Table of Contents

Foreword ix
Acknowledgments xiii

1 Looking at language in use—some preliminaries 1
   1.1 Introduction 1
      1.1.1 Thinking about goalless, shall and cars 1
      1.1.2 Clues from a corpus—the BNC 4
   1.2 Why read this book? 10
   1.3 Organization of the book 11
   1.4 How to use this book 12

2 Corpus linguistics: some basic principles 13
   2.1 Outline 13
   2.2 Introduction 13
   2.3 Representativeness in corpora 15
   2.4 What is corpus linguistics? Why use a corpus? 18
   2.5 A brief—and more advanced—excursion: description vs. theory 20
   2.6 Types of corpora 24
   2.7 Further reading 26

3 Introducing the British National Corpus 27
   3.1 Outline 27
   3.2 Introduction 27
   3.3 Written material 28
   3.4 Spoken material 32
   3.5 More than text 38
      3.5.1 Part-of-speech tags 38
      3.5.2 Headwords and lemmas 40
      3.5.3 Words & sentences versus w-units & s-units 41
   3.6 Format 42
   3.7 Errors 43
   3.8 More information 45
   3.9 Is it Present-day English? 45
   3.10 Exercise 46
4 First queries with BNCweb

4.1 Outline 47
4.2 Introduction 47
4.3 Getting started: your first query 48
  4.3.1 Planning your query 48
  4.3.2 Running the query 49
  4.3.3 Getting basic frequency information 50
4.4 Exploring the concordance 50
  4.4.1 Navigating through a query result 51
  4.4.2 KWIC view and Sentence view 52
  4.4.3 Random order and corpus order 53
  4.4.4 Viewing the larger context of an example 54
  4.4.5 Obtaining more information about the source of an example 57
4.5 Running a query for a word sequence 58
4.6 Restricting your query to selected portions of the BNC 59
4.7 Accessing previous queries 64
  4.7.1 Query history 64
  4.7.2 Save current set of hits 65
4.8 Browse a text 66
4.9 Exercises 68

5 Some further aspects of corpus-linguistic methodology

5.1 Outline 69
5.2 Introduction 69
5.3 Comparing results: normalized frequencies 69
5.4 Normalized frequencies—some further issues 73
5.5 Precision and recall 77
5.6 Statistical significance 79
  5.6.1 Confidence intervals 80
  5.6.2 Hypothesis tests for frequency comparison 83
  5.6.3 Using statistical software 86
5.7 Further reading 90
5.8 Exercises 90

6 The Simple Query Syntax

6.1 Outline 93
6.2 Introduction 93
6.3 Basic queries: searching words and phrases 94
6.4 Using wildcards 97
6.5 A short tour of the Simple Query Syntax 99
6.6 Advanced wildcard queries 103
6.7 Queries based on part-of-speech and headword/lemma 106
6.8 Matching lexico-grammatical patterns 109
6.9 Proximity queries 114
6.10 Matching special characters 116
6.11 Exercises 117

7 Automated analyses of concordance lines—Part I: Distribution and Sorting 119
7.1 Outline 119
7.2 Distribution 119
  7.2.1 A lovely example: distributional facts about the users of lovely 119
  7.2.2 Frequency distribution by genre 124
  7.2.3 Dispersion & File-frequency extremes: checking the influence of idiosyncratic texts on frequencies 128
7.3 Sort 131
  7.3.1 Sorting a query result on preceding or following context 131
  7.3.2 The Frequency breakdown function 133
  7.3.3 Sorting on the query hit 136
7.4 Exercises 137

8 Automated analyses of concordance lines—Part II: Collocations 139
8.1 Outline 139
8.2 Introduction 139
8.3 Understanding the concept of collocational strength 140
8.4 Steps in collocation analysis 142
8.5 Which association measure should I use? 149
8.6 Calculating collocations in sub-sections of the BNC 158
8.7 Further reading 159
8.8 Exercises 159

