Briefly Noted

Statistics for Corpus Linguistics

Michael P. Oakes
(University of Lancaster)

Edinburgh: Edinburgh University Press
(Edinburgh Textbooks in Empirical
Linguistics series), 1998, 287 pp; hardbound,
ISBN 0-7486-1032-4, £40.00; paperbound,
ISBN 0-7486-0817-6, £16.95

Corpora are frequently used to investigate authentic language use and a variety of large-sample statistical procedures are currently employed and developed for this purpose. This book provides a largely comprehensive and succinct overview of the most commonly employed statistical procedures with their mathematical foundations and is thus useful for people interested in gaining a deeper understanding of statistical theory and practice in corpus linguistics.

In a clear and precise manner, the author outlines the most common univariate and multivariate statistical procedures and discusses sample case studies to show how these procedures can be useful for corpus linguistics. The book includes chapters on statistical foundations, factor analysis, clustering techniques, and concordancing, as well as two chapters on information theory and literary detective work. The author covers a broad range of statistical techniques while outlining the major research questions and methodologies in corpus linguistics.

The formatting of the book facilitates an understanding of the rather complex material, because key terms are bolded, tables are clearly and consistently laid out, and formulas are set apart from the main text using sufficient space. The key terms are further collected in a glossary that contains concise definitions and is useful as a quick reference guide and refresher. The book also includes tables of important distributions and helpful references to additional readings at the end of each chapter. Moreover, each chapter concludes with a number of exercises and their solutions are listed briefly on one page in the appendix.

There are only a few aspects of the book that could be improved in future editions. First, although the author provides brief syntheses of relevant corpus studies, he often does not provide directions for the application of statistical procedures beyond the context of those studies. Applications of statistical procedures are frequently misunderstood in scholarly research because the researchers are not sufficiently informed about the underlying assumptions of the procedures and the consequences of violating them. It would thus be helpful to readers who are not familiar with a statistical procedure to include tables that briefly summarize its assumptions or lists that summarize its main goals.

Second, it might be helpful to reduce some of the elaborate and detailed descriptions of computational algorithms, because nowadays statistical work typically relies on computers and statistical software that automatically perform the salient computations. It would be beneficial for the novice statistical researcher to be provided with more information about the interpretations of model parameters and the main goals of each procedure and be warned about common misconceptions about them. Finally, given the relative complexity of the concepts in the book, it would be helpful to include some of the computational steps for the exercises in the appendix along with the solutions.

In summary, Oakes provides a comprehensive, principled, and mostly succinct introduction to the use of statistical procedures and their applications to investigations of large language databanks. The book is thus a useful reference guide for both corpus linguists and those who are interested in becoming one, but the reader of this book should be prepared to deal with a large array of dense material that includes numerous tables, formulas, and mathematical symbols.—*André Rupp, Northern Arizona University*

Focus: Linguistic, Cognitive, and Computational Perspectives

Peter Bosch and Rob van der Sandt
(IBM Institute for Logic and Linguistics, Heidelberg and University of Nijmegen)

Cambridge University Press (Studies in
natural language processing), 1999,
xviii+368 pp; hardbound, ISBN
0-521-58305-5, $69.95

"This book presents a collection of writings on the issue of focus in its broadest sense. While commonly being considered as related to phenomena such as presupposition and anaphora, focusing is much more widespread, and it is this pervasiveness that the current collection addresses. This volume brings together theoretical, psychological, and descriptive approaches to focus, at the same time maintaining the overall interest in how these notions apply to the larger problem of evolving some formal representation of the semantic aspects of linguistic content.

"The chapters in this volume have been reworked from a selection of original papers presented at a conference held in 1994 in Schloss Wolfsbrunnen in Germany."—*From the publisher's announcement*

Machine Translation: Theory, Applications, and Evaluation

Nico Weber (editor)
(Fachhochschule Köln)

St Augustin: Gardez! Verlag
(Sprachwissenschaft, Computerlinguistik,
und neue Medien, edited by Nico Weber,
band 1), 1998, 192 pp; paperbound, ISBN
3-928624-71-7, DM 49.90

The contents of the volume are as follows: