1 Introduction
Goals and methods of the corpus-based approach

1.1 Studying language: structure and use

Studies of language can be divided into two main areas: studies of structure and studies of use. Traditionally, linguistic analyses have emphasized structure – identifying the structural units and classes of a language (e.g., morphemes, words, phrases, grammatical classes) and describing how smaller units can be combined to form larger grammatical units (e.g., how words can be combined to form phrases, phrases can be combined to form clauses, etc.).

A different perspective – which is the focus of this book – is to emphasize language use. From this perspective, we can investigate how speakers and writers exploit the resources of their language. Rather than looking at what is theoretically possible in a language, we study the actual language used in naturally occurring texts.

Many studies of language use focus on a particular linguistic structure, investigating the ways in which seemingly similar structures occur in different contexts and serve different functions. For example, in English that-verb-complement clauses and to-verb-complement clauses are similar in their structural characteristics and can be similar in meaning, as in sentences such as:

(1) I hope that I can go.
(2) I hope to go.

In addition, that-clauses can occur with the that omitted:

(3) I hope I can go.
A structural analysis would describe the grammatical similarities and differences among these three sentences. All three options are equally grammatical ways to complete the meaning of the verb. However, an analysis of language use goes beyond traditional grammatical description to ask why the language should have multiple structures that are so similar in their meaning and grammatical function.

Answers to this question should consider a range of factors. For example, do spoken varieties versus written varieties have different preferences for one of the forms over others? Are the forms usually used with different verbs? Are the forms used preferentially for different specialized meanings? These are some of the kinds of questions that can be addressed in studies of use. In fact, in the illustrative analyses presented in Chapters 3 and 4, you will see that there are strong patterns in the preferred use of these different structures.

In addition to analyzing the language use patterns for a linguistic structure, studies of use can focus on the language of a text or a group of speakers/writers. For example, interest in an individual author's style or in the language used by different social groups has been a common motivation for studies of use, considering questions such as: How does the language used by a particular author compare to the language used by his contemporaries? How does the language used by women differ from the language used by men?

Equally important are investigations comparing the language of different texts or groups of texts. Many times each day we use different varieties of language as we participate in different situations – from talking to a family member, to reading a newspaper, to writing a letter to a friend, to reading an academic article. The varieties of language that we use in different situations are referred to as registers, and describing the characteristics of these registers is an important area of study. However, it is also a complex one because many different grammatical and lexical choices come into play. How can we find the patterns in the language used in conversation, newspapers, academic prose, personal letters, etc.? How can we characterize the language used in these different varieties? Questions such as these are also an important aspect of studies of use, and are addressed in Part II of this book.

For all these studies of use, analysts attempt to uncover typical patterns rather than making judgements of grammaticality. There are two central research goals in such analyses of use: (1) assessing the extent to which a pattern is found, and (2) analyzing the contextual factors that influence variability. For example, in an analysis of that-complement clauses versus to-clauses, we would want to consider whether speakers have a preference for one kind of clause and writers have a preference for the other kind of clause. Further, we would want to consider a range of contextual factors, such as the typical verbs that each clause type is used with.

Finding patterns of use and analyzing contextual factors can present difficult methodological challenges. Because we are looking for typical patterns, analyses cannot rely on intuitions or anecdotal evidence. In many cases, humans tend to notice unusual occurrences more than typical occurrences, and therefore conclusions based on intuition can be unreliable. Furthermore, we need to analyze a large amount of language collected from many speakers, to make sure that we are not basing conclusions on a few speakers' idiosyncrasies. However, with a large amount of language, it is time-consuming to carry out the analyses and difficult to keep track of multiple contextual factors. If you wanted to compare the language used in conversation and academic articles, for example, imagine how difficult it would be to keep track of even twenty different linguistic structures in ten texts from each register – let alone figure out the ways that these structures are interrelated with a range of contextual factors.

Because of these difficulties, until recently many investigations of language use were either unfeasible or simply impossible. The corpus-based approach, however, provides a means of handling large amounts of language and keeping track of many contextual factors at the same time. It therefore has opened the way to a multitude of new investigations of language use.

