

# Contents

---

<i>List of Figures</i>	x
<i>List of Tables</i>	xi
<i>Acknowledgements</i>	xii
<b>1 Introducing Online Corpora</b>	1
Using a corpus	1
Choosing your corpus	4
The tradition of corpus analysis of English	9
Five online corpora of English	10
Analysing online corpora	16
The organisation of this book	19
<b>2 Interpreting Corpus Data</b>	21
Quantitative and qualitative analyses	22
Representativeness	22
Frequencies	28
Normalisation of frequencies	30
Mutual Information	32
Other measures of collocation	34
Key words	37
Other statistical tests	42
Summary	44
<b>3 Exploring Lexis with Corpora</b>	45
What is a word in a corpus?	46
Obtaining lexical information from a corpus	48
Analysing lexical data	52
The lexicogrammatical environment of words	53
The semantic environment of words	59
The wider social environment of words	63
Summary	66
<b>4 Exploring Grammar with Corpora</b>	67
Attitudes to grammar	67
What is grammar?	68

Word categories	71
Exploring lexical items with a corpus	72
Exploring grammatical items with a corpus	76
Exploring phrases with a corpus	78
From phrase to clause	83
Complex sentences	88
The grammar of speech and writing	91
Delexicalised verbs	92
Colligation	93
Verb systems	96
Data-driven grammar versus intuition	99
Summary	100
<b>5 Exploring Discourse with Corpora</b>	<b>101</b>
What is discourse?	101
Using corpora to explore discourse	105
Analysing spoken discourse	111
Analysing written discourse	116
Intertextuality	121
Summary	122
<b>6 Exploring Pronunciation with Corpora</b>	<b>124</b>
Describing different accents	125
The international phonetic alphabet	126
Consonant sounds	126
Consonant symbols	127
Vowel sounds	129
Using online corpora to explore pronunciation	131
Accent variation	132
Accents of English online	136
Exploring vowels online	137
Fast talking	143
The acquisition of variation	147
Lexicalisation	149
Intonation	152
Summary	153
<b>7 Contextualising Corpus Texts</b>	<b>155</b>
Text and co-text	155
Text and context	157
Metadata	158
Corpora and sociolinguistics	159
User-related variables	165
Use-related variables	169
Summary	172

<b>8 Conclusion: Issues in the Use of Corpora in Teaching and Research</b>	173
Using corpora in teaching	173
Developing your own research interests	178
The future of online corpora	179
<i>Appendix: Online Corpora</i>	
<i>Bibliography</i>	183
<i>Glossary</i>	188
<i>Index</i>	193
	201

# Introducing Online Corpora

# 1

---

This book has a number of intended readerships. It should appeal to

- students who are embarking on a formal course in English language, at upper school level or university;
- teachers who wish to know more about the technology available to support English language education in schools;
- learners and teachers of English as a foreign language, who wish to explore how language is used in a vast quantity of ‘authentic’ texts; and
- general readers who are simply curious about how the English language works and about new methods of exploring this topic.

All of its readers will be united by a common interest in the English language and a desire to explore its many written and spoken forms. However, the book assumes little or no previous formal experience of the study of the English language. A substantial part of the book, therefore, reviews some of the basic concepts that readers will encounter when they begin an exploration of how words work in English, how they form phrase patterns, sentences and ultimately entire texts and discourses. It also reviews some of the basic tools required to study the different accents of English. This introductory chapter sets out a number of concepts, some of which will be discussed in greater detail as the chapters unfold. The glossary at the end also provides an explanation for many of the concepts from linguistics and corpus linguistics which we draw on in the book. Readers, particularly those wishing to take corpus techniques further, may also find Baker, Hardie and McEnery’s *Glossary of Corpus Linguistics* (2006) to be a useful reference work. Finally, readers will find a website to accompany the present book at [www.palgrave.com/language](http://www.palgrave.com/language).

## Using a corpus

One of the most exciting developments in the exploration of English over the past 40 years has been the accumulation of vast electronic archives, or corpora,

of written and spoken texts, texts stored on computers and manipulated easily and quickly by search programs. The availability of such corpora has changed the working practices of linguists, particularly those scholars interested in the meanings and patterns of words and phrases. Before the advent of electronic corpora, it was a long and arduous process to compile and search substantial bodies of data in order to confirm or challenge one's own intuition as a language user. Nowadays vast bodies of data are available at the touch of a keyboard and the click of a mouse. Early beneficiaries of large, electronically searchable, corpora of English were dictionary makers, who suddenly had hitherto inaccessible evidence for the meanings of words, and language learners, who also had access to materials constructed using corpus data that seemed to guarantee relevance and authenticity. Only in the last few years, however, have language corpora begun to be freely available online to the casual browser, language learner and relatively novice student.

A few years ago, Susan Hockey (2000, p. v) noted that 'The World Wide Web is good for looking at material but it does not provide many tools for analysing and manipulating that material.' Since then, happily, the resources available on the Web have improved to the extent that students of language and linguistics can make considerable inroads into linguistic study solely by using freely available corpora, provided that they know where to look, have an appreciation of a few basic notions, and know how to maximise the potential of language corpora, with all their idiosyncrasies and differences. So, as written and spoken corpora become available to an ever wider network of potential users, guidance is needed in the use of them to explore aspects of language. This book is designed to give that guidance.

The careful analysis of corpora can give insights into (i) how language is *really* used, rather than how people *think* it is used and (ii) how it is *commonly* and *typically* used. To see how true this is, you can try jotting down your immediate thoughts about the typical linguistic contexts of a word in real language and then compare your list with the sample of evidence from a corpus such as

### TASK 1.1 Comparing intuitions and evidence

1. Without reference to a dictionary, write down a definition of the adjective *seedy* and some typical examples of the nouns that it describes.
2. On a computer, log on to the BNC interface at BYU: <http://corpus.byu.edu/bnc/>
3. In the 'Display' section, click on 'List'.
4. In the 'Search String' section, type *seedy* in the 'Word(s)' box.
5. The results will show you the number of occurrences of *seedy* in the 100 million words of the BNC.
6. Click on the word *seedy* in the 'Results' section in order to see some of the contexts of usage. Note down some of the nouns that are modified by *seedy*, for example *seedy affair*.
7. To compare your results with 360 million words of current American English, go to [www.americancorpus.org](http://www.americancorpus.org) and repeat Steps 3–6.

the British National Corpus (BNC) or the Brigham Young University (BYU) Corpus of Contemporary American English, both of which we'll introduce in more detail later in this chapter. For example, what sorts of nouns are modified by *seedy*? Now complete Task 1.1.

The results of these searches may confirm your expectations or surprise you. The term *seedy* derives from the notion 'gone to seed' and out of context it might suggest shabbiness or lack of care. Some of the results of the corpus search confirm such a definition. For example,

Harold Macmillan's seedy and stagnant Britain

However, the majority of examples in the BNC suggest something that is sexually unsavoury about the term *seedy*. For example,

seedy affair  
 seedy porn photos  
 seedy abortionist  
 a rake's progress of late nights, seedy bars and relentless beer bellies  
 the seedy world of prostitution  
 the film-maker's seedy little wife, pompously and unsexually nude  
 seedy northern beauty contest  
 a seedy image of people who use porn  
 gave up her seedy career and wrote an exposé of the porn business.

