An Introduction to Corpus Linguistics

Graeme Kennedy
(Victoria University of Wellington)

London: Addison Wesley Longman
(Studies in language and linguistics,
edited by Geoffrey Leech and Jenny
Thomas), 1998, xii+315 pp; hardbound,
ISBN 0-582-23153-1, £42.00; paperbound,
ISBN 0-582-23154-X, £16.99

Reviewed by
Vincent B.Y. Ooi
National University of Singapore

This timely book joins the growing number of leading introductory volumes on corpus linguistics, including McEnery and Wilson (1996) and Biber, Conrad, and Reppen (1998). The former volume represents the first-ever introductory textbook on corpus linguistics; the latter, also recently published, is strong on reporting many of the authors' own analyses of speech and writing and providing "methodology boxes" for the computer processing of text using the corpus-based paradigm. Space constraints do not permit the comparison of all three volumes, especially in relation to the use of the corpus-based approach for natural language processing. In the present instance, it is to his credit that Graeme Kennedy has single-handedly achieved the difficult task of systematically providing an historical overview and evaluating the significance of many important studies that currently define the corpus-based approach.

The book is organized as follows:

1. Introduction (12 pages); the chapter includes the notion of a corpus and the scope of corpus linguistics.
2. The design and development of corpora (75 pages); the chapter includes the types of corpora, corpus design, representativeness, and markup.
3. Corpus-based descriptions of English (116 pages); the chapter details the results of various studies of corpora from lexical, grammatical, pragmatic, and language-variation perspectives.
4. Corpus analysis (64 pages); the chapter includes some common corpus-based software, including software for corpus annotation, search, concordancing, and retrieval.
5. Implications and applications of corpus-based analysis (27 pages); the chapter includes a treatment of language as "possibility" and "probability" and, from these two complementary perspectives, how the corpus-based approach relates to the practice of computational linguistics, language teaching, and computer-assisted language learning.

Kennedy's choice of treatment of these chapters clearly indicates the precedence of linguistic issues over computing ones. Indeed, for him, "the most important skill is not to be able to program a computer or even to manipulate available software ...

[but] to be able to ask insightful questions which address real issues and problems in theoretical, descriptive and applied language studies" (p. 3). Notwithstanding this preference, Chapter 3 does introduce the reader to a range of computer software for the linguistic analysis of text, including software for lemmatization, word-class tagging, semantic tagging, parsing, and concordancing. While Kennedy's main aim is to introduce the reader to the linguistic rationale behind the processing and refer the reader to the relevant sources for further understanding of the software, he does take pains also to introduce the reader to the actual operations of the software. For instance, in his treatment of Lancaster's CLAWS word-tagging system, he details the evolution of this software from the first version of the tag set to the C5 tag set. He also details the modular components of the software and provides a number of figures of the processed output in order to enhance the reader's understanding (pp. 212–221). However, most of the treatment of CLAWS is based on material written by the Lancaster team between 1987 and 1994; the latest version of CLAWS, using the C7 tag set (as detailed by Garside, Leech, and McEnery 1997), is not included in the book. To be fair, a piece of linguistic software such as CLAWS is continually developed, and it would be hard to keep pace with its latest release. Nowadays, a central means of resolving such problems is to get the reader into the practice of using the World Wide Web for updates of the latest technological developments. Therefore, by the same token, Kennedy's documentation of the ENGCG Parser of English (which is software for word-class tagging and syntactic parsing; see pp. 241–243) could be improved upon by providing the relevant hypertextual links. The reader should know that analysis is available via e-mail (engcg@lingsoft.fi), and the latest documentation and parsing demos are now on the Web (http://www.ling.helsinki.fi/~tapanain/cg/index.html). Instead, throughout the book, there are very few hyperlinks (which include, for instance, one link to the British National Corpus on page 54 and another to the WordSmith Tools software on page 267). Perhaps the evolving nature of the Internet and the World Wide Web means that it would be less wise to provide Web links, which tend to get outdated after a while, since computer technology will (without our vigilance) sooner or later tend to outdate us all. But, then one can always look forward to the next book edition.

Kennedy's main focus is detailing the workings of language using the corpus-based approach. Chapter 3, by far the longest, is devoted to an elucidation of various linguistic investigations of the probabilistic nature of language. Language is rightly seen as involving not only the Chomskyan set of possible or impossible structures but (perhaps, more importantly) the set of probable or improbable structures. This point concerning the nature of language is also made by no less a well-known linguist than Michael Halliday (1992), who views lexicogrammar as an inherently probabilistic system.

It is precisely the ability to interrogate or train the corpus for its relevant statistical probabilities that has allowed the emergence of natural language processing systems that are more robust and successful in achieving language disambiguation. This central fact has made the corpus so relevant to the practice of computational linguistics nowadays. In his usual systematic manner, Kennedy traces the developments that have led to the much closer relationship between corpus linguistics and computational linguistics (pp. 276–280).

An Introduction to Corpus Linguistics does not set out to be a book that details linguistic programming for corpus analysis. Rather, the richness of the book is the author's vast experience and knowledge in evaluating the development of corpora in linguistic research; the book, in the author's own words, looks back "to the pre-electronic age as well as to the massive growth of computer corpora in the electronic age." A strong feature of the book is the inclusion of many useful figures and tables that capture the research findings in a concrete manner for the reader. This excellent book should be required reading for students and teachers involved in corpus-based research and is generally useful to anyone who seeks a more comprehensive understanding of the resurgence of empirical linguistics.

References

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press.

Garside, Roger, Geoffrey Leech, and Tony McEnery (editors). 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Addison Wesley Longman, London.

Halliday, M. A. K. 1992. Language as system and language as instance: The corpus as a theoretical construct. In Jan Svartvik, editor, *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82.* Mouton de Gruyter, Berlin, pages 61–77.

McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics.* Edinburgh University Press.

Vincent B.Y. Ooi is Senior Lecturer in the Department of English Language and Literature, National University of Singapore. He is also presently cochairman of the Information Technology Committee of the Faculty of Arts and Social Sciences. His teaching and research interests include empirical linguistics, lexicogrammar, computer processing of linguistic texts and new varieties of English. Ooi's address is: Department of English Language and Literature, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260; e-mail: vinceooi@nus.edu.sg; WWW: http://www.fas.nus.edu.sg/ell/Vincent