REVIEW


Yamuna Kachru

*University of Illinois at Urbana-Champaign*

ykachru@uiuc.edu

The two volumes reviewed here present two different facets of corpus linguistics. The book edited by Knowles, Wichmann, & Alderson presents studies conducted with, and theoretical issues arising from, the Spoken English Corpus (SEC) that resulted from a collaboration between the Unit for Computer Research on the English Language (UCREL) and IBM UK Scientific Center (UKSC). The McEnery & Wilson volume is intended to be used as a textbook for courses in corpus linguistics and does an excellent job of introducing this field of research to a wide range of audiences. Corpus and computational linguistics, of course, are among the fastest growing subareas of linguistics; it is no wonder a number of introductions to these fields have appeared since the McEnery & Wilson volume, e.g., Kennedy 1998, Biber, Conrad, & Reppen 1998, Lawler & Dry 1998, among others.

The Knowles, Wichmann, & Alderson volume is divided into four sections: Section A: Compiling and Transcribing a Corpus of Speech, Section B: Evaluation, Section C: Analysis, and Section D: Application in Speech Synthesis. The sections are preceded by List of Contributors, Foreword titled The Spoken English Corpus in its Context by Geoffrey Leech, Acknowledgements, Abbreviations and Acronyms, and Availability of Spoken English Corpus. References and Index following the four sections complete the volume.

The foreword by Geoffrey Leech puts the Lancaster/IBM Spoken English Corpus (SEC) in its context. SEC was preceded by a number of machine-readable English text corpora such as the Brown University corpus of written American English, the Lancaster/Oslo-Bergen corpus (LOB) of written British English, and the London-Lund corpus (LLC) of spoken British English. SEC is small in comparison to some of these and the label ‘spoken’ is used in a special sense; it is only 52,637 words long, represents 339 minutes of recording time, and the recordings are of prepared monologues, read out in an accent close to BBC English.
or RP. It is, however unique in that the data are transcribed by trained phoneticians using a prosodic transcription developed for the project, and the records of speech are provided for research in a number of parallel versions. The detailed description of the corpus is available in Knowles, Williams, & Taylor 1996.

The first chapter by Briony Williams, after a preliminary discussion of the need for data-driven, as opposed to theory-driven, research on language, provides more details about the type of data the SEC provides. Two transcribers, Knowles and Williams, transcribed one half of a text, and both transcribed passages from the middle of a text. Thus, about one-fifth of the corpus (nearly 10,000 words) was transcribed by both transcribers. The two transcriptions of these ‘overlap passages’ were compared, but no revisions were made. These overlap passages with their sets of transcriptions are available for research by those who may be interested in questions of agreement between transcribers, individual variation in speech, speech perception, etc. The SEC uses a simpler prosodic transcription system as compared to the LLC. The SEC, in addition to audio recordings, is available in the following forms: orthographic text, grammatically tagged and prosodically transcribed forms (in both machine-readable and book forms). The overlap passages are available in F0 plot form also, which makes it easier to compare the two transcriptions of overlap passages on the one hand, and the transcriptions and audio recordings on the other.

The second chapter by Lita Taylor goes into the details of the compilation of the corpus. SEC consists of the following genres: commentary, news broadcasts, general public lecture, expert lecture to restricted audience, religious broadcast, magazine-style reporting, fiction, poetry, dialogue, propaganda, and miscellaneous (e.g., weather reports, road condition reports, details of programs on the radio). Female speech accounts for 30 percent of the corpus (17 of the 53 texts). The following versions of SEC are available to researchers: spoken recordings, unpunctuated transcription, orthographic punctuated transcription, prosodic transcription, grammatically tagged version, and syntactically parsed version. The paper describes the special symbols used to represent prosodic features, and gives samples of all six versions.

In the third chapter, Briony Williams details the nature of the intonation-transcription system adopted for the project. SEC is a corpus of spoken English; it was felt that the corpus should be presented as speech rather than a set of written texts. Little is known about prosody of texts, therefore, a system of prosodic transcription had to be developed keeping in mind the purpose of the transcription. The chapter presents a comparative assessment of the existing intonation-transcription systems, describes the SEC system, and discusses the potential applications of the system in the area of intonation transcription of natural texts and resynthesis from transcription to perceptually-equivalent intonation of select utterances.