9 "Adding value" to a concordance using customized annotations 161
9.1 Outline 161
9.2 Introduction: why annotate your concordance data? 161
9.3 Annotation within BNCweb: using the "Categorize hits" function 162
  9.3.1 Setting up a category for analysis 163
9.3.2 Categorizing concordance hits 165
9.3.3 Analyzing data categorized in \textit{BNCweb} 167
9.3.4 Re-editing your annotations 169
9.3.5 Advantages and disadvantages of categorizing queries within \textit{BNCweb} 169
9.4 Summarizing and presenting results of customized annotations 170
9.5 Exporting a \textit{BNCweb} query result to an external database 174
  9.5.1 Downloading from \textit{BNCweb} 175
  9.5.2 Importing into database software 177
  9.5.3 Annotating the database 179
  9.5.4 Analyzing the database 180
  9.5.5 Advantages and disadvantages of the database approach 181
9.6 Reimporting an analyzed database into \textit{BNCweb} 181
9.7 Further reading 183
9.8 Exercises 183

10 Creating and using subcorpora 185
  10.1 Outline 185
  10.2 Introduction: why create subcorpora? 185
  10.3 Basic steps for creating and using a subcorpus 186
    10.3.1 Defining a new subcorpus via Written metatexual categories 186
    10.3.2 Running a query on your subcorpus 188
  10.4 More on methods for creating subcorpora 190
    10.4.1 Selecting a narrower range of texts for a subcorpus 190
    10.4.2 Defining a new subcorpus via Spoken metatexual categories 191
    10.4.3 Defining a new subcorpus via Genre labels 193
    10.4.4 Defining a new subcorpus via Keyword/title scan 195
    10.4.5 Defining a new subcorpus via manual entry of text IDs or speaker IDs 198
    10.4.6 Modifying your subcorpora 200
  10.5 Saving time by using subcorpora 201
  10.6 Exercises 202

11 Keywords and frequency lists 205
  11.1 Outline 205
  11.2 Introduction 205
11.3  The Keywords function  
11.3.1  About keywords  
11.3.2  Producing keyword lists  
11.3.3  Interpreting and adjusting keyword list settings  
11.3.4  Finding items contained in only one frequency list  
11.4  The Frequency lists function  
11.5  Exercises  

12  Advanced searches with the CQP Query Syntax  
12.1  Outline  
12.2  Introduction  
12.3  From Simple queries to CQP syntax—a primer  
12.4  Regular expressions  
12.5  Part-of-speech and headword/lemma queries  
12.6  Lexico-grammatical patterns and text structure  
12.7  Advanced features of CQP queries  
12.8  Exercises  

13  Understanding the internals of BNCweb: user types, the cache system and some notes about installation  
13.1  Outline  
13.2  BNCweb users: standard users and administrators  
13.3  Additional information available to administrator users  
13.3.1  Overview  
13.3.2  Administrator access to the Query history feature  
13.3.3  Administrator access to user-specific data stored by other features  
13.4  Customizable settings in BNCweb  
13.4.1  Configuration settings available to standard users  
13.4.2  Configuration settings available to administrator users  
13.5  The cache system  
13.5.1  General description  
13.5.2  Maintenance of the cache system  
13.6  Installation of BNCweb  
13.6.1  Prerequisites  
13.6.2  Time and disk-space required  
13.6.3  Configuration of the Perl library bncConfigXML.pm
1 Looking at language in use—some preliminaries

1.1 Introduction

1.1.1 Thinking about goalless, shall and cars

Let's start by having a look at the following three questions:

a) What is the meaning of goalless?

b) How is the word shall used in Present-day British English? Suggest one or two typical examples to illustrate your description.

c) Who talks more about cars, British men or British women?

Question a) concerns the meaning of a single word—this type of question could, for example, be asked by a learner of English as a foreign language who has come across goalless without sufficient context to fully understand its meaning. In contrast, the second question goes beyond lexical meaning; shall is a modal verb (like will, must and can) and is therefore normally used together with other verb forms (like run, sing and be). In other words, rather than simply asking a question about the meaning of a certain word, question b) is about how this word can be combined with other elements of the English language to express a particular grammatical relationship or function. This question might for example be asked by an English teacher who is preparing a lesson on modal verbs. Question c), finally, broadly deals with the relationship between language and society. It is admittedly a bit of an odd question—calling up common clichés and stereotypes about the difference between the two sexes—and you are probably more likely to meet questions of this form during a dinner table conversation than as part of a linguistic enquiry. But there's a deeper reason for asking this question here, which will become apparent when we discuss possible answers, so let's just for the time being assume that this is a perfectly sensible thing to ask.