1.2 What is the corpus-based approach?

At this point, you might be wondering: what actually is the corpus-based approach, and what makes it different from other analytical approaches in linguistics? The following sections address
these questions. We begin by identifying the essential characteristics of corpus-based analyses in Section 1.2.1. Underlying these characteristics is a new perspective on language use: studying the use of language characteristics by considering the relevant “association patterns.” This notion, which forms the basis for all subsequent analyses in the book, is introduced in Section 1.2.2. Association patterns represent quantitative relations, measuring the extent to which features and variants are associated with contextual factors. However, functional (qualitative) interpretation is also an essential step in any corpus-based analysis, and so we discuss the relationship between quantitative and qualitative techniques in Section 1.2.3. Finally, in Sections 1.2.4 and 1.2.5, we turn to a comparison of the corpus-based approach with other analytical approaches in linguistics, and summarize the research areas that can be studied using this approach.

1.2.1 The characteristics of corpus-based analyses

The essential characteristics of corpus-based analysis are:

- It is empirical, analyzing the actual patterns of use in natural texts;
- It utilizes a large and principled collection of natural texts, known as a “corpus,” as the basis for analysis;
- It makes extensive use of computers for analysis, using both automatic and interactive techniques;
- It depends on both quantitative and qualitative analytical techniques.

Taken together, these characteristics result in a scope and reliability of analysis not otherwise possible. Several of the advantages of the corpus-based approach come from the use of computers. Computers make it possible to identify and analyze complex patterns of language use, allowing the storage and analysis of a larger database of natural language than could be dealt with by hand. Furthermore, computers provide consistent, reliable analyses—they don’t change their mind or become tired during an analysis. Computers can also be used interactively, allowing the human analyst to make difficult linguistic judgements while the computer takes care of record-keeping.

Goals and methods of the corpus-based approach

Finally, it is important to note that corpus-based analyses must go beyond simple counts of linguistic features. That is, it is essential to include qualitative, functional interpretations of quantitative patterns. In each chapter of this book, you will find that a great deal of space is devoted to explanation, exemplification, and interpretation of the patterns found in quantitative analyses. The goal of corpus-based investigations is not simply to report quantitative findings, but to explore the importance of these findings for learning about the patterns of language use.

1.2.2 Association patterns in language use

Many early studies in corpus linguistics simply counted the occurrence of linguistic items. For instance, some lexical studies compared the frequency of particular words, or of two-letter, three-letter, and four-letter words. Some grammatical studies counted the frequency of nouns, verbs, and adjectives. Studies of this type can be useful in providing reference materials (such as identifying the fifty most common words) or for providing simple stylistic indicators (such as the relative frequencies of nouns and verbs in a text).

However, a representative corpus, if properly exploited, can provide many additional kinds of information about language use. In particular, a corpus-based approach allows researchers to identify and analyze complex “association patterns”: the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features.

There are two main kinds of research question that can be investigated in terms of association patterns, as shown in Table 1.1. The first is to focus on the use of a linguistic feature—either a lexical item or a grammatical construction (Part A of Table 1.1); the second is to focus on the characteristics of texts or varieties (Part B of Table 1.1).

Linguistic analyses have traditionally focused on a particular linguistic feature, either a word or grammatical construction. However, the use of such features can be further investigated by considering their systematic associations with other features. Two main kinds of associations are important here: linguistic associations and non-linguistic associations.
Table 1.1 Association patterns in language use

A. Investigating the use of a linguistic feature (lexical or grammatical)
   (i) Linguistic associations of the feature
       - lexical associations (associations with particular words)
       - grammatical associations (associations with particular grammatical constructions)
   (ii) Non-linguistic associations of the feature
       - distribution across registers
       - distribution across dialects
       - distribution across time periods

B. Investigating varieties or texts (e.g., registers, dialects, historical periods)
   (i) Linguistic association patterns
       - individual linguistic features or classes of features
       - co-occurrence patterns of linguistic features

Linguistic associations fall into two major categories:

1. lexical associations – investigating how the linguistic feature is systematically associated with particular words;
2. grammatical associations – investigating how the linguistic feature is systematically associated with grammatical features in the immediate context.