The next two chapters, by Brian Pickering, Briony Williams, and Gerry Knowles (4) and Gerry Knowles (5), go into the details of the prosodic transcription. Chapter 4 compares the transcription by the two transcribers in the overlap
passages and specifies where the disagreements lie. Chapter 5 discusses the problems of transcription from the transcribers' perspective. Both chapters raise the interesting question of whether perceived categories are discrete or not. The transcription reveals that there is considerable fuzziness in deciding whether syllables are stressed, whether the stressed syllables are accented, and whether the accented syllables have level tones. Obviously there is need for more research in these areas of textual prosody.

The subsequent chapters are concerned with analyses of prosodic transcription from four different viewpoints. Chapter 6 examines the distribution of the prosodic marks (i.e., the tonetic stress mark, the boundary mark, and the stress mark) in the overlap passages. Chapter 7 by Lita Taylor traces the correspondence between tone group and punctuation. The punctuators worked on the SEC version of unpunctuated texts and not on the recordings. Therefore, the fact that the punctuation marking correlates largely with the prosody suggests either that there is a high degree of predictability from prosody to punctuation, or that there is some other aspect of text structure available to readers that makes it possible to assign textual prosody and punctuation. In Chapter 8, Gerry Knowles relates grammatical structure to prosodic structure and describes a system whereby it is possible to assign prosodic patterns to natural texts. One complication in predicting prosodic structure on the basis of grammatical structure is the unpredictability of speaker choice as discussed in chapter 8. Another major complication is introduced by style of speaking, which is tackled by Anne Wichmann in Chapter 9. The chapter demonstrates that "there are prosodic features which can be related systematically to context, in particular to the roles of the participants in the interaction and the relationship between them" (187). This leads to the conclusion that prosody can not be studied exclusively as a structural system, and that textual prosody adds another dimension to discourse analysis in general and conversation analysis in particular.

The three chapters in the fourth section, chapters 10-12, describe how the SEC corpus has been used to develop the prosody component of a text-to-speech synthesis system. Chapter 10 by Briony Williams and Peter Alderson describes an experiment in which a model that synthesizes British English intonation patterns from text annotated with TSMs was carried out with impressive results. In Chapter 11 by Brian Pickering a further experiment is described which examines the naturalness and acceptability of the intonation contour generated by the model. In the final chapter, Nick Campbell examines the timing aspect of speech, which, though not as independent as pitch, nevertheless contributes to the rhythm of speech.

The book makes a valuable contribution to the study of textual prosody. As a result of the work reported in the volume, it becomes clear that the interface between phonological structures and prosody presents many challenging issues for research.

The book by McEnery & Wilson has seven chapters following Acknowledgements and List of Abbreviations: 1. Early Corpus Linguistics and the

The first chapter explains why the Chomskyan revolution was skeptical of large-scale corpora capturing the essence of language. It then presents the opposing views that demonstrate how certain kinds of linguistic data are not accessible to introspection alone, e.g., data regarding the frequency of certain grammatical features, and grammaticality judgments based on introspection alone that have been falsified by data attested in corpora. The chapter is followed by suggestions for further reading and study questions.

The same format, chapters followed by suggestions for further reading and study questions, is followed throughout the book, which makes it useful as a course book for corpus linguistics.

The second chapter describes various corpora and explains the difference between corpora and machine-readable texts. It also illustrates various systems of annotation of machine-readable texts that make the task of linguistic researcher easier and less time-consuming. Suggestions for further reading point out the more recent advances in annotations, and study questions offer an opportunity to think critically about issues in annotation of corpora.

The third chapter looks at the relationship between quantitative and qualitative approaches to corpus analysis. In order for any quantitative analysis to be valid, care has to be taken that the data are not skewed in favor of any particular sample of language use. The issue of corpus representativeness thus assumes a great deal of importance. This topic has been discussed in some detail in the chapter and is followed by illustrations of various statistical methods applied to corpora to investigate different phenomena. Further readings guide readers to other sources of information on various methods of statistical treatment of linguistic data, and the study questions set some simple problems for training readers to think in terms of statistics.

