

1 Introducing corpora and corpus analysis tools

Simply speaking, corpus linguistics is an approach or a methodology for studying language use. It is an empirical approach that involves studying examples of what people have actually said, rather than hypothesizing about what they might or should say. As we will see, corpus linguistics also makes extensive use of computer technology, which means that data can be manipulated in ways that are simply not possible when dealing with printed matter. In this chapter, you will learn what a corpus is and you will read about some different types of corpora that can be used for various investigations. You will also get a brief introduction to some of the basic tools that can be used to analyse corpora. Finally, you will find out why corpora can be useful for investigating language, particularly LSP.

What is a corpus?

As you may have guessed, corpus linguistics requires the use of a corpus. Strictly speaking, a corpus is simply a body of text; however, in the context of corpus linguistics, the definition of a corpus has taken on a more specialized meaning. A corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria. There are four important characteristics to note here: ‘authentic’, ‘electronic’, ‘large’ and ‘specific criteria’. These characteristics are what make corpora different from other types of text collections and we will examine each of them in turn.

If a text is authentic, that means that it is an example of real ‘live’ language and consists of a genuine communication between people going about their normal business. In other words, the text is naturally occurring and has not been created for the express purpose of being included in a corpus in order to demonstrate a particular point of grammar, etc.

A text in electronic form is one that can be processed by a computer. It could be an essay that you typed into a word processor, an article that you scanned from a magazine, or a text that you found on the World Wide Web. By compiling a corpus in electronic form, you not only save trees,

you can also use special software packages known as corpus analysis tools to help you manipulate the data. These tools allow you to access and display the information contained within the corpus in a variety of useful ways, which will be described throughout this book. Essentially, when you consult a printed text, you have to read it from beginning to end, perhaps marking relevant sections with a highlighter or red pen so that you can go back and study them more closely at a later date. In contrast, when you consult a corpus, you do not have to read the whole text. You can use corpus analysis tools to help you find those specific sections of text that are of interest – such as single words or individual lines of text – and this can be done much more quickly than if you were working with printed text. It is very important to note, however, that these tools do not interpret the data – it is still your responsibility, as a linguist, to analyse the information found in the corpus.

Electronic texts can often be gathered and consulted more quickly than printed texts. To gather a printed corpus, you would probably have to make a trip to the library and then spend some time at the photocopier before heading home to sit down and read through your stack of paper from beginning to end. In contrast, with electronic resources such as the Web at your disposal, you can search for and download texts in a matter of seconds and, with the help of corpus analysis tools, you can consult them in an efficient manner, focusing in on the relevant parts of the text and ignoring those parts that are not of interest. Because technology makes it easier for us to compile and consult corpora, electronic corpora are typically much larger than printed corpora, but exactly how large depends on the purpose of your study. There are no hard and fast rules about how large a corpus needs to be, but we will come back to this issue when discussing corpus design in more detail in Chapter 3. Basically though, ‘large’ means a greater number of texts than you would be able to easily collect and read in printed form.

Finally, it is important to note that a corpus is not simply a random collection of texts, which means that you cannot just start downloading texts haphazardly from the Web and then call your collection a ‘corpus’. Rather, the texts in a corpus are selected according to explicit criteria in order to be used as a representative sample of a particular language or subset of that language. For example, you might be interested in creating a corpus that represents the language of a particular subject field, such as business, or you might be interested in narrowing your corpus down even further to look at a particular type of text written in the field of business, such as company annual reports. As we will see, the criteria that you use to design your corpus will depend on the purpose of your study, and may include things like whether the data consists of general or specialized language or written or spoken language, whether it was produced during a narrow time frame or spread over a long period of time, whether it was produced by men or women, children or adults, Canadians or Irish people, etc.

Who uses corpora?

Corpora can be used by anyone who wants to study authentic examples of language use. Therefore, it is not surprising that they have been applied in a wide range of disciplines and have been used to investigate a broad range of linguistic issues. One of the earliest, and still one of the most common, applications of corpora was in the discipline of **lexicography**, where corpora can be used to help dictionary makers to spot new words entering a language and to identify contexts for new meanings that have been assigned to existing words. Another popular application is in the field of language learning, where learners can see many examples of words in context and can thus learn more about what these words mean and how they can be used. Corpora have also been used in different types of socio-linguistic studies, such as studies that examine how men and women speak differently, or studies comparing different language varieties. Historical linguists use corpora to study how language has evolved over time and, within the discipline of linguistics proper, corpora have been used to develop corpus-based grammars. Meanwhile, in the field of computational linguistics, example-based machine translation systems and other natural language processing tools also use corpus-based resources. Throughout the remainder of this book, you will learn about other applications of corpora, specifically LSP corpora, in disciplines such as terminology, translation and technical writing.