Lexical associations are illustrated in Chapter 2 through an analysis of the words big, large, and great. Specifically, that analysis considers the collocates of these three words – that is, the words that tend to co-occur with each target word. For example, big commonly co-occurs with toe, while large commonly co-occurs with number. Although these three words are nearly synonymous in isolation, the analysis shows that they tend to be used with very different kinds of words. Thus, this analysis looks at “lexical-lexical” association patterns and finds them to be quite different for each of these three words.

In Chapter 4, on the other hand, we investigate “lexical-grammatical” associations. For example, we compare the nearly synonymous adjectives small and little, showing how they have very different grammatical associations with attributive versus predicative positions (e.g., the small boy versus the boy is small). The opposite type of research question – focusing on a grammatical feature and considering its lexical associations – is also exemplified in the book. For instance, in Chapter 4 we compare the verbs that are most commonly used with that-clauses versus to-clauses (such as think co-occurring with that-clauses versus want commonly co-occurring with to-clauses).

In addition to its linguistic associations, the use of a linguistic feature can be studied in terms of its non-linguistic associations. Three major factors are relevant here: how a lexical item or grammatical construction is distributed differentially across 1. varieties defined by situation (registers), 2. varieties defined by social group (dialects), or 3. periods of time. For example, Chapter 3 includes an investigation of how nominalizations are distributed differently across academic prose and conversation – an example of the association between a grammatical feature (nominalizations) and non-linguistic feature (register).

It is important to realize that linguistic and non-linguistic association patterns are not independent. Rather, they interact. Thus, most sample analyses in this book include both kinds of association patterns. For example, when we consider lexical-lexical associations for big, large, and great, we also consider their distributions across different registers.

Instead of focusing on particular linguistic features, it is also possible to describe the characteristics of texts or varieties in terms of association patterns (Part B of Table 1.1). In this case, corpus-based studies attempt to characterize registers, dialects, styles, or individual literary works in terms of their linguistic association patterns. These linguistic associations can be either individual features or classes of features. In Chapter 6, for example, we characterize different spoken and written registers with respect to their use of dependent clauses.

However, to characterize varieties more thoroughly, another kind of linguistic association pattern is important: the ways in which groups of linguistic features commonly co-occur in texts. For example, nouns, prepositions, long words, and attributive adjectives tend to co-occur in certain registers. Why is this so? What function do these features share? What other features tend
to occur in texts when these features are rare? These and related questions are addressed in the second half of this book.

Though many different kinds of association patterns can be investigated with corpus-based studies, all of these patterns share an important characteristic: they represent continuous relationships. That is, the patterns are not absolute statements about what always happens or never happens in language use; rather, these patterns occur to differing extents. We might think of certain patterns as very common or very rare— but what does “common” or “rare” signify? Making comparisons between association patterns requires a more precise characterization of the extent to which different patterns exist—that is, quantitative measures. The next section discusses the use of quantitative analyses in corpus-based research, as well as the complementary role of qualitative, functional interpretation.

1.2.3 The role of quantitative analyses and functional interpretations

In the last section we reviewed different kinds of association patterns that can be investigated in corpus-based studies and noted that these relationships are continuous constructs. Therefore, quantitative techniques are essential for corpus-based studies. For example, if you wanted to compare the language use patterns for the words big and large, you would need to know how many times each word occurs in the corpus, how many different words commonly co-occur with each of these adjectives (the collocations), and how common each of those collocations is. These are all quantitative measurements.

In all the sample analyses in this book, you will find quantitative analyses. For many of the examples, particularly in the early chapters of the book, we present only frequency data—how often a certain pattern occurs relative to other patterns. In many cases strong patterns can be observed directly from these frequencies, and in order to keep examples straightforward and accessible to a wide audience, we do not use statistical procedures in these cases.