While by no means all the examples conform to this pattern of meaning (there is, for example, an innocent reference to British actor Will Hay's creation of the role of a *seedy, blustering and ineffectual teacher* and even a reference to *seedy grass-heads caught in my socks*), many do. These instances are enough to suggest that in contemporary English the 'sexually unsavoury' element of the meaning of *seedy* is strong enough to carry into contexts in which it is not necessarily explicitly mentioned; for example, *business friends who ran seemingly dowdy or seedy little second-hand enterprises in shops dark and dusty*. This element of meaning has not yet, however, found its way into the online Oxford English Dictionary.

At this very basic level, then, checking your intuitions against corpus data can help you confirm or challenge your preconceptions about words, what they mean and how they are used. We build up our intuitions, of course, from long experience of language read and heard in a linear fashion. We come across instances of words like *seedy* from time to time, perhaps once in a million words that we read or hear spoken. From each of these contexts we build up an evolving picture of what the word can mean in various contexts. By bringing a substantial number of these instances together in a small space – a corpus – we can become aware of patterns that remained below the surface of our consciousness. As John Sinclair, one of the pioneers of modern corpus linguistics, noted, 'The language looks rather different when you look at a lot of it at once' (Sinclair 1991, p. 100).

## Choosing your corpus

There are some important issues that every corpus user should be aware of when choosing a corpus that will help you answer your own questions about language. We will return to some of them again in later pages. First of all, we need to consider what kind of collection of texts a corpus is, and what differentiates it from other online textual resources.

### The nature of a corpus

A corpus may be described quite simply as a body of texts; in fact, this is the literal meaning of the word. Most corpus linguists, however, prefer to be rather more specific, and describe a corpus as a large principled collection of texts, that is, one which has been created for a purpose. This is now a widely accepted definition. Even more precisely, a modern corpus is a sample of naturally occurring language, in electronic form, which has been designed to represent a language, language variety, register or genre. The key word here is 'designed', and this is what distinguishes a corpus from its close relative, the text archive. While an archive may have no predetermined structure and is not intended to represent something larger, a corpus is motivated, created with a linguistic purpose in mind. The Web as a whole may be used as a resource for linguistic exploration, and with the arrival of online tools such as WebCorp, it has become much more straightforward to do so. Given its constantly changing size and nature, however, it is better treated as a massive archive rather than a corpus. Nevertheless, it is certainly possible to create a corpus from material on the Web, by selecting texts according to particular criteria.

It is important to bear in mind the principled nature of corpora when using online examples for research. Since corpora are usually compiled with a purpose, it is necessary to match your needs as a corpus user against the interests and goals of the corpus designers. Although many corpora do indeed contain vast quantities of data, it may not be the kind of data you are interested in. For example, the TIME corpus brings together 100 million words of text, easily putting into the shade the 4 million words of the Scottish Corpus of Texts & Speech (SCOTS) and the 1.8 million words of the Michigan Corpus of Academic Spoken English (MICASE). However, as its name suggests, the 100 million words of the TIME corpus consist entirely of texts taken from *TIME* magazine, and while this is a fascinating source of a particular type of journalistic data, it does not contain any spoken English, or other varieties of written English. On the other hand, the MICASE corpus contains exclusively spoken English, but only from the domain of academia – genres such as lectures, seminars and consultations with tutors. Alongside its written data, the SCOTS corpus contains 800,000 words of spoken data across a range of situations, from lectures to spontaneous child–parent interactions, but given its size it does not necessarily give you much linguistic data from each individual situation. Each corpus clearly has its uses, and as a user you need to think about which one is most appropriate for your own explorations. Fortunately, most online corpora explicitly set out their design criteria

and contents, so that you can assess their suitability for your purpose before you begin, and then carry out a small, pilot investigation to see how searches work in practice.

This book contains a substantial number of tasks and activities that aim to use the most appropriate online resource or resources in each case. You should therefore find it easy to design your own follow-up studies and develop your own interests. It will normally be possible to attempt the same activity with a different corpus, or to use the same corpus but subtly modify the research question. This will give you a feel for the nature, design and strengths of the corpus you are using, and allow you to appreciate its (inevitable) weaknesses.

### Representativeness

As hinted in the previous section, the issue of representativeness is a crucial one. It is possible to make valid generalisations about a language or language variety only if the corpus is a fair sample of that language or language variety. Just as no one would claim to be able to make valid conclusions about Singaporean English from analysing a corpus of Scottish English, so we cannot draw valid findings about the use of the word *like* in spoken language from a corpus which contains only written language, and vice versa. Similarly, we cannot assume that if a grammatical construction is common in a corpus made up of scientific articles then it is also commonly used in the language as a whole. Nonetheless, such a finding may form a good hypothesis on the basis of which further research can take place.

Also, it is important to bear in mind that a corpus can only provide positive evidence of a usage or construction. In other words, it cannot tell you what is *never* said or written. The philosopher Karl Popper neatly illustrated this point by observing that a naturalist who came upon any number of white swans would still not be in a position to make the claim that black swans do not exist (Popper 1959). However, it would only take a single black swan to come along to falsify the claim. The force of the argument lies in the seductive nature of corpus data, which might, for example, show thousands of examples of *had* in the active voice. An observer might be tempted therefore to claim that *had* only ever appears in the active voice. However, it only takes a single example of a passive construction – such as *a good time was had by all*, which is attested in the SCOTS corpus – to negate that claim.

A representative corpus should, however, be constructed so as to reduce the possibility of making false claims about the language based on partial or skewed data. A word of caution is also necessary here. Corpus linguists talk frequently of the representativeness of corpora, but rarely agree on whether it is even possible to attain perfect representativeness, let alone what would constitute the make-up of this ideal corpus. Occasionally, the issue is easy to address. If you want to investigate the use of greeting formulae in published correspondence in the UK in the eighteenth century, then the population of texts is finite and so, copyright matters notwithstanding, you could in principle select a fair and representative sample of the overall population, employing such criteria

as decade of writing, gender and age of writer, and the level of formality of the letter.

Most of the time, however, the issue of representativeness is far from straightforward, and it is particularly thorny if you want to make generalisations about a language as a whole. What are the ‘right’ proportions of written language and spoken language? How do we deal with the fact that some texts are very widely read or heard, over either a long or a short period of time (for example, the Bible, the Queen’s speech or the Presidential address, the fiction of J.K. Rowling, front-page newspaper articles), while others are read or heard by few people, or small groups of people (for example, specialised scholarly monographs, horse-racing reports, sermons in a small village church). How can we make sure that the voices of people of all ages, genders, occupations, races, religions, types of upbringing and so on are fairly represented?

