Exploring Textual Data

Ludvic Lebart, André Salem, and Lisette Barry
(CNRS, Université de la Sorbonne Nouvelle, and L. Barry Associates, Inc.)

Dordrecht: Kluwer Academic
Publishers (Text, speech and language
technology series, edited by Nancy Ide
and Jean Véronis, volume 4), 1998,
xi+246 pp; hardbound, ISBN
0-7923-4840-0, $108.00, £65.00,
Dfl 190.00

Reviewed by
Douglas Biber
Northern Arizona University

This ambitious book introduces readers to the study of texts using statistical methods, focusing especially on multivariate statistics. The book is written to address the concerns of students and scholars, from many different disciplines, who need to analyze texts in their research. The authors specifically identify literary texts, historical texts, scientific texts (e.g., for information retrieval), sociological texts (responses to socioeconomic surveys), and various types of interviews (in marketing, applied psychology, etc.) as types of texts that could be analyzed using these techniques.

The first chapter briefly surveys some of the ways in which texts are approached in various disciplines (linguistics, content analysis, and artificial intelligence) and introduces the reader to the statistical processing of texts. A major part of the chapter is devoted to the processing of "special texts": responses to open questions on surveys. This type of text is especially important in the present book, since most of the examples presented in subsequent chapters are based on corpora of responses to open-ended questions in socioeconomic surveys.

Chapter 2 raises the important question of what unit of analysis to use for textual statistics. It addresses the choice between using orthographic words and lemmas, and briefly introduces some content-based units of analysis. Surprisingly, the chapter discusses only research questions that characterize smaller linguistic units (rather than complete texts); thus, there is no discussion of techniques that use whole texts as the unit of analysis. Most of Chapter 2 deals with vocabulary distributions, including a section on repeated sequences of words. Chapter 2 also briefly introduces concordance displays and tagged corpora.

Chapters 3–8, then, present various multivariate statistical techniques used for text analysis. These include correspondence analysis, cluster analysis, textual time series, and discriminant analysis. While the authors give statistical formulas for many of the procedures, detailed mathematical proofs are given only in the appendices, making the main text accessible to a wider audience. In addition, these chapters contain many specific examples showing how each statistical technique is applied to actual research questions. Chapter 5 is especially noteworthy in this regard, in that it works through examples in detail, showing how correspondence analysis and cluster analysis can be applied to describe associations among words and among demographic groups.

My only major criticism of this book is that the actual presentation of material is narrower than the intended goals and audience presented in the preface and introduction. The book would make an ideal introduction to statistical methods for applied

NLP work in the social sciences. Nearly all examples in the book focus on the analysis of survey responses to questions such as "What is the single most important thing in life for you?" and "What are some examples of what you would consider an 'ideal' meal?" These responses are analyzed to investigate research questions such as whether female and male respondents from different age groups identify different things as important in their life, or whether respondents living in Paris, New York, and Tokyo have different behaviors, habits, and food preferences. These are important research questions for the social sciences, and using computational text analysis combined with multivariate statistical techniques allows kinds of investigation not feasible before.

However, the book is framed as a general introduction to statistical text analysis, and in this regard it is somewhat less successful. Although many research questions in stylistics, discourse analysis, and linguistics could be answered using these techniques, the book itself provides few examples of this type. Thus, for the most part, readers from these other disciplines will need to figure out for themselves how their research questions relate to the kinds of examples discussed in the text, to determine which techniques are appropriate for their applications. Nearly all examples in the book use words or sequences of words as the observations to be analyzed. In contrast, linguists and discourse analysts are often interested in grammatical structures, discourse segments, and even entire texts as observations. For example, linguists often investigate the ways in which grammatical variants are distributed systematically across grammatical and discourse contexts. It seems that this type of research question could be addressed through correspondence analysis, although it is very different from the sociological research questions exemplified in the book (such as the life choices of males versus females). Further, there is no discussion of the relation between these techniques and other techniques such as VARBRUL, which have been used widely by sociolinguists and discourse analysts for decades.

A more minor criticism is that the book is needlessly abstract at some points, especially in the early chapters. For example, a table in Chapter 2 assigns a numeric code to each of the words in a corpus, and these numbers are used in subsequent discussion. We are told that this coding simplifies processing, but there is no indication of how that is the case. This chapter also includes considerable discussion based on an artificial text corpus, where each word is an arbitrary capital letter rather than an actual word. Similarly, we are given the frequencies of some of the most common words in Joyce's *Ulysses* (the 10th, 100th, 1,000th, and 10,000th most common words), but we are not told what the words themselves are. In cases such as these, I would have found the presentation much more accessible if actual words and corpus extracts had been used.

However, despite these criticisms, this book should become an essential reference work for any researcher interested in the quantitative analysis of textual data. I know of no other book that covers the same range of topics with this amount of detail. Although the examples are restricted primarily to a single domain of investigation, researchers from any field who use quantitative techniques to process texts should find much here that will be of use.

Douglas Biber has written many articles and books that use quantitative and corpus-based methods to investigate linguistic issues, including *Variation across Speech and Writing* (Cambridge University Press, 1988), *Dimensions of Register Variation: A Cross-Linguistic Comparison* (Cambridge University Press, 1995), and *Corpus Linguistics: Investigating Language Structure and Use* (with Susan Conrad and Randi Reppen, Cambridge University Press, 1998). Biber's address is: Department of English, Northern Arizona University, Flagstaff, AZ 86011-6032; e-mail: Douglas.Biber@nau.edu.