Task:
Spend a few moments thinking about possible answers to the questions above. Then ask some fellow students or friends the same questions and compare their answers to yours. Do you all agree on what the correct answers are? If not, think about the reasons why these differences may have occurred.
If you are a native speaker of English (or a highly proficient speaker of English as a second or foreign language), you may feel that your intuitions about the language will be fully sufficient to provide answers to all three of them. However, and this may have been confirmed if you did the above task as a group of people, even native speakers quite often disagree about certain aspects of language and its use, and these three questions may be no exception. For example, when answering question a), many people immediately think of goalless as meaning 'aimless, purposeless; having no destination'. Interestingly, typically only few people think of a second meaning of the word, namely that which is used in football to refer to 'a game in which no goals were scored on either side'.

Moving on to question b), your intuition may have told you that shall is quite old-fashioned and slowly dying out, while speakers nowadays prefer will and other future time expressions such as going to or gonna. You may also have worked out that the modal auxiliary shall is followed by the infinitive without to, and perhaps even that shall is used most frequently when the subject is a first person pronoun (that is, I or we). As a result, the typical example you gave might have looked something like this:

(1) I shall ring you up as soon as I arrive.

Alternatively, you might also have thought of a use of shall in offers, suggestions, requests for instructions, and requests for advice. This use takes the form of a question, i.e. the subject (e.g. I) follows the modal shall. A typical sentence is shown in (2).

(2) Shall I carry your bag?

When asked about the level of formality of this second type of use, people are usually quite undecided. However, the majority have the impression that this is a particularly polite—and therefore formal—usage. Furthermore, when asked about which of the two structures is more frequent, people often don't feel confident in providing a clear answer.

As for question c), most people would answer this by stating that men talk more about cars than women.

This quick summary clearly shows that the intuition-based approach can result in a considerable range of possible answers, and it is not clear how close to the "truth"—or perhaps better, how close to actual usage—they really are. In order to determine this, you may therefore want to find independent confirmation. Let us consider some ways in which this could be done. For example, dictionaries will easily help you with question a). Indeed, the Oxford English Dictionary (OED) lists both of the meanings of goalless that were mentioned above. Yet
you may also want to know which of the two senses is more common in Present-day English: unfortunately, the OED does not give you any help there.¹

For the second question, grammar books are an obvious source of additional information. However, in this context it is important to ask what authority the author of a particular grammar book has for writing up his or her description. If its contents are heavily based on the author's intuitions about the English language, they may in fact also not fully reflect actual usage, even considering that an author of a grammar book is likely to be very knowledgeable about such matters.²

Another way of trying to find answers to at least the first two questions is by asking a wide range of informants who are native speakers of English. This is best done by giving them apparently unrelated questions whose context will trigger the use of the feature in question (e.g. shall vs. will). This method of "informant testing" is often more accurate than a direct appeal to native speaker intuitions, as the information provided is less likely to be influenced by factors such as self-censorship or accommodation. For example, when asked directly, an informant may opt to use I will or I'll—instead of I shall—because he or she does not want to give the impression of being old-fashioned. However, the same informant may not have any problems with using I shall in situations where they are not aware of the fact that the questions or tasks are designed to extract information about their use of will vs. shall. Although this informant-based method is clearly more informative than relying purely on the intuitions of a single speaker, it is obviously also much more difficult and time-consuming to carry out.

Finally, you could simply decide to observe what's happening around you and draw your conclusions on the basis of the data you collect. Every time someone talks about a car, you take note of the speaker's sex. Every time someone uses shall, you look at the type of construction in which it is used. And every time you read or hear goalless, you use the context to find out more about the meaning of this word. Once you have noted down a sufficient number of instances, you will have a reliable basis for a description of what is really going on with goalless, shall and talk about cars in today's English. However, there are two major problems with this method. First, with fairly infrequent words and expressions (e.g. goalless), you will have to wait a very long time before you have enough data to make any general claims. Secondly, and more importantly, your language experience may differ dramatically from that of other people who also use English. If, for example, you are a student at a British university, a large

¹ However, some learner dictionaries (e.g. the Collins COBUILD Advanced Learner's English Dictionary 2006) do indicate whether certain senses are particularly common or rare.

