the single correct approach. In fact, research questions for corpus-based studies often grow out of other kinds of investigations. While it is not our purpose here to describe the full range of linguistic methodologies, it is useful to briefly identify some of the interrelations between corpus-based research and other approaches.

Research questions for corpus-based analyses can be related to other approaches in several ways. First, many questions grow out of prior structural analyses. For example, curiosity about the use of related structures – such as *that*-clauses compared to *to*-clauses – first comes from knowing that such similar structures exist. Research questions may also grow out of a hypothesis or theoretical framework. For example, it has been hypothesized that the lack of time for planning and editing in spontaneous speech makes it impossible for spoken language to be as structurally complex as written language. In Chapter 6 we investigate a research question related to this hypothesis, looking at the frequency and distribution of different kinds of complex structures across spoken and written registers – with surprising results for the hypothesis. Similarly, a theoretical framework might outline the linguistic features that should be acquired by children at different ages, and this could be tested empirically with a corpus-based study.

Intuition and anecdotal evidence can also lead to interesting corpus-based investigations. For example, university students often develop the impression that research in different academic disciplines is written up in very different ways. Chapter 6 presents a corpus-based analysis that investigates this possibility, studying differences across research articles in biology and history. Similarly, some elementary school teachers have felt that their second-language students consistently make many grammatical errors in essays, even after years of using English; Chapter 7 addresses a research question comparing the errors of second-language and native-speaker elementary students. Thus, there are many ways that our experiences with language and previous research in linguistics can motivate corpus-based studies.

Finally, language use can be studied through detailed analyses of specific linguistic features in particular texts, complementing the findings from analyses of large corpora. For example, early investigations that documented the importance of information ordering (e.g., given information before new) were necessarily based on intensive qualitative analysis of single texts. Similarly, micro-analysis of interactions in small segments of conversation, as with conversation analysis, can also provide different perspectives on language use that are not covered in the corpus-based approach. We argue throughout this book that comprehensive analyses of language use require a corpus-based approach, but such investigations are often framed in terms of the constructs and hypotheses resulting from earlier micro-analyses of individual texts.

1.2.5 Areas of linguistics that can be addressed with the corpus-based approach

You may have noticed from the discussion to this point that corpus-based methods can be used to study a wide variety of topics within linguistics. In the last few pages, we have mentioned numerous investigations, focusing on individual words, grammatical features, men's and women's language, children's acquisition of language, author style, register patterns. In fact, another of the strengths of the corpus-based approach is that it can be applied to empirical investigations in almost any area of linguistics.

The core areas of linguistic structure, such as lexicography (the study of words) and grammar, can be studied from a use perspective by applying corpus-based techniques. For example, in lexicography corpus-based techniques enable examination of the linguistic and non-linguistic associations of particular words. In the past, dictionary makers generally limited their task to identifying the possible meanings of a word. Now they can also include information about the most common uses, the frequency of related words, and the contexts in which words and meanings are most commonly found. Grammatical structure can similarly be analysed from a use perspective by applying corpus-based techniques.

A variety of issues in other areas of linguistics can also be addressed with corpus-based studies. For example, within sociolinguistics, corpus-based techniques allow investigations of dialect and register patterns that previously could not be addressed, such as the complex co-occurrence patterns among features in different registers. In the past, language acquisition was typically investigated through detailed case studies that relied on a small number of subjects. Now, as corpora of learners' language are compiled, studies can be based on a large number of learners, and general patterns across learners can be examined.

Studies of style, too, are facilitated by the corpus-based approach. Individual authors or styles across historical periods can be investigated in a more comprehensive way than in the past, examining more texts and more language features. In fact, historical corpora now offer the opportunity to investigate the use of many linguistic features across historical periods, or to examine the development of registers over time.

Corpus-based studies are also applicable to educational linguistics. The results of large-scale studies of use are helpful in designing effective materials and activities for classroom and workplace training, allowing us to help students with the language that is actually used in different target settings. Within educational linguistics, the field of language testing, too, can benefit from results of corpus-based studies, making tests which conform to the actual language that students will be using on a regular basis.

In sum, almost any area of linguistics can be studied from a use perspective – and the corpus-based approach provides a suite of tools and methods that are particularly effective for such investigations.

