Tony McEnery and **Andrew Wilson**. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996. ISBN 0-7486-0808-7 (hardback); ISBN 0-7486-0482-0 (paperback). Reviewed by **Charles F. Meyer**, University of Massachusetts at Boston.

The publication of *Corpus Linguistics* is noteworthy: as the first volume in the new series 'Edinburgh Textbooks in Empirical Linguistics', this textbook reflects not only the increasing importance that empirically-based studies of language are coming to play in linguistics but the prominent role that corpus linguistics has assumed among the many different empirically-based approaches to language study. In *Corpus Linguistics*, McEnery and Wilson (hereafter MW) very clearly introduce the field of corpus linguistics to students, providing a very effective overview of the key linguistic and computational issues that corpus linguists have to address as they create corpora and conduct analyses of them.

Corpus Linguistics is divided into seven chapters that focus on a number of topical issues in corpus linguistics, issues ranging from the theoretical underpinnings of corpus linguistics to the various annotation schemes that have been developed to tag and parse corpora, the quantitative research methods used to analyze corpora, the types of linguistic studies that have been carried out on corpora, and the contributions that computational linguistics has made to the creation and analysis of corpora. Each of these topics is approached in a clear and readable format that will make this text valuable not just to students but to specialists in other areas of linguistics interested in obtaining information about corpus linguistics.

After noting in the opening chapter ('Early Corpus Linguistics and the Chomskyan Revolution') that corpus linguistics is more a methodology (a way of approaching language study) than a sub-discipline in linguistics, MW continue with a discussion of the methodological assumptions that characterize corpus linguistics and distinguish it from Chomskyan approaches to language study. They note the difference between rationalist and empiricist approaches to language study, and detail the classic Chomskyan arguments that have been leveled over the years against empiricist studies of language. Because corpora contain data reflecting 'performance', they are of little value in studying 'competence', the most important area for linguists to study. In addition, 'corpora are "skewed"' (p 8), in the sense that they do not contain all of the possible structures that exist in a language. These objections led Chomsky to

value introspection as the best way of describing a language, and to reject descriptions of actual language use based on analyses of corpora.

Although MW acknowledge some validity to Chomsky's objections to corpus analyses, they counter these objections with a number of arguments in favor of corpus linguistics. A corpus, for instance, can be used to verify introspective judgments, and to overcome the problem of basing grammatical arguments on 'artificial data' (p 12). Moreover, œorpora can provide important information on the frequency of grammatical constructions, and the sophisticated software developed to analyze corpora can give the linguist access to much important information on grammatical structure present in corpora that have been tagged and parsed.

Although Chapter 2 ('What is a corpus and what is in it?') purports to describe what a corpus is, it is primarily a chapter about what corpora look like—specifically the annotation schemes that have been developed to tag and parse them. MW only briefly discuss the issues one must confront when creating a corpus (eg the size of the corpus), and while they discuss many methodological concerns throughout the book, it would have been desirable to have grouped these issues together in a single chapter and to have discussed how the representativeness of a corpus is influenced by such variables as its length, the genres it contains, and the types of individuals whose speech and writing are included in the corpus.

The strength of Chapter 2 is its discussion of annotation schemes, which is detailed and very well illustrated. MW provide a very clear overview of the TEI (Text Encoding Initiative), illustrating how the various tags developed by TEI can be used to create 'headers' (in which information about authors/speakers, titles, dates of publication, etc can be recorded) and to mark up texts themselves with information on paragraph boundaries, type faces, and so forth. The remainder of the chapter focuses on the various schemes that have been developed to annotate linguistic information in corpora. MW first compare tagging schemes from corpora as diverse as the British National Corpus and the CRATER Corpus of Spanish, and then describe the process of developing the CLAWS tagging schemes at Lancaster University. The chapter concludes with a discussion of parsing schemes and of how corpora can be annotated with markup revealing their semantic, discoursal, and prosodic structure.