What about the sorts of language that are by nature private – personal diaries, intimate conversations, the words you mutter to yourself when working through the instructions to build a self-assembly bookcase? Can we justify not including them in a corpus? Most large corpora contain some ephemeral texts, spoken or written, that might seem odd out of context. On its own, for example, the following excerpt from the spoken section of the BNC might cause the casual user to question the principles of the designers:

mm mm. Mm. Mm. Mm. Mm. Ma ah ah! Daddy! Daddy! Hello! Dee dee dee, dee dee dee dee, dee dee dee dee dee, dee dee dee! What is she meant to think about? Pauline? Ah yea yea yea yea yea yea! Ah, yea yea yea yea yea yea yea! Yea yea yea yea yea yea yea! Yea! Ooh ooh ooh ooh! Ooh ooh! Ooh! Ooh ooh! you didn’t tell me it was on yet. Will you, shall I take her up tonight? If she’ll let me. What? [BNC text KDG, conversation].

This spontaneous effusion in isolation might well tell us very little, but as part of a larger, representative corpus, it can function as a legitimate thread in the larger tapestry of language.

Representativeness can simply be thought of as the inclusion in a corpus of a large number of texts in a large number of registers and genres. However, there are also statistical measures of representativeness, which can guide the corpus designer – and these are discussed more fully in Chapter 2. Most general corpora will at least try to capture a snapshot of different types of speaking and writing, from casual conversation to carefully composed legal documents. Further discussion of the issue of representativeness can be found in most introductory books (see the further reading at the end of this chapter), or the articles by Biber (1993) and Clear (1992).

## Size

The issue of size is a separate but related issue to representativeness. For applications in lexicography, it is important that corpora should be large, usually tens if not hundreds of millions of words in size. O’Keeffe, McCarthy and Carter

(2007, p. 4) observe that to obtain a sufficient range of prepositions that follow *bargain*, they needed to search a corpus of about 10 million words. Online corpora range from the 360 million words of the BYU Corpus of Contemporary American English (with more words being regularly added) to the 1.8 million words of MICASE. As O’Keeffe, McCarthy and Carter go on to state (*ibid*, p. 4),

In terms of what constitutes a large or a small corpus, it depends on whether it is a spoken or a written corpus and what it is seeking to represent. For corpora of the spoken language, anything over a million words is considered to be large; for written corpora, anything below five million is quite small.

Nevertheless, a lot of interesting research can be done on the more frequent or ‘core’ words, constructions and features of a language with corpora of even a few tens of thousands of words. Size, along with design, is a feature of a corpus which should suit the purpose to which it is put. In the context of corpora for language learning, Chambers (2007, p. 9) notes that

It is clear that, despite the corpus linguist’s need for corpora to be as large as possible in support of the researcher’s quest for the elusive quality of representativity, several of those who focus on classroom applications and who have experience of working with learners have become aware of the usefulness of the small corpus.

Several online corpora are very large indeed, but many others are relatively small and precisely focused. Rather than necessarily having a large corpus, it is more important to have a good understanding of the nature of the corpus being used, and therefore the level of faith which one can put in findings from it. A list of currently available free online corpora, with some details about their nature, is given in the Appendix.

### **Type of corpus**

It should already be clear that the features of online corpora for linguistic study can be diverse in nature. While it is not necessary to provide a comprehensive survey of corpus types here, it is helpful to be aware of a few general distinctions.

One important distinction is between synchronic and diachronic corpora, that is, between corpora which contain texts from a particular time period (such as English from the 1990s) and seek, therefore, to provide a snapshot of language usage, and those which can be used to investigate language change over time, as would be the case for a corpus of personal correspondence between 1700 and 1900, for example, and is the case for the TIME corpus, which spans most of the twentieth century. Diachronic corpora can be used to investigate neologisms or new coinings, or changes in grammatical constructions over time. A specific type of diachronic corpus, known as a monitor corpus, is one to which new texts

are continually added: the BYU Corpus of Contemporary American English is an example of a monitor corpus.

Some corpora have been specifically designed to study translation, and there is a distinction to be drawn here between parallel and comparable corpora. Although the terminology varies, a parallel corpus is generally taken to be one which contains the same texts in a number of language versions, while a comparable corpus contains texts which are functionally equivalent in two or more languages. So while a collection of translations of the same novel would, by this definition, constitute a parallel corpus, a collection of (original, not translated) newspaper articles in different languages about the same issue would be termed a comparable corpus. Actual translation corpora may have features of both types.

### Annotation

Annotation refers to the process of adding information to a corpus, so that it becomes possible to search for features that lie below the surface of language. Consider the problem: it is easy to search a corpus for the sequence of characters *e-a-r-n*, but the computer will not know to retrieve instances of the inflected forms *earns*, *earned*, *earning* as well. If you want to retrieve these forms of the verb from an unannotated corpus, you would have to either type each form into the search facility in turn, or use wildcards if this facility is available (for example, *earn\** to retrieve *earn* followed by any number of characters, including zero – but note that this would also find the unrelated *earnest*). Alternatively, you could use a corpus which was lemmatised, that is, annotated so that each form of the verb is attached to a headword or ‘lemma’, which is generally the form of the verb found in the dictionary. Similarly, if you are interested in the word *search* as a verb but not as a noun, then a corpus in which every word form in context is tagged with the right part of speech would allow you to specify that you want only the verb spelt *s-e-a-r-c-h* and not the noun with an identical form.

Lemmatization and part-of-speech tagging are two of the most common types of annotation, but in theory any feature of language can be tagged. Imagine you wanted to find in a corpus all of the words which had to do with the concept of EATING. In that case a semantically tagged corpus would help: searching for EATING might retrieve *eat* itself, *devour*, *munch*, *gobble*, *feed*, *graze*, *dine*, *nibble* and so on. The possibilities for corpora annotated with semantic information are limitless: a first step in this direction is the possibility of searching the BNC, TIME corpus and BYU Corpus of Contemporary American English for synonyms as identified by a thesaurus.

Annotation is generally expensive and time-consuming, especially if it has to be done manually rather than automatically. Because of the work involved, tagging tends to be reserved for small corpora which are used to investigate specific research questions but are not made generally available. Larger corpora, like the three just mentioned, have been tagged automatically. However, with a bit of ingenuity, and trial and error, you can find out a lot about words, meanings and grammar even with an untagged corpus.

## The tradition of corpus analysis of English

If we understand corpus linguistics to mean simply the empirical analysis of language, that is, the exploration of language using authentic texts as data, then it has a long history. Manual analysis of texts has been used for over a century for research into the areas of lexicography, dialectology, anthropology and grammar. To take just one example, Charles Fries' *Structure of English* (1952) is a description of English grammar that took as its data a quarter of a million words of recorded and transcribed telephone conversations. The advent of the computing age revolutionised the means, if not the methods, of corpus analysis. In the late 1950s, work began on huge mainframe computers, with data painstakingly entered by hand on punched cards. In the following years, advances in technology paved the way for such landmarks in corpus linguistics as

- the Brown corpus of 1 million words of written American English published in 1961, created at Brown University, Rhode Island, USA;
- its British English counterpart, the LOB Corpus (Lancaster–Oslo/Bergen Corpus); and
- the work on the Survey of English Usage at University College London in the 1960s, which led to major grammatical descriptions of English (for example, Quirk et al. 1985).