1.3 Corpora and corpus analysis tools used in this book

1.3.1 Corpora

As noted above, a corpus is a large and principled collection of natural texts. In subsequent chapters of this book you will be introduced to some of the important issues relating to size and representativeness in corpus design (see especially Methodology Boxes 1 and 2). For some of the example analyses in the book, we have used corpora especially designed to address specific research questions. However, many of the example analyses use well-known corpora which are publicly available. There are four corpora (or parts of corpora) that we have used repeatedly in this book: the London–Lund Corpus, the Lancaster–Oslo/Bergen (LOB) Corpus, the conversation register from the British National Corpus (BNC), and certain registers from the Longman–Lancaster Corpus. Each of these is described in Table 1.2 and briefly reviewed below. Further information about them, including how to obtain them, is provided in the appendix.

Two of these corpora represent spoken registers. The first of these is the London–Lund Corpus, which contains several different kinds of spontaneous and prepared speech. Together, the texts total approximately 500,000 words. For analyses requiring larger databases of spoken language, we use part of the British National Corpus (BNC). These texts total approximately 4,000,000 words of conversation. (The entire BNC includes c. 100 million words of text, with 90 percent of the corpus from written registers and 10 percent from spoken registers.)

The other two corpora contain only written texts. The LOB Corpus consists of 15 registers, as listed in Table 1.2. Together, the texts in this corpus total approximately 1,000,000 words, with individual registers ranging from 12,000 to 160,000 words. When analyses require larger databases of written prose, we use two registers from the Longman–Lancaster Corpus: about 3,000,000 words each of academic prose (including books and articles from a variety of disciplines) and fiction (including novels and short stories).

All texts in the London–Lund, BNC, and LOB corpora are from British English. The academic prose and fiction texts in the Longman–Lancaster Corpus come from authors of both British and American English. Given the scope of the present book, and the large number of linguistic investigations already included, we have chosen to disregard national dialect differences here. However, the techniques introduced in this book could also be used to investigate differences across national dialects, and we would expect such a study to uncover interesting patterns of variation.

1.3.2 Analysis tools

Two kinds of tools are used for the analyses described in this book: commercially available packages and computer programs that we developed for specific analyses. Publicly available software

Table 1.2 *Corpora used in analyses throughout the book*

LONDON–LUND CORPUS

Category	Number of texts	Approx. number of words
Face-to-face conversations or discussions	65	235,000
Telephone conversation	110	60,000
Public conversations, discussions, interviews	20	85,000
Spontaneous commentary (radio broadcasts)	20	55,000
Spontaneous oration	12	30,000
Prepared oration	12	35,000
TOTAL	239	500,000

BRITISH NATIONAL CORPUS **(BNC)** – conversation only

Category	Number of texts	Approx. number of words
Face-to-face conversation	160	4,000,000

LANCASTER–OSLO/BERGEN **(LOB)** CORPUS

Category	Number of texts	Approx. number of words
Press reportage	44	88,000
Editorials	27	54,000
Press reviews	17	34,000
Religion	17	34,000
Skills and hobbies	36	72,000
Popular lore	48	96,000
Biographies and essays	75	150,000
Official documents	30	60,000
Academic prose	80	160,000
General fiction	29	58,000
Mystery fiction	24	48,000
Science fiction	6	12,000
Adventure fiction	29	58,000
Romantic fiction	29	58,000
Humor	9	18,000
TOTAL	500	1,000,000

Table 1.2 *Cont'd*

LONGMAN–LANCASTER CORPUS – two categories only

Category	Number of texts	Approx. number of words
Academic prose	98	2,700,000
Fiction	144	3,000,000
TOTAL	242	5,700,000

packages are referred to as "concordancing" programs. These programs allow the user to search for specific target words in a corpus, providing exhaustive lists for the occurrences of the word in context. They thus enable the analysis of lexical collocations (i.e., lexical–lexical association patterns), and also provide frequency information. Concordancing programs have been available for many years, and new ones join the market each year. Some of these, such as TACT and Lexa, are available at a very small or no cost. In the early chapters of the book, we illustrate analyses using concordancing software – the kinds of analyses that everyone can do once they are familiar with a concordancing program and have a corpus to work with. More information about commercially available software packages is included in the appendix.

However, many interesting research questions involve investigating complex grammatical constructions or complex association patterns. Concordancing programs are not made for these sorts of investigations. For example, it is impossible with a concordancing program to conduct a thorough investigation of when *that* is omitted from *that*-clauses; and it would be even more difficult to look at the complex co-occurrence patterns of linguistic features in different registers. Instead, these investigations require computer programming skills (see Methodology Box 3 for a more detailed discussion of these issues).

