Charles F. Meyer. *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press, 2002. xvi + 168 pages. ISBN 0 521 80879 0 (hardback). ISBN 0 521 00490 X (paperback). Reviewed by **Claudia Claridge**, University of Kiel.

English Corpus Linguistics joins a number of other introductory corpus-linguistics books published in recent years. However, what distinguishes this publication from others available is that, instead of dealing with the field as a whole (e.g. McEnery and Wilson 1996/^{2}2001; Kennedy 1998) and/or pursuing a particular research agenda (e.g. Stubbs 1996; Biber et al. 1998), it can be described as

a kind of basic manual for corpus construction and analysis, with the emphasis on the former. Thus, it fills a gap in the existing literature.

The structure of the book falls into five sections. First, there is a preface presenting basic definitions and aims, followed by a first chapter linking corpus linguistics with linguistic theory and (practical) applications of corpus linguistic research. Then come three chapters (2-4) describing corpus construction from planning, via collection and computerization to corpus annotation, and one chapter (5) presenting a detailed case study of corpus analysis. Finally, a very brief sixth chapter both sums up and highlights possible future developments of the areas dealt with in the book. The whole is rounded off by two appendices listing available corpus resources and concordancing programmes.

In the preface, Meyer states his view of corpus linguistics as essentially a methodology, not a linguistic theory, and argues that, therefore, an increased awareness of methodological assumptions and procedures on the part of both corpus creators and users is vital for the progress of corpus linguistics (p. xiv). Corpus linguistics is indeed probably best viewed as a methodology; however, some further discussion of how the choice of a particular methodology correlates with broad, pre-existing theoretical assumptions about language and has potential theoretical repercussions or – to mention a clearly contrary view – can in fact be seen as a linguistic paradigm in its own right (cf. corpus-driven linguistics, Tognini-Bonelli 2001), would have provided a more balanced and informative approach. The preface defines a corpus as "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (p.xi). This definition at first seems overly brief and general, but the approach is narrowed down to the creation of "balanced corpora" and their use in "descriptive linguistic analysis" (p. xv), thus excluding most corpus research in computational linguistics/natural language processing, for example. This seems a wise restriction as the corpus-linguistic views and needs of the approaches just mentioned differ considerably and would have made the book unwieldy. The intended audience of the book seems to be the beginner in corpus linguistics: although Meyer does not explicitly state this (speaking only of "corpus linguists" as such, p. xiv), the structure and content, including numerous very basic aspects, as well as the study questions at the end of each chapter, imply this readership.

Chapter 1 discusses the relationship of corpus linguistics to generative linguistics and to functional theories of language, concluding – unsurprisingly – that it is the latter, not the former, that shows any interest in corpus linguistics. While Meyer gives examples to show that corpus linguistics can in fact contribute not insignificant insights to generative theory (p. 4f.), he thinks it unlikely

that generative linguists will ever develop much interest in using corpora. If this is so, it prompts the question why corpus linguists repeatedly feel the urge to one-sidedly topicalize this ultimately not very fruitful issue. The greater part of Chapter 1 is devoted to an overview of the place of corpus-based research in various fields, ranging from grammar- and dictionary-writing to language pedagogy, and taking in historical linguistics and contrastive analysis on the way. The treatment here is necessarily cursory, but it serves the purpose of highlighting the wide range of the possible applications of corpora and of stimulating further interest in corpus linguistics in readers of many different linguistic persuasions.

Chapter 2 is concerned with the planning stage of corpus construction. Meyer stresses the importance of careful initial planning in setting up the criteria for collection, which are determined by the future uses of the corpus, while at the same time retaining flexibility for adjustment in the compilation process. The chapter presents a comprehensive and clear discussion of the following compilation criteria: size of corpus, genres, length of text samples, number of texts, range of speakers, time frame, native vs. non-native speakers, and sociolinguistic variables (age, gender, dialect, education). Throughout the discussion, alternative approaches are evaluated and problematic points highlighted, e.g. the difficulties probability sampling can present (p. 43f.). However, not all of the aspects are treated as thoroughly as one might wish, a case in point being the question of the inclusion of complete texts or of text samples. Discussion of this aspect is biased towards the latter solution, without a clear statement of the potential advantages of using complete texts, among them the uneven distribution of linguistic features throughout texts as well as the general consideration that text-linguistic studies (beyond register comparison) should also be possible with corpora. The chapter uses the *BNC* as its example for illustrating the various criteria, which does not seem to be the most logical or useful choice: how many *beginning* corpus linguists would start with compiling a corpus of that scale – and thus have corresponding problems? It might also have been helpful to list more clearly those corpora that are in some way representative in their treatment of one or the other criterion discussed, so that the interested reader could have a closer look for her/himself at corpus linguistic problems and solutions.