Since the 1980s, corpora have developed in two parallel ways: some have become larger, and others have remained fairly small but become focused on closely delimited types of language, such as child language, the language of learners and the many private corpora created to answer particular research questions. Legal constraints still mean that many corpora are available only to researchers (perhaps for an annual fee, sometimes with password system), but the impetus for this book is that more and more corpora can be accessed immediately and freely on the Internet. There is a long way to go, especially as regards corpora for languages other than English, specialised corpora of particular language varieties and multimedia corpora, but freely accessible online corpora promise to open up hitherto inaccessible riches of language data to the user at much earlier levels of language study. Of course, the user of online corpora is more or less at the mercy of the corpus compilers in terms of the textual content and design of the corpus and the analysis which can be carried out on the corpus through its integrated search tools. For the relative beginner, however, this can be a good thing, as having a limited number of possibilities can make getting to grips with corpus analysis rather less daunting.

It will no doubt have occurred to the reader that there are inevitable uncertainties that are bound to plague any book dealing with online corpora. In particular, there is the risk that particular corpora discussed here will cease to be available as corpus projects come to an end, technical support is no longer available or copyright licences expire. Other online corpora (such as the BYU Corpus of Contemporary American English, which went online literally days before the completion of the first draft of this book) promise to be regularly

updated – which means that search results will change over time. Search tools might be enhanced, change their appearance or even disappear. Moreover, on the plus side, in the current climate, it is likely that new corpora will appear, existing corpora will be expanded and annotated, new portal sites will allow various corpora to be accessed together and online analysis tools will become more sophisticated. Such changes have occurred with regard to several corpora in the course of writing this book. Even so, the kinds of basic principles outlined in the book and the searches we recommend should be easily adaptable to new corpora and new corpus interfaces.

## Five online corpora of English

Below we introduce five of the main freely available online corpora of English, which are frequently referred to in this book. Alongside each we suggest a straightforward task which will allow you to begin to find out more about the nature of these corpora and how they can be analysed. These corpora, along with several others, are also listed in the Appendix.

### British National Corpus

The British National Corpus (BNC; [www.natcorp.ox.ac.uk/](http://www.natcorp.ox.ac.uk/)) – because of its large size, level of annotation and availability – has become the gold standard among corpora of British English. It was completed in 1994, and contains

#### TASK 1.2 Using the British National Corpus

Using your intuition, jot down the words you expect to follow *high and* \_?

1. Go to the BNC site at [www.natcorp.ox.ac.uk/](http://www.natcorp.ox.ac.uk/)
2. Type 'high and' in the 'Look up' box under 'Search the corpus' and hit Return.
3. You will be given a random sample of 50 occurrences of the sequence from the thousand or so examples in the BNC.
4. Go through the examples, and pick out those where *high and* seems to you to be part of a fixed expression. How many expressions can you find? How many had you previously thought of? Did the corpus produce any surprises?
5. Look again at some of the examples which you have decided are not examples of fixed expressions. Can you explain in grammatical terms why you do not consider these to be fixed expressions?
6. Choose a small number of examples and try to identify what sort of text the example has come from. Click on the text ID code at the beginning of each example to see how accurate you were. Assuming that your guesses were at least partially correct, what does this indicate about the language of different sorts of text?

You will find more on performing word searches in online corpora in Chapter 3.

100 million words of British English texts from the late twentieth century. Ten per cent of the corpus is made up of transcribed spoken language of various genres, and ninety per cent is written material, again of a wide range of genres. Importantly, the corpus has been automatically part-of-speech-tagged. The BNC website contains further information and allows simple searches (of strings of characters, or words restricted by part of speech) to be carried out. The search results indicate the total number of hits, although the online system returns a random sample of only 50 examples in the context of the sentence in which they occur. From the search results, the user can click to find out details of the text in which an example occurs.

### **BYU-BNC: The British National Corpus**

Another means of accessing the BNC online is through an interface created and maintained by Mark Davies at Brigham Young University (BYU; <http://corpus.byu.edu/bnc/>). This interface offers a number of additional features over the BNC's own online search system, in particular full results, the possibility of identifying collocates, comparing the use of a word across different registers (for example, spoken, fiction, news) and searching for synonyms. For reasons of copyright, it is not possible to see complete texts, and as with the BNC's own site, the surrounding text (the co-text) of a search word is limited. Given its ease of use, and additional features, BYU-BNC is the online version of the BNC that is used most in this book.

#### **TASK 1.3 Using the BYU-BNC**

From your own experience of language, do you think that passive constructions (for example, *he was followed by a big dog*, *the book was proofread twice*) are more common in spoken language or in written language?

1. Go to the BYU-BNC site at <http://corpus.byu.edu/bnc/>
2. Under 'Display' click 'Chart', and under 'Search String' type 'was \*ed by' in the 'Word(s)' field. Click 'Search'.
3. Look at the resulting chart. In which of the five main registers (Spoken, Fiction, Newspaper, Academic, Miscellaneous) does the sequence have the largest raw frequency?
4. In which of the registers does the sequence have the largest normalised frequency, that is, the greatest frequency of occurrence per million words? Why is this answer different from the answer to Question 3?
5. Do the results confirm or challenge your expectations?
6. What words do you expect to occur most frequently in place of '\*ed' in this search? Again, jot down your intuitions, and then check them against the corpus by selecting 'List' under 'Display' and rerunning the query.

You will find more on exploring grammar in Chapter 4, and more on interpreting quantitative data in Chapter 2.

### BYU Corpus of Contemporary American English

Using the BNC as a model, Mark Davies has also designed an online corpus of American English currently totalling a massive 360 million words (and to be regularly updated; [www.americancorpus.org](http://www.americancorpus.org)). It uses similar search tools to the BNC and so offers a new and fascinating opportunity to compare British and American usages. For copyright reasons, it is again impossible to see complete texts, although short extracts around the search word are available for viewing.

#### TASK 1.4 Using the BYU Corpus of Contemporary American English

From your own experience of language, what adjectives do you expect to follow the adverb *real* in American English? Do you expect such expressions (for example, *real good*, *real important*) to occur more commonly in speech or in writing?

1. Go to the BYU Corpus of Contemporary American English site at [www.americancorpus.org](http://www.americancorpus.org)
2. Under 'Display' click 'Chart', and under 'Search String' type 'real' in the 'Word(s)' field. Click on 'POS List' and select 'adj.ALL' from the drop-down list, so that '[\*]' appears after 'real' in the 'Word(s)' field. Click 'Search'.
3. Look at the resulting chart. In which of the five registers (Spoken, Fiction, Magazine, Newspaper, Academic) is the construction most common?
4. Does there appear to be any change in frequency over time (for example, between the earliest texts in the corpus and the latest)?
5. Have a closer look at some occurrences, by clicking on a bar in the chart, to see a list of the most common instances of the search string. What adjectives occur most commonly with *real* in each register?
6. Compare your results with the British English data in the BYU-BNC. Does British English also make use of *real* as an adverb to qualify adjectives?

You will find more on exploring the connections between language and context in Chapter 7.

### TIME Corpus

This third major online corpus site created by Mark Davies (<http://corpus.byu.edu/time/>) contains more than 100 million words of text from the US magazine *TIME* from 1923 to the present, and is accessible through the same interface as the BYU-BNC and the BYU Corpus of Contemporary American English. However, a novel feature of this corpus is the possibility of looking at frequencies of words as they appear decade by decade, allowing us to trace key cultural concepts, and the rise and dissemination of neologisms over almost a century of American journalism.