² It has to be pointed out, however, that many modern descriptions of English are no longer purely intuition-based. Instead, grammar books nowadays are often based on exactly the kind of data and methodology that we will describe in this book.
part of your language use will take place in interactions with other students and a considerable part of what you read will be academic texts (like the one you are reading right now). This is very different from the language experience of an average coal miner, lawyer, or jazz musician. And maybe the experience of these other types of language users will be particularly different from yours just in the context of the three questions you are trying to answer.

This book is about a method—and a tool—that will allow you to eliminate these two major problems to a very large extent. Suppose you had access to a huge collection of texts and conversations produced by a cross-section of today’s population in Britain—i.e. by students, lawyers, jazz musicians, coal miners and a whole range of other types of language users. Further suppose that you would have access in such a way that it is possible to easily search the complete collection in a matter of seconds, and that you would also be able to get further information about the search results that are retrieved (e.g. about the type of speaker or writer, the kind of context in which it was produced, etc.). This is exactly what the British National Corpus (BNC) and BNCweb will give you.

1.1.2 Clues from a corpus—the BNC

The BNC is a 100 million word collection of samples of written and spoken language from a wide range of sources. It was put together to represent a wide cross-section of current British English, and contains a large number of language samples from different kinds of texts, produced by different kinds of language users and made available in different ways. A more detailed description of the corpus—including an account of how it was compiled, what type of texts it contains and what additional information is available about these texts—will be given in Chapter 3. BNCweb is a user-friendly web-based interface that was created to search (or as we say, to query) the data contained in the BNC. It gives you easy access to a wide range of functions that allow you to linguistically analyze the results of your queries. Originally developed at the University of Zurich by Hans Martin Lehmann, Sebastian Hoffmann and Peter Schneider (see Lehmann et al. 2000), BNCweb is nowadays maintained and further extended by Sebastian Hoffmann and Stefan Evert. The functionality of BNCweb is described in detail in the remaining chapters of the book.

To whet your appetite, let us quickly return to our three questions and see what clues we can find with the help of the BNC and BNCweb. A quick search for goalless shows that there are only 86 instances in the whole corpus. So on average, the word occurs less than once in every million words. Figure 1.1 displays how BNCweb will present the results of the search—or query—to you. This kind of output is generally referred to as a concordance.
Looking at language in use—some preliminaries

Looking at this concordance, it is immediately obvious that football appears to be the predominant context in which British English speakers make use of the word *goalless*. In fact, if you were to look at all 86 instances in more detail, you would find that every single one is from the field of sports. Now, this does not of course mean that the other meaning of *goalless*—i.e. 'aimless'—does not exist at all in Present-Day English. After all, although the BNC contains nearly 100 million words, it is actually quite tiny in comparison with the totality of language use in Britain, and it is entirely possible that some very infrequent features are not represented at all in the corpus. However, you can now safely say that the 'aimless' meaning of *goalless* is very marginal indeed. The other obvious point to note from this list of results is that *goalless* often co-occurs with *draw*, referring to a game during which no goals are scored. Of the total of 86 instances, 51 (59 per cent) co-occur with *draw*. If you are a learner of English as a foreign language, this is useful information because it will not only allow you to understand the most common meaning of the word but it will also give you the opportunity to notice how it is used idiomatically by native speakers.

What can the BNC tell us about the second question, i.e. how *shall* is used in Present-day English? A simple lexical search of *shall* gives you many more hits than you will want to look at: there are 19,505 instances of *shall* in the whole

---

3 At least this is the case in British English. Speakers of other varieties of English may prefer the expression *goalless tie* instead.
BNC. However, we could restrict our investigation by looking at the spoken part of the corpus only. A good reason for doing this is that we suspect that *shall* is becoming less common nowadays: it is widely assumed in linguistics that when something changes in a language, that change generally starts in the spoken rather than the written variety.