In some of the examples in later chapters of the book, however, statistical procedures are important for investigating complex association patterns. For some analyses, tests of statistical “significance” are also included. Significance tests show how likely it is that quantitative results could have occurred by chance, and thus they should always be reported in research articles describing a corpus-based study. However, it is not our purpose here to teach you how to carry out statistical tests. We do provide conceptual introductions to the statistical procedures used in our example analyses, and we provide some methodological details in the methodology boxes included in Part IV. The discussion in this book will allow you to understand the purpose and importance of the statistical procedures that we have used, but we have not set out to comprehensively discuss statistical techniques. You should consult a statistics textbook for more complete coverage of that kind.

In addition, as you read the sample analyses in this book you will find much more than quantitative and statistical findings. As noted above, a crucial part of the corpus-based approach is going beyond the quantitative patterns to propose functional interpretations explaining why the patterns exist. As a result, a large amount of effort in corpus-based studies is devoted to explaining and exemplifying quantitative patterns. In a textbook this size it is not possible to provide a full functional interpretation of every analysis. However, we do consistently outline the major aspects of such interpretations, emphasizing the importance of this step in all corpus-based analyses.

1.2.4 The corpus-based approach compared to other approaches in linguistics

So far in this introduction, we have been emphasizing the distinctive features of the corpus-based approach. In particular, we have emphasized its strengths in investigating language use, as opposed to traditional studies of language structure. We have noted that comprehensive studies of use cannot rely on intuition, anecdotal evidence, or small samples; they rather require empirical analysis of large databases of authentic texts, as in the corpus-based approach.

However, corpus-based analysis should be seen as a complementary approach to more traditional approaches, rather than as
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, microanalysis 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

Goals and methods of the corpus-based approach

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 30 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

<table>
<thead>
<tr>
<th>LONDON–LUND CORPUS</th>
<th>Number of texts</th>
<th>Approx. number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face-to-face conversations or discussions</td>
<td>65</td>
<td>235,000</td>
</tr>
<tr>
<td>Telephone conversation</td>
<td>110</td>
<td>60,000</td>
</tr>
<tr>
<td>Public conversations, discussions, interviews</td>
<td>20</td>
<td>85,000</td>
</tr>
<tr>
<td>Spontaneous commentary (radio broadcasts)</td>
<td>20</td>
<td>55,000</td>
</tr>
<tr>
<td>Spontaneous oration</td>
<td>12</td>
<td>30,000</td>
</tr>
<tr>
<td>Prepared oration</td>
<td>12</td>
<td>35,000</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>239</strong></td>
<td><strong>500,000</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BRITISH NATIONAL CORPUS (BNC) – conversation only</th>
<th>Number of texts</th>
<th>Approx. number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face-to-face conversation</td>
<td>160</td>
<td>4,000,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LANCASTER–OSLO/BERGEN (LOB) CORPUS</th>
<th>Number of texts</th>
<th>Approx. number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Press reportage</td>
<td>44</td>
<td>88,000</td>
</tr>
<tr>
<td>Editorials</td>
<td>27</td>
<td>54,000</td>
</tr>
<tr>
<td>Press reviews</td>
<td>17</td>
<td>34,000</td>
</tr>
<tr>
<td>Religion</td>
<td>17</td>
<td>34,000</td>
</tr>
<tr>
<td>Skills and hobbies</td>
<td>36</td>
<td>72,000</td>
</tr>
<tr>
<td>Popular lore</td>
<td>48</td>
<td>96,000</td>
</tr>
<tr>
<td>Biographies and essays</td>
<td>75</td>
<td>150,000</td>
</tr>
<tr>
<td>Official documents</td>
<td>30</td>
<td>60,000</td>
</tr>
<tr>
<td>Academic prose</td>
<td>80</td>
<td>160,000</td>
</tr>
<tr>
<td>General fiction</td>
<td>29</td>
<td>58,000</td>
</tr>
<tr>
<td>Mystery fiction</td>
<td>24</td>
<td>48,000</td>
</tr>
<tr>
<td>Science fiction</td>
<td>6</td>
<td>12,000</td>
</tr>
<tr>
<td>Adventure fiction</td>
<td>29</td>
<td>58,000</td>
</tr>
</tbody>
</table>