**TASK 1.5 Using the TIME Corpus**

Using your experience of language, jot down a list of adjectives which you think are likely to have changed dramatically in frequency or usage over time; that is, adjectives which you think have only become popular recently, or have fallen into disuse or have significantly changed in meaning. Here are a few to get you started: *funky, gay, pious, spooky, spunky*.

1. Go to the TIME Corpus site at <http://corpus.byu.edu/time/>
2. Under 'Display' select 'Chart', and then type one of your adjectives in the 'Word(s)' field. Click 'Search'.
3. Explore the adjective in terms of its frequency – a quantitative analysis. The tallest bar in the chart indicates the register in which the search term occurs with the highest normalised frequency, that is, the greatest frequency per million words (see Chapter 2 for further discussion of normalisation). Is there a clear tendency for your adjective to become more or less frequent, relatively speaking, over time?
4. Explore the adjective in terms of its meaning and common collocates – a form of qualitative analysis. Go back to the search box, and tick 'Show Sections'. In the 'Search String' section, click on 'Context' and use the 'POS List' below to identify nouns in the vicinity of the search word (try five words to each side).
5. Explore how collocation patterns with different nouns have changed over time. Do the collocates suggest a change in meaning in your adjective, or simply a change in use?
6. Go back and repeat Steps 2–5 for each of your adjectives.

You will find more on quantitative and qualitative analysis of corpora in Chapter 2.

**Michigan Corpus of Academic Spoken English**

The Michigan Corpus of Academic Spoken English (MICASE; <http://quod.lib.umich.edu/m/micase/>) contains close to 2 million words of audio recordings and transcripts of academic speech events collected at the University of Michigan, USA. MICASE is an example of a small but very focused corpus of spoken English in a restricted range of settings. The transcripts can be searched according to various criteria such as the academic role of speaker, type of speech event, academic discipline and so on. Complete transcripts can be viewed and also downloaded. Searching for a word displays a concordance view that can be sorted in various ways, and much more context can be shown than a few words on each side of the search item.

**TASK 1.6 Using MICASE**

Using your experience of language, and knowledge of academic situations, think about how you ask questions in the classroom, or in lectures, seminars and one-to-one sessions in a university context. How do you frame your question?

1. Go to the MICASE site at <http://quod.lib.umich.edu/m/micase/>
2. Click 'Search MICASE' and type the word 'question' in the search box. Leave the Speaker Attributes and Transcript Attributes at the default settings. When prompted, choose 'View all results'.
3. Scan the concordance lines to identify instances of *question* where the speaker (whether student or teacher) is using the word to signal that he or she is about to ask a question: for example, *I have a question, my question is, I guess my question for you is*. How many examples can you find?
4. Why do you think the speaker chooses to announce his or her question in this way, rather than just, say, asking the question bluntly? In other words, what purpose does the framing serve? You will probably find it useful in answering this question to read the wider context of some of the occurrences.
5. Spend some time looking more closely at a couple of transcripts. Go back to the Search page, and specify that you want speech events which are 'Highly interactive' in the 'Interactivity Rating' category. What other ways can you find of indicating that you are going to ask a question?

You will find more on exploring features of discourse through online corpora in Chapter 5.

**Scottish Corpus of Texts & Speech**

One of the main online corpora we shall be drawing on in this book is the Scottish Corpus of Texts & Speech (SCOTS; [www.scottishcorpus.ac.uk](http://www.scottishcorpus.ac.uk)). Compiled and maintained by a team at the University of Glasgow, SCOTS includes texts in Scottish English and varieties of contemporary Scots, plus a few texts in Scottish Gaelic. At 4 million words, SCOTS is small compared to the BNC and the BYU Corpus of Contemporary American English, but it offers advantages which in part compensate for its size, depending on the type of research to be undertaken. Four in particular are given below:

- SCOTS includes 800,000 words of spoken text, in audio and, in some cases, audio-visual format. These recordings are made available as audio-visual files (requiring Apple QuickTime™ to view them), with a synchronised orthographic transcription.
- Most of the texts in SCOTS are available as complete texts; they have been copyright-cleared, and while they cannot be republished without permission, they can be downloaded and analysed for educational and research purposes.

- Extensive sociolinguistic metadata is made available with every text and can be used to refine a search.
- The Advanced Search facility in SCOTS allows users to browse available documents, written or spoken, and download them in bulk to their own computer as plain text files. Other search software can then be used on this subcorpus.

In some respects, however, the user must still exercise caution in using the SCOTS resource. The corpus contains a wide range of genres (from spoken conversations and interviews to written prose fiction, poetry, correspondence, documents from the Scottish Parliament and so on), but the corpus is not balanced – that is, the quantities of texts in each genre do not represent the proportions of these which are produced in the respective language varieties. Nor is there a perfect geographical balance, although SCOTS does aim to cover as much as possible of Scotland and a map facility allows the user to see this coverage at a glance. In Teubert and Čermáková's terms (2007, p. 70), SCOTS is an 'opportunistic' corpus.

### TASK 1.7 Using the SCOTS Corpus

Think about the characteristic Scottish features of pronunciation of which you are aware. Jot down some words which you think would help you to identify a speaker's accent as strongly Scottish.

1. Go to the SCOTS site at [www.scottishcorpus.ac.uk](http://www.scottishcorpus.ac.uk). Click on the 'Advanced Search' option.
2. Build up a complex query in the following way. Under 'Select your criteria', click 'Spoken', choose 'Participant details' and 'Region of residence'. From the drop-down list, select 'Moray', and hit Return. Then under 'General', click 'Word search' and 'Word/phrase (concordance)'. Type 'fitt' in the search box, and hit Return.
3. Scroll down the page to the concordance of *fitt*, click on a few examples in turn to be taken to the document page, from which you can listen to the recording. What do you think *fitt* means?
4. Many speakers alternate between *fitt*, *what* and *whit* in the north-east of Scotland. Try the query again using these other forms. Can you find examples where it is actually difficult to say whether the speaker is saying *fitt* or *whit*?
5. Now repeat Steps 2–4 to listen to examples of *what/whit/fitt* by speakers living in other regions of Scotland.
6. Explore pronunciation variation in Scotland further by building up similar search queries based on the words you identified as likely to identify a speaker's accent as Scottish.

You will find more on exploring features of pronunciation through corpora in Chapter 6.

### The DIY option: building your own corpus

In addition to the freely available online corpora detailed above and in the Appendix, you may find that your institution has a subscription to other corpora, which can be used for comparative purposes. It is also possible that you might want to build your own corpus. This might be the case if your interest lies in a very specific type of language, which is not included in any of the corpora described above, or if you want to create a new corpus to act as a point of comparison with an existing one. Additional online corpora can be accessed through the Sketch Engine program, available at [www.sketchengine.co.uk](http://www.sketchengine.co.uk), which can also help you to build your own corpus from web material. As the focus of this book is on using existing online corpora, we shall not dwell here on corpus building. Students seeking detailed advice on this should see McEnery, Xiao and Tono (2006), which contains an excellent unit on 'DIY corpora'.

