English Computer Corpora: Selected Papers and Research Guide

Stig Johansson and Anna-Brita Stenström (editors)
(University of Oslo and University of Stockholm)

Berlin: Mouton de Gruyter (Topics in
English Linguistics 3, edited by Jan
Svartvik and Herman Wekker), 1991,
vii + 402 pp.
Hardbound, ISBN 3-11-012395-9 and
0-89925-675-9, DM $158.00

Reviewed by
Douglas Biber
Northern Arizona University

Although corpus linguistics has been around for at least three decades,[1] it is only in the last several years that a corpus-based, probabilistic approach has become widely accepted. Thus, as recently as 1987, one of the goals stated in Leech's (1987) introduction to a book on corpus-based processing was that readers would come away from the book "with less inclination ... to assume that the paradigm of AI is the only one to be successfully applied to natural language processing." This situation has changed dramatically in the last few years, with empirical, corpus-based approaches currently being widely used for probabilistic part-of-speech taggers, syntactic parsers, and computational lexicography.

However, because of the earlier predominance of AI approaches in CL, many corpus linguists came to think of themselves as comprising a distinct subfield. Although the earliest work in this area is usually associated with Brown University in the United States (see footnote 1), most corpus-based research has been carried out in Great Britain and Scandinavia. In 1977, a separate organization was established for corpus linguistic research—the International Computer Archive of Modern English (ICAME)—and the first ICAME conference was held in 1979.

The book under review here, *English Computer Corpora: Selected Papers and Research Guide* (henceforth, *ECC*), comes from this research tradition; it contains the conference proceedings from the Tenth ICAME meeting, held in Bergen in 1989. According to the introduction by Johansson (p. 4), the chapters are not intended as state-of-the-art papers, but they do provide "good indications of the sort of work currently undertaken" in this tradition.

The book is divided into eight major sections: probabilistic grammatical analysis, syntax, lexis, speech, regional/social variation, specialized corpora, software, and a reference section. The book ranges from articles with primary computational emphases (e.g., on techniques for probabilistic parsing, or descriptions of software tools for corpus analyses) to papers with primary linguistic emphases (e.g., functional analyses of infinitival complement clauses and modals).

Two of the chapters in the first section, dealing with probabilistic tagging and parsing techniques, are written by some of the first researchers to demonstrate the

[1] In the early 1960s, Francis and Kučera began work on the Brown Corpus, a computer-based collection of 500 published texts of American English from 15 major text categories, comprising one million words of text; the first user manual for the corpus was published in 1964 (Francis and Kučera 1964).

importance of a probabilistic approach for natural language processing: DeRose, and Leech and Garside (cf. DeRose 1988; Garside, Leech, and Sampson 1987; Church 1988). The paper by DeRose briefly describes several tests of his linear-time algorithm for word tagging (Volsunga), on English and Greek texts; for example, he compares the results based exclusively on relative word-tag probabilities with those based on collocational probabilities (achieving over 90% accuracy in both cases). Leech and Garside describe the history of their project to develop a large 'treebank' of parsed English texts, with the eventual goal of using the treebank to derive a probabilistic phrase structure grammar; four stages of research are described, using both hand parsing and automatic parsing. (The chapter by Souter and O'Donoghue describes research in progress on a "Realistic Annealing Parser.")

There are three chapters that focus on syntactic description, illustrating quite different analytical approaches. De Haan uses quantitative statistical techniques to analyze the functions of nominal postmodifying clauses, while Mair argues (from an analysis of infinitival complement clauses) that qualitative corpus-based analyses are equally important because the corpus provides "authentic and realistic data." Ihalainen compares variation in the form of grammatical subjects in spoken corpora of educated and dialectal English.

Lexical matters are dealt with by Brekke, Vossen, Altenberg, and Collins. Brekke analyzes the various senses of the word *wall* in the Brown and Lancaster–Oslo/Bergen (LOB) corpora, and Vossen analyzes the treatment of polysemy in the *Longman Dictionary of Contemporary English*. The paper by Altenberg describes amplifier collocations in spoken English (based on the London–Lund Corpus), emphasizing the repetitive nature of speech and the high frequencies of some recurrent amplifier combinations. Collins compares the functions of *will* and *shall* in Australian, British, and American English. (The papers by Knowles and Wichman deal with the marking of prosody in transcriptions of speech; software tools are described in the chapters by Brodda, Hofland, and Nol.)

The large number of corpora analyzed in the papers in *ECC* reflects the rapid expansion in the availability of on-line text collections. The chapter by Faber and Lauridsen describes the careful design principles used to build a parallel Danish–English–French corpus in contract law; Leitner describes the Kolhapur Corpus of Indian English; and Ljung analyzes the distribution of vocabulary in a corpus of English secondary school textbooks used in Sweden (the GYM corpus). Other corpora used for the analyses in this book include: the Brown Corpus, the LOB Corpus, the COBUILD Corpus, the Lancaster/IBM Spoken English Corpus, the Nijmegen Corpus, the Australian Corpus, and the Helsinki Corpus of Modern English Dialects. The chapter by Taylor, Leech, and Fligelstone provides a useful survey of English machine-readable corpora available in 1989. Finally, a survey article by Altenberg presents a bibliography of publications based on English computer corpora, including approximately 650 entries.

As the title states, the papers in *ECC* deal exclusively with English corpora. In addition, apart from the chapter by Ljung (which includes analysis of the 18 million words of text in the COBUILD Main and Reserve Corpora), the papers in this volume do not include analyses of very large corpora.[2] Overall, though, *ECC* offers a good overview of the kinds of linguistic and computational research based on English computer corpora.

[2] The meaning of *large* in the collocation *large corpus* has shifted dramatically over the last several years; in 1989 there were few projects based on corpora larger than 1 million words, while at present several projects are analyzing corpora of 100 million words (or larger).

References

Church, Kenneth W. (1988). "A stochastic parts program and noun phrase parser for unrestricted text." In *Proceedings, Second Conference on Applied Natural Language Processing*. Austin, TX, 136–143.

DeRose, Steven J. (1988). "Grammatical category disambiguation by statistical optimization." *Computational Linguistics*, 14(1), 31–39.

Francis, W. Nelson, and Kučera, Henry (1964). "Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers." Department of Linguistics, Brown University.

Garside, Roger; Leech, Geoffrey; and Sampson, Geoffrey, editors (1987). *The Computational Analysis of English: A Corpus-Based Approach*. Longman.

Leech, Geoffrey (1987). "General introduction." In *The Computational Analysis of English: A Corpus-Based Approach*, edited by Roger Garside, Geoffrey Leech, and Geoffrey Sampson, 1–15. Longman.

Douglas Biber is the author of *Variation across Speech and Writing* (Cambridge University Press, 1988), a corpus-based study of the linguistic characteristics of 23 spoken and written genres. Biber's address is: Department of English, Northern Arizona University, Flagstaff, AZ 86011-6032; e-mail: biber@nauvax.ucc.nau.edu