---

Table 1.2 Cont’d

<table>
<thead>
<tr>
<th>LANCASTER–OSLO/BERGEN (LOB) CORPUS (cont.)</th>
<th>Number of texts</th>
<th>Approx. number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romantic fiction</td>
<td>29</td>
<td>58,000</td>
</tr>
<tr>
<td>Humor</td>
<td>9</td>
<td>18,000</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>500</strong></td>
<td><strong>1,000,000</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LONGMAN–LANCASTER CORPUS – two categories only</th>
<th>Number of texts</th>
<th>Approx. number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic prose</td>
<td>98</td>
<td>2,700,000</td>
</tr>
<tr>
<td>Fiction</td>
<td>144</td>
<td>3,000,000</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>242</strong></td>
<td><strong>5,700,000</strong></td>
</tr>
</tbody>
</table>

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
Methodology Box 6
Norming frequency counts

When corpus-based studies examine the frequency of features across texts and registers, it is important to make sure that the counts are comparable. In particular, if the texts in a corpus are not all the same length, then frequency counts from those texts are not directly comparable. For example, imagine that you analyzed two texts and found that each one has 20 modal verbs. It might be tempting to conclude that modals are equally common in the texts. However, further imagine that the first text has a total length of 750 words, and the second text is 1,200 words long — in this case, your conclusion would be wrong. Because the second text is longer, there are more opportunities for modals to occur, and therefore simply comparing the raw counts does not give an accurate account of the relative frequencies of modals in the two texts.

"Normalization" is a way to adjust raw frequency counts from texts of different lengths so that they can be compared accurately. The total number of words in each text must be taken into consideration when norming frequency counts. Specifically, the raw frequency count should be divided by the number of words in the text, and then multiplied by whatever basis is chosen for norming. To continue with the example above, the counts in the two texts could be normed to a basis per 1,000 words of text as follows:

\[
\text{Text 1} \\
(20 \text{ modals} / 750 \text{ words}) \times 1,000 = 27.5 \text{ modals per 1,000 words}
\]

\[
\text{Text 2} \\
(20 \text{ modals} / 1,200 \text{ words}) \times 1,000 = 16.7 \text{ modals per 1,000 words}
\]
You can see from these normed counts that the raw frequencies are very misleading in this case—that is, modal verbs are actually considerably more common in Text 1 than in Text 2.

In the above example, counts were normed to a basis of 1,000 words since both texts were about this long. In a corpus with shorter texts, counts might be normed to their frequencies per 500 words of text. When working with very short texts, such as the writing of children, it might even be necessary to norm counts to their frequencies per 100 words of text. If a higher basis is adopted, then the counts for rare features can be artificially inflated—sometimes dramatically so. For example, if a student text of 80 words happened to have one passive construction (a generally rare feature in elementary student writing), and the texts were being normed to a basis of 100 words, the text would have a normed score of 1.25 passives per 100 words. However, if that same count was normed to a basis of 1,000 words, the normed value would be 12.5 passives per 1,000 words, which represents a frequency unlikely to be achieved in any extended elementary student writing. Thus, frequency counts should be normed to the typical text length in a corpus.

Methodology Box 7
Statistical measures of lexical associations

The simplest way to identify collocate pairs is by their relative frequency—that is, by how commonly one pair, such as “large number,” occurs relative to another pair, such as “large man.” Such frequency information can give a sense of the most common collocational associations.

However, frequency information alone may present a biased measure of the strength of associations between words. More common words are more likely to occur in a collocate pair simply by chance. Therefore, an alternative way to judge the strength of associations between words is to use measures that account for the likelihood of words occurring together by chance—i.e., statistical measures.

Statistical measures can be used to analyze both the associations between collocate pairs and the differences between the locations of particular words. Below we briefly review a common statistical test for each of these purposes. We explain the principles behind the tests, but for details about the statistical formulas, you should consult the articles listed under “Further reading.”

Mutual information score

The mutual information score or mutual information index gives a measure of the strength of association between two words. It focuses on the likelihood of two words appearing together within a particular span of words (the span is specified for the analysis, e.g., adjacent words, a window of three words, etc.).