This book is about investigating the way people use language in speech and writing. It introduces the corpus-based approach to linguistics, based on analysis of large databases of real language examples stored on computer. Each chapter focuses on a different area of linguistics, including lexicography, grammar, discourse, register variation, language acquisition, and historical linguistics. The chapters follow the same progression of topics, beginning with the identification of important research questions, the steps followed for corpus-based analysis, and finally the interpretation of the findings. Example analyses are presented to provide concrete descriptions of the research methods and advantages of corpus-based techniques.

Ten methodology boxes provide clear and concise explanations of the issues involved in doing corpus-based research and reading corpus-based studies, and there is a useful appendix of resources for corpus-based investigation. This lucid and comprehensive introduction to the subject will be welcomed by a broad range of readers, from undergraduate students to professional researchers.
Corpus linguistics
Investigating language structure and use
Cambridge Approaches to Linguistics

General editor: Jean Aitchison. Rupert Murdoch
Professor of Language and Communication,
University of Oxford

In the past twenty-five years, linguistics – the systematic study
of language – has expanded dramatically. Its findings are now
of interest to psychologists, sociologists, philosophers, anthro-
pologists, teachers, speech therapists and numerous others who
have realized that language is of crucial importance in their life
and work. But when newcomers try to discover more about the
subject, a major problem faces them – the technical and often
narrow nature of much writing about linguistics.

Cambridge Approaches to Linguistics is an attempt to solve
this problem by presenting current findings in a lucid and non-
technical way. Its object is twofold. First, it hopes to outline the
“state of play” in key areas of the subject, concentrating on what
is happening now, rather than on surveying the past. Secondly,
it aims to provide links between branches of linguistics that are
traditionally separate.

The series will give readers an understanding of the multi-
faceted nature of language, and its central position in human
affairs, as well as equipping those who wish to find out more
about linguistics with a basis from which to read some of the
more technical literature in textbooks and journals.

Also in the series

Jean Aitchison: The seeds of speech: language origin and evolution
Charles Barber: The English language: a historical introduction
Jean Aitchison: Language change: progress or decay?

Forthcoming titles include:

Shula Chiat: Understanding children with language problems
Loralne K. Obler and Kris Gjerlow: Language and the brain
Corpus linguistics
Investigating language structure and use

DOUGLAS BIBER
Northern Arizona University

SUSAN CONRAD
Iowa State University

RANDI REPPEN
Northern Arizona University
Contents

Preface xix

1 Introduction: goals and methods of the corpus-based approach 1

Part I Investigating the use of language features 19

2 Lexicography 21
3 Grammar 55
4 Lexico-grammar 84
5 The study of discourse characteristics 106

Part II Investigating the characteristics of varieties 133

6 Register variation and English for Specific Purposes 135
7 Language acquisition and development 172
8 Historical and stylistic investigations 203

Part III Summing up and looking ahead 231

9 Conclusion 233

Part IV Methodology boxes 243

1 Issues in corpus design 246
2 Issues in diachronic corpus design 251
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Concordancing packages versus programming for corpus analysis</td>
<td>254</td>
</tr>
<tr>
<td>4 Characteristics of tagged corpora</td>
<td>257</td>
</tr>
<tr>
<td>5 The process of tagging</td>
<td>261</td>
</tr>
<tr>
<td>6 Norming frequency counts</td>
<td>263</td>
</tr>
<tr>
<td>7 Statistical measures of lexical associations</td>
<td>265</td>
</tr>
<tr>
<td>8 The unit of analysis in corpus-based studies</td>
<td>269</td>
</tr>
<tr>
<td>9 Significance tests and the reporting of statistics</td>
<td>275</td>
</tr>
<tr>
<td>10 Factor loadings and dimension scores</td>
<td>278</td>
</tr>
<tr>
<td>Appendix: commercially available corpora and analytical tools</td>
<td>281</td>
</tr>
<tr>
<td>References</td>
<td>288</td>
</tr>
<tr>
<td>Index</td>
<td>297</td>
</tr>
</tbody>
</table>
Preface

When we set out to write this book, we were faced with many possible goals and emphases. We could have, for instance, presented the history of research done within a corpus-based framework, surveying studies up to the current state-of-the-art. Or we could have concentrated on simple corpus-based studies that readers could conduct on their own. Or we could have focused purely on the methods and analytical techniques required for corpus analysis. Although we give attention to all of these areas, they are not the central focus of this book. Instead, we chose to write a book that reflects the reasons for our own enthusiasm for corpus-based research: this is a book about language as well as corpus linguistics, showing how empirical investigations of corpora can shed new light on previously intractable research questions in linguistics.

In the past several years, on-line corpora and analysis tools have become increasingly accessible, and corpus-based studies have become increasingly common. However, many aspects of corpus-based research remain mysterious for readers, since research articles can rarely afford the space to discuss methodological considerations or analytical procedures in complete detail. In addition, researchers are just now learning how to fully exploit the resources of representative text corpora, and as a result, they are also becoming aware of the many new research questions that can be investigated through corpus-based research. For these reasons, we saw the need for a textbook such as this one: presenting explicit discussions of the methodological issues and techniques used in corpus-based work, with a primary emphasis on
Preface

the interesting investigations of language use possible through this approach.

Much of the research reported here has been generously supported by the National Science Foundation and Addison-Wesley Longman. Work on the ARCHER project, analysing diachronic register variation in English, was funded by NSF Grant BNS-9010893 (Co-PIs: Doug Biber and Ed Finegan; Research Assistants: Dwight Atkinson, Jena Burges, Dennis Burges, Randi Reppen, Ann Beck). Longman ELT Division provided many of the corpora and computers used for the example analyses throughout the book.

In addition, we are grateful to colleagues at Northern Arizona University and Iowa State University for their friendship and stimulating interactions relating to discourse and language use, in particular: Jena Burges, Carol Chapelle, Bill Grabe, John Hagge, Marie Helt, and Susan Wright. We owe a special thanks to the students in many of our MA TESOL and PhD classes over the last two years, who enthusiastically read and commented on many of these chapters in draft form. Finally, we are grateful to the three reviewers who provided detailed, helpful feedback on the draft manuscript of this book.