The fourth chapter describes how corpora have been used in language study. There are brief accounts of corpora in the following fields of linguistic analysis and fields related to linguistics: speech, lexicon, grammar, semantics, pragmatics, discourse analysis, sociolinguistics, historical linguistics, stylistics, teaching languages and linguistics, dialectology and variation studies, psycholinguistics, cultural studies, and social psychology. The study questions encourage readers to think critically about the potential contribution and pitfalls of reliance on corpora for linguistic research. Further readings direct the readers’ attention to a number of valuable studies published between 1982-1994.

The fifth chapter aims at exploring the relationship between corpora and computational linguistics by providing a brief overview of research in natural language processing. The areas covered are part-of-speech analysis, automated
lexicography, parsing, and machine translation. The study questions are all related to one problem specified in question 1, and several publications are listed in the section on further readings that introduce readers to, among other topics, constraint grammars.

The sixth chapter elucidates the nature of sublanguages — or genre-specific restricted languages — by examining some aspects of a particular sublanguage, that of IBM Manuals, in contrast with two corpora of unconstrained language. The two corpora are the Canadian Hansard corpus, a corpus of one million words of transcribed proceedings from the Canadian Parliament in the mid-1970s, and the American Printing House for the Blind corpus of fictional texts, about three hundred thousand words. The IBM corpus is one million words in size. The case study looks at the following features of the three corpora: lexical closure, part of speech closure, and parsing closure, i.e., the rate at which new words, parts of speech and sentence types occur in the three corpora. As hypothesized, the IBM corpus shows less creativity in language use as compared to the other two corpora. The chapter is followed by two study questions and several recommended readings.

The final chapter asks the inevitable question, what next? While considering the future, the pressures related to the size, scope, international concerns, and role in funding of academia and industry and national agencies for building corpora are all discussed. The needs of specific linguistic research goals and computing are also considered. New opportunities presented by multimedia technologies are of great relevance to research on face-to-face interaction. Similarly, smarter software for analysis could simplify the task of corpus analysts. The book ends on an optimistic note: no matter what, advances in corpus linguistics could only provide richer resources for the linguist.

The glossary and appendices are valuable parts of the book, and make it useful as a resource for teaching an introductory course in corpus linguistics. Appendix A lists corpora mentioned in the book and refers to the International Corpus of English (ICE) without giving much detail. Full details about ICE and studies based on the corpora collected under the project are given in Greenbaum & Nelson 1996.

Both the volumes reviewed here are important additions to the fast growing literature in corpus linguistics, and should be read by anyone interested in utilization of large-scale corpora in linguistic research.

REFERENCES


FORTHCOMING:

DIASPORA, IDENTITY, AND LANGUAGE COMMUNITIES

Edited by Braj B. Kachru and Cecil L. Nelson

(Studies in the Linguistic Sciences 31:1, Spring 2001)

The uniqueness of this thematic issue of SLS lies in its interdisciplinary inclusiveness and the visions of seventeen distinguished international scholars representing a variety of languages, regions, and theoretical and methodological insights.

The volume is divided into four parts: PART I: CULTURE, CANON, AND CREATIVITY (Cameron McCarthy, Edwin Thumboo, Shirley Geok-lin Lim); PART II: CONTEXTUALIZING DIASPORAS (Salikoko S. Mufwene, Enrique (Henry) T. Trueba, Aleya Rouchdy, Elabbas Benmamoun, Michael Palencia-Roth, Robert Baumgardner, Nobuko Adachi); PART III: CONSTRUCTING DISCOURSE IN DIASPORA (Pradeep A. Dhillon, Tamara M. Valentine, Robert D. King, Rajeshwari V. Pandharipande, Marc Deneire, Tej K. Bhatia); PART IV: AFTERWORD (Ladislav Zgusta).

In its style and organization, Diaspora, Identity, and Language Communities is ideal as a reference volume and as a text for courses in, for example, Language Contact and Convergence, Language Change, Diaspora Studies, Sociolinguistics, Language and Culture, Language Shift, Language and Politics, and Sociology of Language.

US $15.00 (plus postage)
   To order, contact:

Studies in the Linguistic Sciences
   at our postal address or
e-mail: deptling@uiuc.edu or
   fax: [USA] 217 333 3466