### Analysing online corpora

The previous section introduced five corpora that can be used as online resources for language study. This section previews the kind of questions that linguists ask, in their exploration of how language works. The corpora we focus on in this book all come with integrated analysis tools of various kinds, so that the corpora can be not only accessed but also interrogated online. The kinds of questions that different readers might ask of online corpora are diverse and wide-ranging. For example,

- How is the word *how* used by speakers in different parts of the world?
- How is reported speech signalled in conversation?
- How are different sounds typically pronounced by speakers in different parts of the English-speaking world?
- What prepositions commonly follow particular words, like *result* in English?

There are various ways to begin to answer these questions and others like them: indeed some are probably best answered using a variety of methods. It is difficult not to start by drawing on your own intuition, built up from years of experience as an expert user of English. Dictionaries and reference grammars may confirm or challenge your instincts. Widening the net beyond your personal expertise, you might devise a questionnaire to be completed by a wide range of native speakers and/or expert users of English as a second or other language.

A further perspective is offered by an appropriate language corpus. As we have seen, a language corpus is – ideally – a carefully designed collection of examples of language which were not originally written or spoken for that purpose. To that extent, at least, a corpus represents 'authentic' language use. In practice, some corpora owe their construction as much to accident as design, and others include language that is more or less spontaneous and more or less 'rehearsed'.

Even so, corpora are rich resources by which we can extend our knowledge of language from personal intuition to a large set of data that shows what people under different circumstances actually say and write.

For example, to answer the questions above, the interested reader could go to a computer with an Internet connection and search an online corpus such as SCOTS, which we mentioned above. Searches of the SCOTS corpus can quickly address some of these questions.

### How is the word *how* used by speakers in different parts of the world?

By searching the SCOTS corpus, in ways that we shall shortly discuss, you will be able to find many different examples of the use of *how* by Scottish speakers. One sample conversation illustrates two possible uses of *how*:

- (1) F643: //This used tae//  
 M608: //uh huh//  
 F643: be a post-office, *that's how we had a post-box*.
- (2) M642: //Aye well that // covers your bare wire  
 F643: //That's right.// But then ye see the rats were eatin the wires as well.//  
 M642: //See, *that's how ye joined them*.

In the first example, *how* expresses a reason, and in other varieties this might well be expressed by *why* (as in *that's why we had a post-box*). In the second example, *how* expresses instrumentality, showing the way the wires were joined. In this case it cannot be paraphrased by *that's why ye joined them*. These two possible uses of *how* are typical of speech in Scotland; a search of other corpora would confirm whether these two uses are also found in other speech varieties.

### How is reported speech signalled in conversation?

The means used to signal reported speech in conversation vary from generation to generation and from place to place. Searches of online corpora such as the SCOTS corpus, again, can show us some of the resources used to mark off reported speech in conversation:

- F1049: //But you know that way, when you're workin,// and then like somebody asks you something, so you go and get something and then before you've come back to them somebody else asks you //something, and *you're like* 'Wait a minute'. [laugh]//
- M1048: //Uh-huh, so you get nothing done.// 'Hang on I'm dealin with a customer just now', and then you just get a look. [laugh]
- F1049: Well I was dealin with this woman and *she had asked where the wine was* but I was on the way to give this vodka to another woman,
- M1048: Uh-huh.
- F1049: and then another *woman asked me where the cold meat was*, //and //

M1048: //Uh-huh.//

F1049: *just went* 'It's down that, it's on aisle twenty-three or whatever it //was',//

M1048: //Yeah.//

F1049: and she went, like that. And *I was like* 'If you want to wait a wee second I'll take you to it, but I've got two customers just now'.

M1048: [laugh]

F1049: And *she went* 'Hm', like that. //And then this other customer//

M1048: //[laugh]//

F1049: 'I'll show you where it is', and *I was //like* 'She'll show you where it is', I don't//

M1048: //[laugh]//

F1049: care, [laugh]

In this conversation, the speakers either report the speech indirectly (for example, *she had asked where the wine was*) or report it directly, using the signals *like* and *went*. This kind of observation is useful to those who are interested in the way speakers dramatise their interactions with others in conversation; it is also useful to those who are learning and teaching conversational strategies.

### How are different sounds typically pronounced by speakers in different parts of the English-speaking world?

Variation in accent is one of the most fascinating topics of language study. Many early corpora focused on written texts, and only in recent years have searchable archives of spoken English become readily available online. These allow you to hear speakers either interacting more or less spontaneously, or reading from a prepared text. Both types of corpus have their uses. The latter type of speech archive allows for easy comparison between different accents because the content of their speech is identical. The former type is more representative of the features associated with everyday speech. Chapter 6 focuses on ways of exploring different types of English pronunciation using online resources.

### What prepositions commonly follow particular words, like *result* in English?

Some corpora contain information about grammatical parts of speech. The BYU Corpus of Contemporary American English, for example, currently allows the browser to search 360 million words of American English for sequences such as *result* + preposition. The most frequent preposition following *result* is *of*, mainly in sequences in which *result* functions as a noun, as in *chemicals that build up as a result of stress*. The second most frequent preposition to follow *result* is *in*, mainly in examples in which *result* functions as a verb, as in *it did result in division within the Cabinet*. Learners of English as a foreign language, for whom prepositional usage is a perpetual challenge, can compare a sample of the 7516 current instances of *result* + *in* in the Corpus of Contemporary

American English with the 60 instances of *result + on*; for example, *that may be the most important result on Sunday*. Here, *on* forms a phrasal unit with *Sunday* rather than with *result*. There are a few instances where *on* seems to follow *result* and indicate the nature of the result; for example, *I am extremely sorry that a poor choice of words on my part in any way would result on dishonour cast upon you*. The extreme rarity of these instances suggests that they might be slips of the tongue rather than genuine examples of possible variant uses.

The questions discussed above give just a taste of the way that corpora can be used to explore the use of English around the world. The corpora give evidence of variation in the use of vocabulary and grammar, and show patterns of usage that will interest the student and be useful to teachers and learners of English. We will pick up some of the topics touched on above and investigate them in greater detail in later chapters.

## The organisation of this book

This introductory chapter has set out some of the essential information needed to begin corpus-based study of language, and sketched a brief history of the empirical exploration of English in the digital age. It has surveyed a number of freely available online corpora, focusing closely on the corpora which will feature most strongly in the case studies and tasks in the central chapters of this book, and discussing the kinds of language project to which each is most suited. The rest of the book is structured in the following way. Chapter 2 introduces you to some of the basic search techniques in corpus studies and discusses the strengths and limitations of qualitative and quantitative analysis. It also establishes the pattern of taking the reader through a number of case studies, each of which can be modified using different linguistic features and different corpora, to provide a means of approaching all sorts of research questions.