With *BNCweb*, it is easy to restrict searches to sub-parts of the corpus, e.g. spoken texts only. This part of the BNC contains about 10 million words, but *shall* still occurs 2,735 times. This suggests that *shall* is still in common use in Present-day English—compare this to the 86 instances of *goalless* in the whole corpus—and that it is still a long way from vanishing from the language altogether. Figure 1.2 shows a screenshot of the first five hits that are returned by *BNCweb*.

As you can see, both types of uses mentioned above are found in these first few sentences, e.g. *shall* *we listen* *to you* (no. 1, where the personal pronoun follows *shall*) and *I shall be contacting him* (no. 4, where the personal pronoun is placed first). But which of the two patterns is more frequent, and can we find out more about preferences among particular (types of) speakers?

![Figure 1.2: Result of a search for *shall* in the spoken component of the BNC](image-url)
One way of proceeding from here would now be to look at every single one of the 2,735 instances of *shall* returned by the search, always noting down information about the speaker (if available) and the grammatical pattern in which it is used. However, this would be very tedious and time-consuming. Fortunately there are quicker and more convenient ways of seeing patterns in the way *shall* is used. Let's for example consider the age of speakers who use *shall*. Our intuition might tell us that older speakers are typically more conservative and might therefore more likely use an old-fashioned form. If this were true we might then expect the use of *shall* to be more frequent among older speakers than among younger ones. *BNCweb* allows you to test this hypothesis in just a few simple clicks (using the so-called DISTRIBUTION feature).

![Figure 1.3: Distribution of *shall* over the category "Age of speaker" in the spoken component of the BNC](image)

As you can see in Figure 1.3, the data is not conclusive: older speakers do not use *shall* more frequently than younger ones; in fact, it is the youngest group that can be seen to use this modal most often, while the oldest age group is found somewhere in the middle of the table. Clearly, this finding does not support the view that *shall* is archaic and in the process of dying out.

But let's dig a little deeper. Another thing you can do with *BNCweb* is to find out which words occur particularly often before or after *shall*. In this way, you could confirm your hunch—if this is what you came up with in response to question b)—that the first person pronoun subjects *I* and *we* are very frequent both before and immediately after *shall*. It turns out that nine out of every ten instances of *shall* occur together with *I* or *we*. The interesting question now is whether there are any differences among the various age groups with respect to the two possible sentence types, i.e. *I/we shall* vs. *shall I/we*. Again, *BNCweb* gives you this type of information very quickly—the results are shown in Figures 1.4a and 1.4b.
As you can see, the two patterns show an opposite trend: I/we shall is most often used by older speakers (182 instances, on average 160 times per million words), but the same group of speakers use shall I/we the least (103 instances—about 91 instances per million words). The reverse is true for the youngest speakers, who use shall I/we most often (175 instances, 454 instances per million words) but hardly use I/we shall at all (only 11 instances).

Now that you have obtained these findings—or DESCRIPTIVE STATISTICS—you have quite a good foundation for answering the second of the three questions at the start of this chapter. First of all, you can say that shall is still quite frequent in Present-day English—although of course you haven't yet checked how much more frequent will is. Secondly, you can say that one of the two uses, i.e. I shall or we shall is predominantly used by older speakers, suggesting that the declarative form may indeed be old-fashioned. Furthermore, you can say that the other type of use, which includes offers, suggestions and requests for instructions expressed by shall I? or shall we?, is mainly used by younger speakers. Finally—and most crucially—you could look at this age distribution as a snapshot of a change in the English language that is still ongoing, and from this predict what the future of this use might be. Think about it: what will happen
when the young speakers represented in the BNC will be sixty or older? Will they have started using I/we shall more frequently by then because that's simply what older speakers do? Probably not. A much more likely interpretation of the data is that the declarative use is slightly dated and indeed slowly leaving the language—it is dying out. The use of shall for offers and suggestions, on the other hand, is probably going to increase even further. If this is true, perhaps it would make sense for teachers of English as a second or foreign language to introduce this type of use first, and only later go on to present the more marginal and archaic uses.

