps be regarded as the Eskimo assessing the qualities of a fridge), they are nevertheless brought to the actual intended audience of the book, in a very "refreshing" manner, introducing them to or reminding them of lively debates which are stimulating both for laymen and experts. In addition, I particularly welcomed the many references for further reading which, at the time of publication, covered many of the most recently available studies and developments in tagging and data gathering.

In this way, this book not only puts corpus linguistics in the limelight as a very interesting way of carrying out linguistically relevant research, it also foregrounds the various disciplines in which it is used and stimulates the reader to think about related issues and to formulate other interesting research questions in the field of lexical studies, grammar, sociolinguistics etc.

On a more general level, the introduction-extension-exploration template is obviously a very practical and fruitful way of introducing and teaching corpus linguistics in the classroom. The introduction can pave the way to the students' own reading and critical evaluation of the – preferably entire – articles in the extension section, whereas the exploratory section allows practical application and provides a stimulus for further experimentation and practice.

I will now address some minor points of criticism. First of all, while the authors stress the importance of representativeness, balance and sampling, it is only at a later stage that they acknowledge that attaining representativeness is not always feasible in practice. Not only do issues of copyright – which are in fact very briefly discussed – limit the possibilities or goals one has in mind, the very nature of the data itself seriously affects the size and diversity of the data one can process. One only needs to imagine the vast amount of spoken data that is produced at this very instance by native and non-native speakers of English to realise how underrepresented spoken data is in actual corpora. While the BNC is presented as a balanced corpus in Unit A2 (p. 17), the authors do not, at that point, address the imbalance between spoken (10%) and written data (90%). It is not until the section on DIY that the authors acknowledge that "[i]t is also important to note that the lower proportion of spoken data in corpora such as the BNC does not mean that spoken language is less important or less widespread than written language. This is simply so because spoken data are more difficult and expensive to capture than written data. Corpus building is of necessity a marriage of perfection and pragmatism" (p. 73). To be honest, pragmatism often gets the upper hand out of sheer necessity, a point which the authors could have made earlier.

Secondly, although of course not all recent developments or recently built corpora can be mentioned – as the authors themselves are the first to admit – I miss references to important projects which are aimed at taming the Web (GlossaNet and WebCorp, for instance) alongside the tools that are mentioned to retrieve web data in the overview section in Unit A7. In addition, while tools are presented to retrieve Web-based data, the authors themselves do not stress the inherent danger in using web data for linguistic purposes, which in view of the target audience might have been a useful reminder. Apart from obvious advantages of web data (its being freely available and constantly updated and fed with new material – not subject to the same delays in the creation of designed corpora), there are obvious disadvantages as well, such as the abundance of errors, made by both native and non-native speakers and the fact that the source of the data cannot always be traced. Additionally, using frequency data from a search engine is much more problematic than corpus-based frequencies, which seriously affects the validity of quantitative statements, the application of statistics and reliability in terms of representativeness and balance. See for example

Brekke (2000), Lawrence and Giles (1998), Meyer *et al.* (2003) and Renouf (2003) for more pros and cons of internet data.

Finally, a brief comment with respect to the case study on swearwords. The aim of this case study is to demonstrate the use of corpora in sociolinguistic studies and language variation by exploring differences in spoken and written registers based on sociolinguistic variables such as gender, age and social class. While the study itself shows a statistically significant difference in the use of swearwords (i.e. their frequency) for many of these parameters, it runs the risk of oversimplification. First of all, the output of an informant/informants is clearly not determined by one sociolinguistic variable at the time, but by the combination of these variables: they are of a certain age, belong to a certain social class, have followed a particular kind of education and are either male or female. In my view therefore, observations about language with respect to one variable can only be made if the others are kept constant. Now, even though the authors do combine some of the parameters, the data is not extensive enough to combine all and achieve statistical significance at the same time. Secondly, one's linguistic output is not only determined by one's own specific sociolinguistic parameters, but it is also influenced by those of the interlocutors. In fact, the analysis of the parameter 'intended audience' in written language showed significant quantificational differences between the use of swearwords for an all male intended audience and the use of swearwords for an all female intended audience (p. 282). The authors, however, do not transpose this finding to the results of the spoken data in which such a parameter is clearly operative as well. Whom one is talking to – male, female, young, old, education level and social class and the presence or absence of social distance – is at least as important as one's own sociolinguistic features, especially when it comes to using swearwords. This is one area where the results gained by corpus-based analysis should be positioned, interpreted and put into the perspective of a wider sociological context if one does not want to underemphasize the importance and complexity of the social dimension.

None of these minor flaws, however, diminishes the intrinsic value of this book in any serious way. It is a very fruitful marriage of theory, practice and up-to-date technical knowledge and a very useful course book which I would definitely consider using in teaching corpus linguistics. While the material covered may not shake the world of experienced corpus linguists (for whom it is not primarily intended in any case), this book is indeed a working refrigerator for anyone who wants to start teaching or doing corpus linguistics.

References

Atkins, Sue, Jeremy Clear and Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7: 1–16.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8: 243–257.

Brekke, Magnar. 2000. From the BNC toward the Cybercorpus: A quantum leap into chaos? In J.M. Kirk (ed.). *Corpora Galore: Analyses and techniques in describing English. Papers from the Nineteenth International Conference on English Language Research on Computerised Corpora* (Language and Computers 30), 227–247. Amsterdam and Atlanta: Rodopi.

Lawrence, Steve and C. Lee Giles. 1998. Searching the World Wide Web. *Science* 280: 98–100.

Meyer, Charles, Roger Grabowski, Hung-Yul Han, Konstantin Mantzouranis and Stephanie Moses. 2003. The World Wide Web as linguistic corpus. In P. Leistyna and C.F. Meyer (eds.). *Corpus analysis. Language structure and language use* (Language and Computers 46), 241–254. Amsterdam and New York: Rodopi.

Renouf, Antoinette. 2003. WebCorp: Providing a renewable data source for corpus linguists. In S. Granger and S. Petch-Tyson (eds.). *Extending the scope of corpus-based research. New applications, new challenges* (Language and Computers 48), 39–58. Amsterdam and New York: Rodopi.

Wolfgang Teubert (ed.). *Text corpora and multilingual lexicography* (Benjamins Current Topics 8). Amsterdam and Philadelphia: Benjamins, 2007. ix + 159 pp. ISBN 978-90-272-2238-1. Reviewed by **Christer Geisler**, Uppsala University.

Parallel corpora such as the English-Norwegian parallel corpus are by now well-established, but surprisingly few such corpora have actually been exploited as a source of information in the compilation of bilingual dictionaries. The present book brings together research from one much needed area: the use of corpus linguistic methods in bilingual and multilingual lexicography. It comprises a short preface, twelve articles, and an index. All contributors to the volume partici-