Abstract

Rita C. Simpson and **John M. Swales** (eds). *Corpus linguistics in North America. Selections from the 1999 symposium.* Ann Arbor: The University of Michigan Press, 2001. 241 pages. ISBN 0-472-09762-8. Abstracted by **Federica Barbieri**, Iowa State University.

Corpus Linguistics in North America includes selected papers from the first national symposium devoted to Corpus Linguistics in the United States, held at the University of Michigan in May 1999. In the Introduction, the editors survey major developments in Corpus Linguistics on both sides of the Atlantic over the past 15 years, and outline the structure of the volume. The volume comprises two sections: Part I focuses on corpus development and tools for accessing existing corpora, and Part II presents examples of current corpus-based linguistic analyses.

Part I includes five chapters, the first three of which report on the status of specific corpus projects currently underway in North America. In the first chapter, Charles Meyer reviews the initial design of the International Corpus of English (ICE), and illustrates the complications involved in compiling a corpus in an international context, with a particular focus on ICE-USA, the American component of the corpus. Meyer further shows how a text retrieval program especially developed to analyse texts annotated with ICE tags – the ICE Corpus Utility Program (ICECUP) – can be used to search for all or part of a syntactically analysed parsed tree, and to perform lexical searches and to generate KWIC concordances. In the second chapter, Christina Powell and Rita Simpson report on the progress of the on-line Web-based search interface developed for the Michigan Corpus of Academic Spoken English (MICASE) by the Humanities Text Initiative unit of the Digital Library Production Service of the University of Michigan. The chapter focuses on the transcription and encoding conventions used to create the computerised transcripts for the on-line version of the corpus, and describes the access, functionality and structure of the Web interface, as well as the challenges involved in the design and programming of the Web search functionality. In the following chapter, Douglas Biber, Randi

Reppen, Victoria Clark and Jenia Walter describe the design and construction of the spoken component of the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL Corpus), a corpus in many ways similar to MICASE, also designed to fill in the gap in medium-sized register-specific corpora of spoken and written academic language. In the fourth chapter, Mark Davies illustrates how to create and use multimillion-word corpora from web-based newspapers and magazines. The chapter includes a description of two large corpora that the author has created for Spanish and Portuguese, practical step-by-step instructions for compiling corpora using on-line newspapers, and a discussion of how large multimillion-word corpora can be used to uncover and trace the spread of new and uncommon syntactic constructions. In the last chapter of Part I, Susan Hockey presents an overview of the functions and operations provided by concordance programs for Corpus Linguistics, highlighting desirable features, as well as pitfalls and 'traps for the unwary' (p 76).

Part II consists of seven chapters on a variety of corpus-based studies, including studies of grammar and usage, studies of discourse features from spoken corpora, and studies on applications of Corpus Linguistics to language teaching. In the opening chapter, Douglas Biber, drawing on research done for the *Longman Grammar of Spoken and Written English*, presents three case studies illustrating unexpected and counterintuitive findings about the use of verbs, revealed by corpus-based investigations. In the second chapter, Hongyin Tao presents an in-depth grammatical and discourse-pragmatic analysis of the verb *remember* in spoken discourse. Tao points out that one of the major strengths of Corpus Linguistics lies in 'its potential to make explicit the more common patterns of language use' (p 116), and argues for an integrated approach to the study of language use combining corpus linguistics methodology and sociocultural linguistic insights.

The following two chapters are based on the MICASE and focus on spoken academic English. John Swales and Bonnie Malczewski look at attention getters and 'new-episode flags' (NEFs), a class of discourse markers used either to move the discourse from monologue to dialogue format (or vice versa) or to announce the beginning of a new segment of the discourse. Anna Mauranen explores discourse reflexivity. Distinguishing reflexive targeted expressions between monologic, dialogic, and interactive, Mauranen shows that spoken academic discourse lacks the internal connectors that typically occur in written academic discourse.

The following two chapters illustrate applications of corpus-based research to language teaching and are explicitly aimed at language teachers. Aaron Lawson compares how the subjunctive and the demonstratives are used in the Barnes

Corpus of Spoken French Language (a corpus of standard continental French) with the ways in which they are presented in a number of recent mainstream university-level textbooks of French as a Second Language. The study reveals overwhelming discrepancies between actual usage and textbook descriptions for both of the features considered. Using a similar methodology, Stephanie Burdine examines expressions of disagreement in three genre-specific corpora of spoken English and compares the findings to the coverage of these expressions in intermediate-level ESL textbooks, revealing, as Lawson, significant incongruities between textbook descriptions of language and real language usage. Burdine outlines a lexically based typological framework of disagreement strategies obtained via concordancing, and proposes the approach as a possible application to teaching. In the last chapter, Randi Reppen reports on the results of a corpus-based study investigating the writing development of elementary school children from two different language backgrounds, English and Navajo. The study involves the use of Biber's multidimensional analysis and shows that the corpus-based approach provides more generalizable information about the developmental changes taking place as students become increasingly literate in writing.