Even though we have extracted all sorts of information from the corpus, we have of course not yet answered the question whether the use of shall in offers and suggestions is particularly polite or not. Unfortunately, the tables we have compiled so effortlessly do not help us find this answer. Instead, we will have to look more closely at a sufficient number of instances of this particular use of shall in context. Descriptive statistics are almost always only one side of the coin, and a comprehensive description of a linguistic phenomenon will often require both a quantitative and a qualitative analysis of the data.

Finally, let's have a quick look at the third question—but how do we do this? How can we really answer the question whether men or women talk more about cars? A very basic approach would be simply to look for the word car and to have BNCweb calculate the same kind of distributional statistics as for shall above, just this time for the sex of speakers rather than age. Figure 1.5 displays the result of this calculation. Interestingly, women seem to use the word car more often than men. Notice, by the way, that the number of actual hits is higher for men (1,789 male vs. 1,597 female uses), but we need to take into account that there are more words in this corpus uttered by men than by women. This is why measuring the frequency across the same amount of text—as occurrences per million words, for example—is important: 485 instances per million words (pmw) for women vs. 361 pmw for men. We will—or we shall?—return to this issue again in later chapters.

<table>
<thead>
<tr>
<th>Sex:</th>
<th>No. of words</th>
<th>No. of hits</th>
<th>Dispersion (over speakers)</th>
<th>Frequency per million words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female</strong></td>
<td>3,290,569</td>
<td>1,597</td>
<td>333/1,360</td>
<td>485.33</td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td>4,949,938</td>
<td>1,789</td>
<td>438/2,448</td>
<td>361.42</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>8,240,507</td>
<td>3,386</td>
<td>771/3,808</td>
<td>410.9</td>
</tr>
</tbody>
</table>

Figure 1.5: Distribution of the word car over male and female speakers in the spoken component of the BNC
But what have we actually answered by looking at Figure 1.5? If you think about it, not all that much. First of all, we have forgotten an important part of the use of the word *car*: the plural form *cars*. Secondly, and much more importantly, what does it actually mean to "talk about cars"? Do you always need the lexical item *car* to do so? If someone says *I bought a Merc yesterday*, clearly this is also talking about a car. Conversely, what about mentioning a *car boot sale*? The word *car* is used here, too, but is the speaker really talking about cars? You can probably see that finding a reliable answer to the third question involves much more than a simple search and a few clicks in *BNCweb*—and this is a valuable insight. Some research questions are much easier to answer with the help of corpora than others, and it is important to know both the opportunities and the limitations that the use of corpora involves.

### 1.2 Why read this book?

This book is mainly about the practical steps involved in answering relevant linguistic research questions with the help of the BNC and *BNCweb*. As you will quickly realize, *BNCweb* is a very user-friendly tool: it is easy to perform a simple search of the corpus, and a few mouse-clicks are usually sufficient to give you lots of further information about your query. You might therefore wonder: is it really necessary to read a detailed manual? Our answer to this is: first, this book is not just a software manual—it was written by linguists interested in language study, and goes beyond a description of what the software can do. It is focused on what linguistic questions you can answer using the software and how you can go about interpreting the data generated by it in a meaningful way. The ease of use of *BNCweb* makes corpus-based language study appear simpler and more straightforward than it really is, and masks some considerations that should be part of every enquiry.

First and foremost, it is necessary to know more about the corpus: What is actually in the BNC? How did the compilers of the BNC choose the texts? How much do we know about the speakers and writers of the texts and the conditions of their production?

Second, it is necessary to learn the theoretical bases and methodological steps in corpus-based research: *How do I interpret the results presented by BNCweb? What do they tell me about British English as a whole or the text varieties that I chose to examine? What do they not tell me? How do I compare results from different searches? How can I be sure the results are reliable? How do I know that my searches really are relevant to answering my research questions?* This book will help you answer these important questions, and you will learn about theory and methods as you work your way through the chapters. It will help you avoid the potential problems and pitfalls that could turn the first
steps of a novice corpus user into a potentially frustrating or misguided experience.

