This is a pre-final version. Published version in: Studies in Language 29 (2005): 261-264

Review of:
Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge et al.: Cambridge University Press. xvi + 168 pp.

Carsten Breul

1 *English Corpus Linguistics* (henceforth *ECL*) is an introductory work which emphasises the methodological aspects of corpus linguistics. This emphasis is prepared for by the author's observation at the very beginning of the book (p. xi) that corpus linguistics is rather a methodology than a field of study such as sociolinguistics or psycholinguistics. Consequently, relatively much space of the book's six chapters is devoted to such topics as "Planning the construction of a corpus" (title of ch. 2), "Collecting and computerizing data" (title of ch. 3) and "Annotating a corpus" (title of ch. 4), while the presentation and discussion of results of corpus linguistic research is more restricted in scope. This weighting of topics covered distinguishes *ECL* from other more recent introductions to corpus linguistics. The approach is legitimate both in itself and in comparison to these other works, which *ECL* thus complements.

2 After having defined 'corpus' as "a collection of texts or part of texts upon which some general linguistic analysis can be conducted" (p. xi) in the preface, Meyer (henceforth M) discusses the relation between corpus research and generative grammar on the one hand and functional descriptions of language on the other hand in the first two sections of chapter 1 ("Corpus analysis and linguistic theory"). The bottom line here is that "corpora are much better suited to functional analyses of language" (p. 5) and that "linguistic analysis will benefit if it is based on real language used in real contexts." (p. 11) As an example of a functionally oriented corpus analysis, M presents an extended summary of an earlier study of elliptical coordination. In the remainder of chapter 1, M gives an overview of some other corpus-based works and points out that corpora have also been used in the writing of various reference grammars. Lexicography is presented as a field of research which has benefited immensely from corpus analysis. Other areas in which corpus-based research has been undertaken are language variation, historical linguistics, contrastive analysis, translation theory, language acquistion, and language pedagogy. Within the domain of natural language processing, corpora serve in the development of taggers, parsers and information retrieval systems.

Chapter 2 contains a detailed discussion of the considerations that are involved in the planning stage of corpus construction. The British National Corpus serves as a model here. Some of the issues that are addressed are the following: the number of texts and range of speakers/writers to be included for the corpus to be representative; which time span to choose for the dates of origin of the texts to be included; the problem of making sure that the text producers to be selected are native speakers of the language to be studied; the sociolinguistic variables to be taken into account, such as the gender, age, level of education, dialectal background of the text producers as well as their social relationship in case conversation is to be sampled.

An equally detailed discussion of the techniques and of the problems involved in the collection and computerisation of written and spoken language is presented in chapter 3. The topics range from ethical and legal issues (e.g. surreptitious recordings of speech, obtaining permission to use copyrighted material) over technical options (e.g. optical character recognition, downloading texts from the WWW, equipment for recording, digitising and transcribing speech) to questions such as how to keep records of the sampled texts and how to represent speech in orthographic form. Several addresses of WWW-sites are given which contain downloadable free- or shareware of applications which support the digitisation and transcription of speech or contain otherwise helpful material.

Chapter 4 introduces the reader to SGML-conformant structural markup of corpora and gives an overview of the word class tagging and parsing procedures. Semantic and discourse tagging as well as problem-oriented tagging are also touched upon.

Chapter 5 presents a step-by-step description of a corpus-based study of pseudo-titles. The study starts off with the issue of framing a research question. M appropriately points out that "it is imperative before undertaking a corpus analysis to have a particular research question in mind, and to regard the analysis of a corpus as both 'qualitative' and 'quantitative' research – research that uses statistical counts or linguistic examples to test a clearly defined linguistic hypothesis." (p. 102) After having addressed the question of which corpus is suitable for the investigation of pseudo-titles, M turns to the problems involved in extracting data from the corpus. Next comes a decision about which types of information about the occurrences to be extracted from the corpus should be recorded. In the discussion of how to actually locate and extract relevant instances, M explains the use of standard concordancers and points out that it may be necessary to write one's own programs or to use other publicly available software in addition in order to perform specific tasks not covered by the standard concordancers. After extracting the relevant instances from the corpus and counting the features one is interested in, the data obtained are subjected to statistical analysis. M explains that "by conducting a more rigorous statistical evaluation of their results, corpus linguists can not only be more confident about the results they obtain but may even gain new insights into the linguistic issues they are investigating." (p. 120f.) Several statistical concepts (e.g. standard deviation, kurtosis, skewness, significance) as well as tests and procedures (e.g. chi-square test, loglinear analysis) are touched upon, and the results of their application to the pseudo-title data are interpreted.

In the final chapter 6 ("Future prospects in corpus linguistics"), M reasonably predicts that "[e]asing the creation of spoken corpora remains one of the great challenges in corpus linguistics, a challenge that will be with us for some time in the future." (p. 139) Although they will still require a considerable amount of work, structural markup and (word class) tagging of corpora will become increasingly accurate. This is also true for parsers, but on a much lower level.

Each of chapters 1-5 has a concluding section which summarises the main points and ends with study questions. Most of the questions aim at a recapitulation of some passage(s) in the respective chapter, two or three of them are explorative. *ECL* has two appendices, a list of references and a combined names and subject index. Appendix 1 contains a list and brief description of about 50 corpora plus about ten other items (e.g. corpus projects, organisations, the ICAME bibliography), often with Internet addresses as references. Appendix 2 lists 16 concordancing programs and their Internet addresses.

3 Chapters 2 and 3 of *ECL* are excellent, so that the book can be recommended especially to anyone interested in a clear and thorough description and discussion of issues involved in planning and constructing a corpus. Although the reader gets a concise overview of structural markup, tagging and parsing in chapter 4, one would wish the discussion of word class tagging in particular to be more detailed and problem-oriented. Indeed, M explicitly warns us against erroneous tagging (p. 89). But for the exposition to match the thoroughness of chapters 2 and 3, some more pitfalls of tagging and examples of how erroneous tagging may distort results would have to be presented. The case study of pseudo-titles illustrates well the kinds of insights to be gained by corpus research. However, several passages which deal with the statistical analysis of the data in this study provide insufficient explanation and lack clarity.

Although some of M's statements about the relation between corpus linguistics and generative grammar appear to aim at bridging the gap (see e.g. p. 1), there are others which a generative grammarian might find off the mark and offensive. For example: "Unlike generative grammarians, corpus linguists see complexity and variation inherent in language" (p. 3).

I have detected about 18 typographical (etc.) errors in *ECL*. What I find somewhat disturbing is that the term *enclitic* is given as *enclictic* two times on p. 87 and *proclitic* as *proenclictic* on p. 90.