Chapter 3 begins our linguistic analysis proper, at the level of the word. This chapter introduces one of the most fundamental tools for corpus analysis, the concordance. Chapter 4 then turns to the grammar and syntax of English, exploring how online corpora can provide new insights in this area. Chapter 5 considers the strengths and weaknesses of corpora in the investigation of linguistic organisation above the level of the sentence, looking at the use of certain linguistic features to organise discourse. Next, Chapter 6 focuses on spoken language, examining how some online corpora and resources, through their multimedia nature, allow us objective insights into phonology, the study of pronunciation.

Chapter 7 brings together some of the issues discussed in Chapters 2–6 and further addresses issues of quantitative and qualitative linguistic analysis using corpus evidence. The case studies here examine the strengths and weaknesses of using whole texts versus extracts in corpus study, and show how linguistic data can be combined with metadata (that is, information about the social background to the speech event and participants) to provide a rich sociolinguistic

analysis. Finally, Chapter 8 considers some developing trends in corpus-based language studies.

---

### FURTHER READING

---

- Adolphs, S. (2006). *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies*. Abingdon, UK: Routledge.
- Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.
- McEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- McEnery, T., Xiao, R. and Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.
- Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

# Index

---

- accents, 125–52, 193  
 American, 126, 129, 133–5, 140  
 attitudes to, 135–6  
 English, 133–4, 138, 140–1  
 native and non-native, 136–7  
 rhotic and non-rhotic, 132–5  
 Scottish, 131–52
- acquisition, 147–9
- adjacency pairs, 112–13, 193
- adjectives, 71, 74–6
- Adolphs, Svenja, 114, 163
- adverbials, 83, 87, 89, 193
- adverbs, 71, 74–6, 116  
 conjunctive, 118–20
- allusions, 121–2
- Anderwald, Lieselotte, 161–2
- annotation, 8, 66, 106–7, 193
- ANOVA (Analysis of Variance), 43, 193
- anthropology, 9
- armchair linguists, 178–9
- assimilation, 143–5
- auxiliary verbs, *see* verbs
- average, *see* mean
- Baker, Paul, 1, 31, 37, 42–3, 105
- balance, 15
- Bazerman, Charles, 121
- Biber, Douglas, 23–8, 105
- Blackhall, Sheena, 151
- Braun, Sabine, 159–60, 163
- Brigham Young University British National  
 Corpus (BYU-BNC), *see* British  
 National Corpus (BNC)
- Brigham Young University Corpus of  
 Contemporary American English  
 (BYU-COCA), 3, 7–9, 12, 14, 18–19,  
 106, 109–11, 113, 116, 120, 179,  
 181, 184
- British Academic Spoken English Corpus  
 (BASE), 124, 183
- British Academic Written English Corpus  
 (BAWE), 183
- British National Corpus (BNC), 2, 3, 6, 8,  
 10–12, 14, 26–7, 33–41, 52–62, 64,  
 69–70, 72–7, 81–98, 107, 109–11,  
 113, 116, 120, 159–63, 175, 179,  
 181, 183
- Brown Corpus, 9, 109, 156–7
- Cambridge International Corpus (CIC),  
 29–31
- Cambridge and Nottingham Corpus of  
 Discourse in English (CANCODE),  
 29–31, 108
- Carter, Ronald, 6–7, 29–31, 67, 98–9, 108,  
 173, 179
- Čermáková, Anna, 15
- Chambers, Angela, 7
- children's language, *see* acquisition
- chi-square, 37, 43–4, 193
- Chomsky, Noam, 99, 178
- clauses, 83–91, 199  
 coordinate, 88–9, 194  
 with copular verbs, 84  
 non-finite, 88, 89–90  
 relative, 88, 90–1, 198  
 subordinate, 89–91, 199
- CLAWS, 53
- Cobuild Concordance and Collocations  
 Sampler, 57–8, 60
- coherence, 116–20, 193
- cohesion, 116–20, 193  
 conjunctive, 118–20  
 lexical, 118, 120
- colligation, 46, 52, 58–9, 93–6
- Collins WordbanksOnline, 184
- collocation, 13, 32–7, 46, 49–62, 193

- comparable corpora, 8, 193  
 competence, 99–100, 178, 193  
 Compleat Lexical Tutor, 184  
 complement, 83–7, 194  
   object complement, 85  
 complex sentences, *see* sentence types  
 concordance, 32, 45, 48–52, 157–8, 194  
 conjunctions, 71, 88–9, 116, 118–20  
   coordinating, 88–9  
   subordinating, 89  
 consonants, 125–9, 132–7  
 context, 155–72, 194  
   of situation, 158  
 copyright laws, 181, 194  
 core language, 7  
 Corpus of Modern Scottish Writing  
   1700–1945 (CMSW), 181  
 co-text, 155–6, 158, 194
- data-driven grammars, 99–100, 194  
 Davies, Mark, 11, 12, 183, 186  
 determiners, 71  
 diachronic corpora, 7, 181, 194  
 dialectology, 9  
 discourse, 101–23, 194  
   communities, 105  
   spoken, 102–5, 111–16  
   written, 102–5, 116–20  
 discourse markers, 30–1, 108, 194  
 discourse prosody, *see* semantic prosody  
 dispersion, 43, 194  
 Douglas, Fiona, 114  
 Durham, Mercedes, 149
- elision, 143, 145–6, 195  
 English Language Interview Corpus as a  
   Second-Language Application  
   (ELISA), 46–7, 124, 153, 156, 162,  
   163, 168, 184
- field, 169–72, 195  
 Fillmore, Charles, 178–9  
 Firth, John Rupert, 52, 158  
 Fitt, Matthew, 40–2, 146  
 Fortune, Liane, 149  
 Fowler, Henry Watson, 52
- Francis, Gill, 83  
 Freiburg–Brown Corpus (FROWN),  
   156  
 Freiburg English Dialect Corpus (FRED),  
   185  
 Freiburg–LOB Corpus (FLOB), 156  
 frequency, 28–32, 44  
   raw frequency, 28–30, 42–3, 48  
 Fries, Charles, 9
- Gadamer, Hans-Georg, 111–12  
 genre, 63–4  
 geographical variation, 65–6  
 GlossaNet, 185  
 Grabe, Esther, 153  
 grammar, 9, 16–17, 18–19, 58–9, 67–100,  
   195  
   grammatical items, 71, 76–8  
   pattern grammar, 83  
   spoken and written grammars, 91–3
- Halliday, Michael Alexander Kirkwood, 53,  
   59, 116, 118, 169–71  
 Hardie, Andrew, 1  
 Hasan, Ruqaiya, 59, 116, 118, 169–71  
 headword, 83, 195  
 Hedderwick, Mairi, 83  
 hesitations, 106–8  
 Hewlett, Nigel, 143  
 Hockey, Susan, 2  
 Hoey, Michael, 93–4, 118  
 Hunston, Susan, 83
- idiom principle, 55, 57, 195  
 idioms, 46, 195  
   *see also* multiword units  
 intellectual property law, 181  
 interjections, 71  
 International Corpus of English (ICE), 109,  
   157  
 international phonetic alphabet, 126  
 International Phonetic Association (IPA),  
   126, 195  
 intertextuality, 121–3  
 intonation, 152–3, 195  
 Intonation Variation in English corpus  
   (IViE), 66, 152–3, 185

