Corpus Linguistics: Readings in a Widening Discipline

Geoffrey Sampson and Diana McCarthy (editors)
(University of Sussex)

London: Continuum, 2004, xiv+524 pp; hardbound, ISBN 0-8264-6013-5, $170.00

Reviewed by
Robert Malouf
San Diego State University

Not that long ago, all linguistics was corpus linguistics. For much of the twentieth century, though, changes in theoretical fashion put linguists for whom collections of texts are the primary object of study at the margins of the field. Now things are shifting again, and a new generation of linguists with a wakening interest in linguistic data is returning to the field's corpus-based roots. What they are finding is that technological and methodological advances have quietly revolutionized corpus linguistics. In *Corpus Linguistics: Readings in a Widening Discipline,* editors Geoffrey Sampson and Diana McCarthy have put together a volume aimed at filling in some of what has happened in corpus linguistics while no one was watching.

This collection of 43 reprinted papers, with original publication dates ranging from 1965 to 2002, includes a few of the usual suspects and a number of more-often-cited-than-read classics, making it an ideal source for an upper division course on corpus linguistics, or to supply what new corpus linguists (or their teachers) wish they had learned in grad school. Beyond these core contributions, however, the editors have wisely included an idiosyncratic selection of more-obscure papers guaranteed to stir the imagination of anyone with a professional interest in language. Papers on, for example, second language teaching, the discourse properties of Internet Relay Chat, and "non-indigenous minority languages" (such as South Asian languages spoken in the UK) show some of the possibilities that corpus-based methods have to offer beyond their philological foundations.

While presented in chronological order, the papers in this volume for the most part fall into three classes. The largest is made up of papers on corpus design and methodology, such as Nelson Francis on the construction of the Brown corpus, Burnage and Dunlop on the British National Corpus, and Böhmová and Hajičová on the Prague Dependency Treebank. The second group are papers on the descriptive analysis of corpora. One, an excerpt from Charles Fries's *The Structure of English,* predates computational corpus analysis, but most of the rest are quantitative studies of English grammatical features. This group includes papers on cleft and pseudo-cleft constructions (Collins), *that* vs. null complementizers (Rissanen), and the use of terms of abuse (McEnery et al.). The third class of papers addresses technical issues in computational linguistics. Some of the classics in this group include Gale and Church on smoothing corpus counts, Hindle and Rooth on prepositional phrase attachment, and Briscoe and Carroll on parser evaluation. In this group we also find a number of less-well-known papers, such as contributions on coding dialog moves and on predicting intonation in spoken language corpora. Finally, there are a handful of papers that offer some personal and historical perspective on the field, most notably Sampson's "Reflections of a dendrographer," the text of a lecture originally delivered at a 2001 conference in honor of pioneering corpus linguist Geoffrey Leech.

Beyond the selection of papers, the "value added" material in this collection is uniformly helpful and well done. Sampson and McCarthy have provided each contribution with an introduction that sets the scene for the paper, putting it into its historical context and providing references for any necessary background (both within the collection and to external sources), and pointing out what hindsight has revealed as some of its strengths and weaknesses. In addition, the editors have normalized the notation and references in each paper and compiled a comprehensive index, bibliography, and list of abbreviations. In an interesting modern addition, citations of online sources in each chapter have been replaced with references to a single list of URLs. While this notation is explained in the introduction, on my first pass through the book I didn't realize that "[jcc]" was a reference to Jean Carletta's homepage at Edinburgh. And, inevitably, a reorganization of the School of Informatics's Web space means that the URL given in the book is no longer valid. Perhaps a volunteer could be persuaded to set up a website with up-to-date URLs and other online resources for the papers collected here.

If I had to name a complaint about this volume, it would be the shortage of papers that directly address current issues in theoretical linguistics. Indeed, this collection does not seem to have been assembled with theoretical linguists in mind; readers are directed to Jurafsky and Martin (2000) for a discussion of phrase-structure grammar. It's hard to fault the editors for this, given the principled stance many theoretical linguistics have taken against corpus-based methodology. However, this is rapidly changing (witness any recent issue of *Language*), and I suspect there will be many generativists among the newcomers to corpus linguistics in the coming years. They will find less an offer for them here than others coming from neighboring disciplines.

With that small caveat, I highly recommend *Corpus Linguistics: Readings in a Widening Discipline* to anyone with any interest in linguistic corpora. Besides bringing students and practitioners of other branches of linguistics up to speed on advances in modern corpus linguistics, this volume may as a secondary effect help inform applied, computational, descriptive, and even theoretical linguists about each other's fields. In what other volume is one likely to find papers on both second language curriculum design and methods for constructing treebank grammars? In any event, this collection is a wonderful addition to the currently available textbooks on corpus linguistics, and provides yet another reason that, as the editors say in their introduction, "this is a good time to become a corpus linguist" (page 3).

Reference

Jurafsky, Daniel and James H. Martin. 2000. *Speech and Language Processing*. Prentice Hall, New Jersey.

Robert Malouf is an assistant professor in the Department of Linguistics and Oriental Languages at San Diego State University. His research and teaching focuses on corpus linguistics, statistical natural language processing, and morphosyntactic theory. Malouf's e-mail address is rmalouf@mail.sdsu.edu.