Several sample analyses in this book were done with computer programs written by the authors. These programs are explained conceptually so that you can understand how the analyses were conducted. However, you will not need any familiarity with the details of computer programming to use this book. Our purpose here is to introduce you to the kinds of linguistic issues that can be investigated using corpus-based analysis. Other textbooks are designed specifically to teach programming techniques, and we encourage readers interested in pursuing corpus-based research on their own to take additional courses in computer programming.

1.4 Overview of the book

All of the chapters in this book are developed around example analyses. The introduction to each chapter states the research questions that will be addressed, so that you know exactly what aspects of language use are under investigation and why they are important. The discussion for each example analysis then includes the methodology, results, and interpretation of the results. The sample analyses are used to teach many aspects of corpus-based analysis. You will learn about the new kinds of research questions that can be asked and the new findings uncovered from corpus-based studies. In addition, you will learn about the kinds of analytical procedures needed to address these questions and the kinds of decisions that researchers make during corpus-based analyses.

Although each of the chapters follows this general outline, they address language issues with a slightly different focus, as described in the next section.

1.4.1 Chapter overview

The book is divided into two major parts. This division corresponds to the two major types of research question presented in Table 1.1: Part I: "Investigating the use of language features," Part II: "Investigating the characteristics of varieties."

Part I of the book deals with research questions that seek to characterize the use of individual linguistic features. Chapter 2 focuses on individual words, and Chapter 3 on grammatical constructions. Chapter 4 then looks at associations between lexical items and grammatical constructions. Finally, Chapter 5 considers the analysis of discourse structure – for example, the use of nouns and pronouns for reference in texts, or the distribution of constructions such as active and passive voice over the course of a text.

Part II of the book then deals with research questions that seek to characterize texts and varieties of language. Chapter 6 focuses on register characterizations and investigations applicable to English for Specific Purposes. Chapter 7 addresses language acquisition issues – for native-speaker children and second-language learners. Finally, Chapter 8 addresses questions related to the historical development of language use and individual author style.

In Part III of the book, we review the contributions of the corpus-based approach to linguistic investigations and briefly describe additional kinds of research questions that can be investigated with this approach.

Part IV of the book contains ten "methodology boxes." These boxes address important methodological issues that arise when conducting corpus-based studies – issues such as designing representative corpora, norming frequency counts, using grammatically tagged or parsed corpora, and using certain statistical techniques. The information in each box is applicable to the sample analyses in several chapters, so you might want to refer to them at several different points. The boxes are meant to act as introductions to important material for understanding corpus-based studies and for conducting them in a principled manner. However, these introductions should not be regarded as procedural manuals; you will need to learn more about these areas as you begin to carry out more advanced corpus-based research on your own.

Finally, the book also includes an appendix that lists publicly available corpora and analysis tools, with brief descriptions and addresses for obtaining them.

1.4.2 What this book is and what this book is not

As you can probably tell from this introduction, the field of corpus linguistics covers a great deal of ground, and in writing this book we have had to make many difficult decisions about where to focus attention. Primarily this is a book about an approach to linguistics: it shows how corpus-based analysis can address new kinds of research questions and reveal new information about language use that researchers could not uncover using traditional approaches. We have sought to give a wide sampling of the kinds of investigations and types of analyses that are possible with the corpus-based approach.

In introducing you to the corpus-based approach, we have also sought to be very concrete. Each chapter presents several specific example analyses. The steps for conducting the analyses and reaching the interpretations are clearly described, and for many research questions (especially in the early chapters), you should be able to carry out similar investigations with a concordancing package and a corpus. However, we have not attempted to write a thorough how-to manual for corpus-based analysis. We do not, for example, give you step-by-step instructions for using a concordancing program, and we make no attempt to teach computer programming skills. For those areas, we refer you to software manuals and courses in programming for linguistics.

The value and diversity of corpus-based work is also apparent from a large number of studies already published. Rather than providing a review of this body of literature, however, we have chosen to give you a more dynamic view of the corpus-based approach by presenting sample analyses. For those readers wanting to learn more about previous corpus-based studies, each chapter provides a brief list of further readings to get you started.

Overall, this book will show you the wide range of language issues that can be addressed using the corpus-based approach. It will develop your skills as an educated and critical reader of corpus-based studies. We hope that it will also inspire you in developing your own research questions about language use and in conducting your own corpus-based investigations to answer those questions.