- Kay, Christian, 103–4  
key words, 37–42, 195
- Lancaster-Oslo/Bergen Corpus (LOB), 9, 156
- language change, 65
- Leech, Geoffrey, 154, 179
- Leibniz, Gottfried, 121–2
- lemma, 8, 46–8, 66, 196
- lexicalisation, 149–52
- lexicogrammar, 59, 67, 196
- lexicography, 2, 6, 8, 9, 46, 96, 196
- lexis, 45–66  
‘borrowings’, 173–5  
infrequent, 56  
lexical items, 71, 72–6  
semantically related, 118  
‘vocabulary 3’, 118–20, 200
- Lexware Culler corpora, 185
- log-likelihood, 21, 37–44, 196  
online log-likelihood calculator, 38
- Louw, Bill, 62
- Macafee, Caroline, 149
- McCarthy, Michael, 6–7, 29–31, 67, 98–9, 108, 173, 179
- McEnery, Tony, 16, 21–2, 154, 163
- MacMahon, Michael, 126–30
- Malinowski, Bronislaw, 158
- Matthews, Ben, 143
- maven, 68, 71, 76
- mean, 24–8, 43, 196
- metadata, 15, 63, 136, 156, 158–64
- metaphors, 46, 57, 196
- Meyer, Charles F., 179
- Michigan Corpus of Academic Spoken English (MICASE), 4, 7, 13–14, 64, 78–9, 110, 112–13, 116, 124, 159, 161, 170, 176, 179–81, 185
- mode, 63–4, 169–72, 196
- modifier, 83
- monitor corpus, 7–8
- Moon, Rosamund, 54
- multimedia corpora, 181, 196
- multiword units, 46, 52–5  
multi-word verbs, 76–7, 175, *see also* verbs
- Mutual Information (MI), 32–4, 37, 42–3, 57–8, 196
- node, 48, 49, 51, 197
- non-standard varieties of English, 180
- normalisation, 13, 25, 30–2, 37, 42–3, 48, 197
- nouns, 69–79, 119
- object, 83–7, 197  
direct and indirect objects, 86
- O’Keeffe, Anne, 6–7, 29–31, 98–9, 108, 173, 179
- Omniglot, 131
- open choice principle, 57, 197
- Oxigen, 152
- parallel corpora, 8, 180, 197
- participles, 88–90
- part-of-speech (POS), 71–2, 106, 197  
POS searches, 72–6
- passives, 98–9
- Penn Treebank Project, 83
- performance, 99–100, 178, 197
- phonemes, 132, 152
- phonetics, 125–31
- phonology, 125, 143–6
- phrasal verbs, *see* verbs
- phrases, 78–83  
adjective phrases, 81–2  
adverb phrases, 81–2  
noun phrases, 78–9, 90–1  
prepositional phrases, 79  
verb phrases, 79–80
- PolyU Language Bank, 186
- Popper, Karl, 5
- Post, Brechtje, 153
- pragmatics, 113–16
- predicator, 83–7, 197
- prepositions, 71, 73–4, 76–7
- process-oriented teaching, 176–8
- product-oriented teaching, 176–8
- pronouns, 71, 77–8, 116, 118
- pronunciation, 15, 18, 124–54
- Purves, David, 30
- qualitative analysis, 13, 22, 28, 30, 37, 41–2, 48, 120, 122, 163, 166, 197

- quantitative analysis, 13, 21–45, 48, 66,  
163, 166, 197
- quotations, 121–2
- raw frequency, *see* frequency
- Rayson, Paul, 38, 43
- reference corpus, 37–42
- register analysis, 169–72, 198  
*see also* field; mode; tenor
- reported speech, 17–18, 107
- representativeness, 5–6, 22–8, 44
- Safire, William, 68
- Scobbie, James M., 143
- Scots language, 14–15, 17–18, 38–41, 100,  
102–3, 112, 114–15, 126–7, 129,  
131–6, 132–52
- Scottish Corpus of Texts and Speech  
(SCOTS), 4–5, 14–15, 17–18, 30–2,  
38–41, 48–51, 54, 63–4, 66, 80–6,  
92–3, 102–4, 110–20, 124, 131–6,  
157–60, 162–8, 171, 176, 180–1,  
186
- Scottish English, *see* Scots language
- Scottish Gaelic, 14
- Scottish Parliamentary discourse, 30, 32,  
119
- Scott, Michael, 41–2
- semantic preference, 52, 59–60
- semantic prosody, 46, 52, 60–2
- semi-fixed expressions, 55  
*see also* multiword units
- sentence types, 88–91
- sequencers, 110
- Shakespeare, William, 121–2
- significance, statistical, 43, 199
- Sinclair, John McHardy, 3, 57
- size of corpora, 6–7, 37
- Sketch Engine, 16
- Smith, Jennifer, 147, 149
- sociolinguistics, 159–68, 198
- Speech Accent Archive, 136–43, 186
- standard deviation, 24–8, 43, 199
- Stefanowitsch, Anatol, 44
- subject, 83–7, 199
- subject-predicator-object-complement-  
adverbial (SPOCA), 83–7, 199
- Survey of English Usage, 9
- synchronic corpora, 7, 199
- synonyms, 56, 58, 117
- tagging, 8, 53–5, 66, 83, 88, 106,  
199
- Tagliamonte, Sali A., 159, 161
- tag questions, 111
- teaching applications, 173–8
- tenor, 169–72, 199
- Teubert, Wolfgang, 15
- text archives, 4, 109
- TIME corpus of American English, 4, 7–8,  
12–13, 56, 65, 121–2, 171, 174, 181,  
186
- Tognini-Bonelli, Elena, 157, 159
- Tono, Yukio, 16, 21, 158, 163
- translation, 8
- Truss, Lynne, 68
- t-test, 43, 57–8, 60, 199
- use-related variables, 169–72
- user-related variables, 165–8
- verbs, 69–75, 83–7, 119  
auxiliary, 71, 80–1, 100  
copular, 84  
delexicalised, 92–3, 194  
ditransitive, 86, 194  
intransitive, 86–7, 195  
non-finite, 89–90  
phrasal/multi-word, 76–7, 196  
transitive, 85, 199  
*see also* phrase
- verb systems, 96–9  
aspect, 96–8, 193  
voice, 96, 98–9, 200
- Vienna-Oxford International Corpus of  
English (VOICE), 180
- Virtual Language Centre Web  
Concordancer, 186
- vocabulary, *see* lexis
- Voltaire (François-Marie Arouet),  
121–2

- vowels, 125, 129–31, 137–9
  - diphthongs, 139–40
  - length, 140–3
  - monophthongs, 139–40
  - Scottish Vowel Length Rule (SVLR),  
141–3
  - unstressed, 130–1, 143–4
  - voicing effect (VE), 141
- weak forms, 143
- WebCorp, 4, 65, 111, 168
- Widdowson, Henry George, 63, 158, 175
- Winter, Eugene, 118–19
- word categories, 71
  - see also* lexis
- word forms, 8, 46–8, 66, 200
  - see also* lemma
- Wordsworth, Dot, 67–8, 100
- World Englishes, 180
- Wynne, Martin, 154
- Xiao, Richard, 16, 21, 158, 163