Chapter 3 deals with the practicalities of collecting and computerizing samples of spoken and written English. This is done in a very down-to-earth and helpful way, with close attention paid to technical points (e.g. recording and transcription equipment, OCRs), procedural aspects (e.g. record keeping, materials storage) and ethical/legal issues (recording permission, copyright). Some of

the information given here may become outdated fairly fast (e.g. technical aspects), but raising awareness of the menial and mundane aspects of corpus linguistics is a very necessary and laudable thing to do. However, the chapter could have been more detailed and comprehensive in some respects. Written texts are admittedly less problematic than spoken ones; none the less the treatment they receive here is somewhat too brief and neglects the challenges they potentially represent. A possible reliance on electronically available texts is presented in a rather optimistic light and scanning is too much taken for granted, the latter perhaps due to the double bias resulting from thinking mostly in terms of printed and modern texts. Hand-written modern texts (e.g. letters, student essays) are not mentioned at all, while older texts, and manuscripts especially, are touched on only briefly. The discussion of computerizing speech is more detailed and necessarily shades into annotation matters when intonation is mentioned. What is not mentioned here is the possibility of sound files accompanying the transcription (as is the case with *COLT* and the *Santa Barbara Corpus of Spoken American English*) and alignment of text and sound, a practice which, with increasingly available computer space, might – indeed should – become more common.

Annotation of various types, namely structural markup, part-of-speech tagging and parsing, is the topic of Chapter 4. According to Meyer, annotation is necessary for a corpus to be "fully useful to potential users" (p. 81), which seems to be putting things too strongly. First, there are numerous features which are (fairly) easily retrievable without (grammatical) annotation and many linguistic questions to be pursued which are not affected by the surface features of the text (layout etc.). Secondly, it is not sufficiently highlighted that any form of annotation, but especially grammatical annotation, is already an interpretation (although cf. Meyer's own remark that "tagsets reflect differing conceptions of English grammar", p. 90) – an interpretation, moreover, that might ultimately contribute to obscuring a feature an individual analyst is looking for. A good solution for the corpus creator might actually be to provide both an annotated and a 'bare' text version of a corpus. As to structural markup, this receives rather too brief a discussion; in consequence, the aims and potential linguistic usefulness of this type of mark-up does not become clear. Furthermore, the main example is SGML as used in the *ICE* project, which might not be the best choice, because it is merely SGML-conformant and predates the TEI guidelines. The *BNC* would have served as a better illustration here. Moreover, a more detailed one of the SGML/XML/TEI complex would have been an advantage, in particular as it is the only comprehensive system with aspirations to become a standard. In view of the fact that the book is also intended for the corpus user

(and not only the compiler), a discussion, however brief, of earlier and/or related but supplemented annotation systems (e.g. COCOA, RET) might have been included. The chapter also includes a treatment of speech/intonation annotation. A point that might have been mentioned in that context is that (some) intonation markup conventions can actually make analysis – especially automatic computer analysis – harder, e.g. forms such as *ti=me* in the *SBC* example on page 85. The corpus user perspective is somewhat neglected throughout Chapters 2-4; they would also have profited from a greater number of examples, e.g. showing different annotation systems (for the same text, perhaps) and texts at different stages of annotation. This would have been very useful for the novice corpus linguist in particular.

Corpus analysis, i.e. the user perspective, is the focus of Chapter 5 and is exhaustively illustrated with a single well-chosen case study, Meyer investigating the occurrence of pseudo-titles in the press sub-corpora of seven *ICE* corpora. The comparative approach provides the opportunity to look again in more detail at corpus compilation, representativeness, and available annotation, this time from the analyst's perspective. The chosen feature is one that is not automatically retrievable in an untagged/unparsed corpus (six of the seven corpora used). This may not be very typical of corpus linguistics methodology as a whole, but the choice highlights the point that automatic retrievability should naturally not be a guide to what is being researched. Unfortunately, Meyer does not comment on the manual retrieval procedure and its results, merely mentioning it (p. 119); there are certainly degrees of manual retrieval, and the process can also turn up findings at odds with those of automatic retrieval, as well as findings the researcher did not expect. Meyer argues for combining quantitative and qualitative aspects in the analysis of corpus data, a very important point as the balance can easily become tilted towards the former in corpus linguistics. The chapter works through the whole process of analysis step by step, thoroughly comparing options and motivating the decisions to be taken, and linking the aspect in hand to more general questions wherever possible. The whole research procedure thus becomes highly accessible and comprehensible even for readers with little to no experience in the field.

In conclusion, the work under consideration here is a very welcome addition to the range of corpus linguistic publications. It offers the beginner a brief yet valuable introduction to the basic aims and – especially – the research procedures of corpus linguistics and thus serves a real need. Perhaps the content of the book could have been more clearly reflected in the title in order to attract the attention of its intended readership. It can be argued that certain aspects have not been treated with sufficient explicitness and detail or in adequate depth, in par-

ticular for readers with little previous knowledge (cf. the remarks above), but remedying this point would have considerably increased the length of the book. However, one helpful addition would have been a 'further reading' section after every chapter.

References

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Kennedy, Graeme. 1998. *An introduction to corpus linguistics*. London and New York: Longman.

McEnery, Tony and Andrew Wilson. 1996/2nd ed. 2001. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Stubbs, Michael. 1996. *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford and Cambridge, Mass.: Blackwell.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam and Philadelphia: Benjamins.