Corpus Linguistics

Tony McEnery and Andrew Wilson
(Lancaster University)

Edinburgh: Edinburgh University Press
(Edinburgh Textbooks in Empirical
Linguistics, edited by Tony McEnery
and Andrew Wilson), 1996, x+209 pp;
distributed in the U.S. by Columbia
University Press; hardbound, ISBN
0-7486-0808-7, $70.00; paperbound,
ISBN 0-7486-0482-0, $24.50

Language and Computers: A Practical Introduction to the Computer Analysis of Language

Geoff Barnbrook
(University of Birmingham)

Edinburgh: Edinburgh University Press
(Edinburgh textbooks in empirical
linguistics, edited by Tony McEnery
and Andrew Wilson), 1996, ix+209 pp;
distributed in the U.S. by Columbia
University Press; hardbound, ISBN
0-7486-0848-6, $70.00; paperbound,
ISBN 0-7486-0785-4, $24.50

Reviewed by
John M. Kirk
The Queen's University of Belfast

The appearance of not one but two introductions to corpus linguistics within the same series shows the maturation and diversification of this fledgling subdiscipline within linguistics. McEnery and Wilson offer an overview or annotated report on work done within the computer-corpus research paradigm, including computational linguistics, whereas Barnbrook offers a guide or manual on the procedures and methodology of corpus linguistics, particularly with regard to machine-readable texts in English and to the type of results thereby generated.

Whereas McEnery and Wilson recognize that the distinguishing features of corpus linguistics rest with its computer-aided empiricism, they are eager to line it up alongside cognitive rationalism in an effort to show the complementarity and interdependence of the two. As they argue, the advantages of a corpus-linguistics approach are that it is invariably systematic and rigorous, and that linguistics based on a corpus acts as a yardstick or control to linguistics based on artificial or introspective data. Of these current research paradigms, the authors' discussion offers fair and balanced criticism.

In the central core of the book, McEnery and Wilson present overviews of the theory and practice of corpus linguistics, the relative merits of qualitative versus quanti-

tative approaches to language study, and a report (within linguistics) on subject-based studies using a corpus linguistics approach of which the authors evidently approve. The three central chapters constitute essential reading for every newcomer to the field. In Chapter 2, the authors emphasize the key factors in a corpus-linguistics approach: sampling, representativeness, size, format (and all their many sets of choices). In Chapter 3, by emphasizing the dilemma of interpretation inherent in any claims about a language as a whole that are based on no more than a sample, they focus on the benefits of quantitative information that is statistically reliable and from which robust generalizations may be made.

As graduates of Lancaster, whose mentors and now collaborators have included Professors Garside and Leech, the authors reflect the preferences of their training: they favor well-controlled (usually smaller) corpora (random *stratified* samples rather than simply random samples) with good extra-textual information stored in headers, which can be reused, with the same data yielding increasingly complementary results, such as the Brown and Lancaster–Oslo/Bergen written corpora, or the Lancaster Spoken English Corpus.

The many criteria expounded in these chapters will almost certainly lead others embarking on the use of corpora to justify their many strategic decisions. Although newer investigations might come to upstage some of the topic-centered studies highlighted here, the general criteria seem sufficiently assured to be of service for some time to come—particularly those of verifiability, total accountability, and strength of argument.

The final two chapters deal quite specifically with computational approaches and numerous accompanying issues, and it is doubtful whether they will interest many who are not natural language processors—certainly, there is little for the student of English who is primarily interested in the analysis and description of data, for whom these chapters are almost certainly beyond reach without further specialized training. The chapters understandably reflect the authors' experience as programmers with Lancaster's pioneering projects: the development of automatic techniques for word-class tagging and syntactic parsing (as evidenced by the development of the CLAWS tagger and on the Lancaster Parsed Corpus and the Lancaster–Leeds Treebank), for automatic machine translation (as evidenced by work on the Canadian Hansard Corpus and on the Crater Corpus), and for the automatic identification of sublanguages (as in the IBM manuals study), on all of which they write with conviction and enthusiasm. Whatever the merits of these various approaches, the discussion lacks a demonstration of the results of machine translation in action, so that readers might see how the approach could benefit them. Nor are there any actual samples from the IBM manuals, whose words and sentences and ratios of words to sentences are counted and presented on graphs ad nauseam, nor a passage illustrating the finite subset of the language (referred to as "closure") in action. Indeed, the entire book is short on substantiating examples, so that it rarely emerges what scholar *X*'s work on subject *Y* contributes to *Y*, however much favored by the authors.

Whereas McEnery and Wilson attempt the broad sweep across this burgeoning field, Barnbrook attempts a much deeper analysis of the exploitation of corpora (McEnery and Wilson's Chapter 2). There are separate chapters on using a computer in the first place, issues arising from the choice of data and their capture, the use of frequency lists, the generation of concordances, the analysis of collocations, and the question of tagging, parsing, and other kinds of in-text annotation. All of these chapters are very accessible to students of English because they are richly illustrated with examples. This same is true of Barnbrook's next chapter, which provides an overview of uses of corpora and their exploitation within natural language processing: stu-

dents of English can see instances that relate to their use of computers (as simply in word-processing) or examples of what is automatically analyzed or, in the case of the so-called sublanguage of dictionary definitions, being investigated. Barnbrook's lucidity, his practical guidance in showing the reader what to do and what comes out of doing it, his copious examples and illustrations, all recommend his book highly. Just as McEnery and Wilson reflect their Lancaster training, so Barnbrook's emphasis on linguistic description reflects his association with the COBUILD dictionary and grammar project at Birmingham. My only reservation about Barnbrook is his excessive use of examples of spelling variants from Chaucer; whereas this use is understandable in view of Barnbrook's doctoral research, even many native-speaker students are neither excited nor curious about the language of the past nor about spelling, so that its use undervalues the approach in its perception of relevance, tending, I have unfortunately found, to discourage. Corpus linguistics is so much more than variants within late fourteenth-century spelling! With so much contemporary data running through Birmingham's Bank of English, their use for illustrative or investigative purposes here could only find approval for the book by a much wider and, I'd guess, more widely international audience.

Although these books are undoubtedly playing to the same tune, with much common ground and overlap, their achievement is to reflect the different traditions emerging in the field by which the authors have become influenced and which ultimately differ between "doing computing" and "doing language." Despite their differences, they each show that corpus linguistics has two central planks: the ways in which the computer is usable for language study and, on this basis, the generation of new descriptions and understanding.

John M. Kirk lectures in English at The Queen's University of Belfast. He compiled the Northern Ireland Transcribed Corpus of Speech and is collaborating on the International Corpus of English. He has published several articles in *Corpus Linguistics* and is working on an introductory textbook. Kirk's address is: School of English, The Queen's University of Belfast, Belfast BT7 1NN, Northern Ireland; e-mail: J.M.Kirk@qub.ac.uk