In this book, methodological points are addressed and illustrated in the context of actual investigations of language use. It is this combination of theory with extensive hands-on practice that makes the book different from others in the field of corpus linguistics. The functionality of the various features of BNCweb are explained through "real-life" examples of linguistic issues, combining "how-to" with a discussion of theoretical and methodological considerations.

1.3 Organization of the book

The organization of the chapters is as follows: Chapter 2 introduces some of the fundamental concepts of corpus-linguistic methodology. This is followed by a detailed description of the British National Corpus in Chapter 3. In Chapter 4, we then illustrate the basic search functionality of BNCweb and show how a query result—in the form of concordance lines—can be investigated to gain insights into the use of a particular word or phrase. This is followed by a second methodology chapter—Chapter 5—which covers a number of important issues relating to the comparability and reliability of findings made through BNCweb. We focus on why normalized frequencies are important (and how they are calculated), introduce the concepts of "precision" and "recall", and testing for statistical significance. In Chapter 6, we offer a detailed description of BNCweb's "Simple Query Syntax" and show how it can be used to perform highly sophisticated searches of the corpus.

The next three chapters are then devoted to various ways of further manipulating and analyzing your query result. Chapters 7 (DISTRIBUTION and SORT) and 8 (COLLOCATIONS) cover ways of exploring your query results automatically, i.e. without the need to look at concordance lines individually (or, as it is often called, manually). In Chapter 9, we then turn to the manual annotation of concordance lines and guide you through the process of adding your own classifications to a query result (either within BNCweb itself or with the help of third-party programs such as Microsoft Excel).

For many research questions, it will be necessary to restrict searches to a subsection of the whole BNC—a so-called "subcorpus". Chapter 10 illustrates the various ways in which subcorpora can be defined. Furthermore, we will show how user-defined subcorpora can be employed to make repeated searches of (sub-parts of) the BNC more efficient. BNCweb also offers two additional functions—the FREQUENCY LIST and KEYWORD features—that can be used to explore the corpus data from a more "whole-text" or macro perspective (i.e. without starting from a concordance); these will be covered in Chapter 11.
In addition to the Simple Query Syntax introduced in Chapter 6, *BNCweb* also accepts queries in something called "CQP Query Syntax", whose advanced features allow users to perform even more powerful and flexible searches of the corpus. Given the much less intuitive nature of this query syntax, however, the description offered in Chapter 12 is likely to appeal predominantly to more advanced users. Chapter 13, finally, concerns practical issues in the running of *BNCweb*. It covers such aspects as the difference between standard users and users with administrator rights, and it also describes some internal aspects of the workings of the software that have been designed to optimize access by whole groups of users. The chapter concludes by outlining some issues relating to the installation and maintenance of *BNCweb*.

### 1.4 How to use this book

This book is probably best read while sitting in front of a computer with access to *BNCweb*. This will make it possible for readers to gain hands-on experience in using the tool by following the step-by-step descriptions of the many sample analyses. Each chapter also contains a number of tasks and exercises that will offer further opportunities for enhancing and broadening the practical skills of readers. However, the book has been written in such a way as to make independent reading of its contents a worthwhile experience.

Several of the chapters contain a considerable amount of information—in fact, it may be too much to fully "digest" everything in one sitting. This especially applies to the two chapters which introduce the Simple Query Syntax and the CQP Query Syntax (Chapters 6 and 12), as their descriptions are designed to be useful as a comprehensive reference to the query language. Although it may be informative to read these chapters in one go, you will probably find yourself returning to their contents at some stage in the future, as your need to make more complex searches arises.

A similar comment applies to the chapter describing the BNC (Chapter 3) and to the methodologically oriented Chapters 2 and 5. While we recommend that you consult these chapters thoroughly before you conduct any serious studies on the BNC, we would like to encourage you to explore the different features and options of *BNCweb* at your own pace, so don't worry if you don't fully understand everything the first time around. As you become more experienced and more familiar with the output provided by *BNCweb*, you will likely get a better grasp of the more theoretical aspects of corpus linguistic methods that we discuss in these chapters. They are therefore well worth revisiting. In sum, we are confident that this book will give you a thorough grounding in corpus linguistic theory and methods, as you learn by doing—as we guide you through this powerful yet user-friendly program.