Chapter 3 ('Quantitative data') discusses the importance of using quantitative research methods to analyze corpora. MW first distinguish qualitative from quantitative research methods, and make the very im-

portant point that the linguistic claims one makes about a corpus depend crucially upon whether the corpus being analyzed is valid and representative; that is, has been created in a manner that allows the analyst to make general claims about, for instance, the genres represented in the corpus. MW then describe the major kinds of statistical analyses that can be performed on corpora. The difficulty of a chapter of this type is that statistics is such a vast area that it is hard to determine precisely how much detail needs to be provided. But the level of detail in this chapter is most appropriate, and there is much useful information provided on how corpora can be statistically analyzed — from methods as basic as frequency counts to those as sophisticated as factor analysis and loglinear analysis (as done with programs such as VARBRUL).

Chapter 4 ('The use of corpora in language studies') surveys the kinds of empirical linguistic analyses that corpora can be used to conduct. MW open the chapter with a discussion stressing the importance of empirical studies of language, noting that they 'enable the linguist to make statements which are objective and based on language as it really is rather than statements which are subjective and based upon the individual's own internalised cognitive perception of the language' (p 87). This statement is very convincingly supported in the remainder of the chapter, which contains a very good discussion of how corpora can be used to study language at all levels of linguistic structure (eg phonetics/phonology, syntax, and semantics) and from many different theoretical perspectives (eg pragmatics, sociolinguistics, and discourse study). As each of these areas are described, MW include descriptions of previous studies conducted in the areas to effectively illustrate how work in the area is conducted and has yielded important information.

The first four chapters of *Corpus Linguistics* are concerned with issues relevant to linguists using corpora to carry out purely linguistic studies. Chapter 5 ('Corpora and computational linguistics') moves to an allied discipline, natural language processing (NLP), and discusses issues such as tagging and parsing from a more computational perspective. Although linguists who use corpora for grammatical analysis may not have an immediate interest in NLP, the research in this area has led directly to improvements in recent years of taggers and parsers — software responsible for annotating corpora and making it easier for linguists to extract information from them. The discussion in this chapter is brief, but does an excellent job of summarizing the theoretical issues underlying the development of taggers and parsers and the role that they play in areas such as lexicography and machine translation.

Chapter 6 (‘A case study: sublanguages’) draws upon much of the information presented in the previous chapters to carry out a sample grammatical analysis of three corpora from three distinct genres: a series of IBM manuals from the IBM Corpus, transcriptions of Canadian parliamentary speeches found in the Hansard Corpus, and fiction from the APHB (American Printing House for the Blind) Corpus. MW analyze these three corpora to advance the hypothesis that the language of the IBM Corpus is a ‘sublanguage’; that is, ‘a version of a natural language which does not display all of the creativity of that natural language[and which] will show a high degree of **closure** [emphasis in original] at various levels of description (p 148). Before pursuing this hypothesis, MW discuss the importance of evaluating a prospective corpus to determine whether the genres it contains are appropriate for the analysis being conducted, and whether the manner in which the corpus has been annotated will allow for the retrieval of the grammatical information desired. MW conclude that the corpora they have chosen are appropriate for their study of sublanguages, and they then conduct analyses to determine the degree of lexical closure, part-of-speech closure, and parsing closure that exists in each of the corpora. In general, this analysis verified MW’s hypothesis and demonstrated that the IBM Corpus (in comparison with the other corpora) is a more ‘restricted genre’ and contains fewer different types of words and sentence-types (though the words it contained corresponded to more parts of speech than the words in the other corpora did).

The final chapter (‘Where to now?’) nicely rounds out the book with a discussion of issues that corpus linguists will need to address in the future corpora that they develop, a series of ‘pressures’ to increase the length of corpora, to make them conform to industry as well as academic standards, and to have corpora draw upon evolving computer technologies in their creation, such as the many multi-media currently being developed. This chapter provides a fitting conclusion to a text that provides a very perceptive overview of the field of corpus linguistics that will be a good choice for use in any introductory course on corpus linguistics.