Are there different types of corpora?

There are almost as many different types of corpora as there are types of investigations. Language is so diverse and dynamic that it would be hard to imagine a single corpus that could be used as a representative sample of all language. At the very least, you would need to have different corpora for different natural languages, such as English, French, Spanish, etc., but even here we run into problems because the variety of English spoken in England is not the same as that spoken in America, Canada, Ireland, Australia, New Zealand, Jamaica, etc. And within each of these language varieties, you will find that people speak to their friends differently from the way they speak to their friends' parents, and that people in the 1800s spoke differently from the way they do nowadays, etc. Having said this, it is still possible to identify some broad categories of corpora that can be compiled on the basis of different criteria in order to meet different aims. The following list of different types of corpora is not exhaustive, but it does provide an idea of some of the different types of corpora that can be compiled. Suggestions for how you can go about designing and compiling your own corpora that will meet your specific needs will be presented in Chapters 3 and 4.

General reference corpus vs special purpose corpus: A general reference corpus is one that can be taken as representative of a given language as

a whole and can therefore be used to make general observations about that particular language. This type of corpus typically contains written and spoken material, a broad cross-section of text types (e.g. a mixture of newspapers, fiction, reports, radio and television broadcasts, debates, etc.) and focuses on language for general purposes (i.e. the language used by ordinary people in everyday situations). In contrast, a special purpose corpus is one that focuses on a particular aspect of a language. It could be restricted to the LSP of a particular subject field, to a specific text type, to a particular language variety or to the language used by members of a certain demographic group (e.g. teenagers). Because of its specialized nature, such a corpus cannot be used to make observations about language in general. However, general reference corpora and special purpose corpora can be used in a comparative fashion to identify those features of a specialized language that differ from general language. This book will focus on special purpose corpora that have been designed to help LSP learners, hence we often refer to the corpora in this book as *LSP corpora*. If you would like to find out more about general reference corpora, look at resources such as Aston and Burnard (1998) and Kennedy (1998).

Written vs spoken corpus: A written corpus is a corpus that contains texts that have been written, while a spoken corpus is one that consists of transcripts of spoken material (e.g. conversations, broadcasts, lectures, etc.). Some corpora, such as the British National Corpus, contain a mixture of both written and spoken texts. The focus in this book will be on written corpora, but if you would like to know more about spoken corpora and the challenges involved in transcribing texts, please refer to Leech et al. (1995).

Monolingual vs multilingual corpus: A monolingual corpus is one that contains texts in a single language, while multilingual corpora contain texts in two or more languages. Multilingual corpora can be further subdivided into **parallel** and comparable corpora. Parallel corpora contain texts in language A alongside their translations into language B, C, etc. These are described in detail in Chapter 6. Comparable corpora, on the other hand, do not contain translated texts. The texts in a comparable corpus were originally written in language A, B, C, etc., but they all have the same communicative function. In other words, they are all on the same subject, all the same type of text (e.g. instruction manual, technical report, etc.), all from the same time frame, etc.

Synchronic vs diachronic corpus: A synchronic corpus presents a snapshot of language use during a limited time frame, whereas a diachronic corpus can be used to study how a language has evolved over a long period of time. The work discussed in this book will be largely synchronic in nature, but more information on diachronic corpora can be found in Kytö et al. (1994).

Open vs closed corpus: An open corpus, also known as a monitor corpus, is one that is constantly being expanded. This type of corpus is

commonly used in lexicography because dictionary makers need to be able to find out about new words or changes in meaning. In contrast, a closed or finite corpus is one that does not get augmented once it has been compiled. Given the dynamic nature of LSP and the importance of staying abreast of current developments in the subject field, open corpora are likely to be of more interest for LSP users.

Learner corpus: A **learner corpus** is one that contains texts written by learners of a foreign language. Such corpora can be usefully compared with corpora of texts written by native speakers. In this way, teachers, students or researchers can identify the types of errors made by language learners. Granger (1998) provides more details about learner corpora.

Are there tools for investigating corpora?

Once a corpus has been compiled, you can use corpus analysis tools to help with your investigations. Most corpus analysis tools come with two main features: a feature for generating word lists and a feature for generating concordances. We will introduce these tools here briefly so that you have some idea what they can do, but they will be examined in greater detail in Chapter 7.

A word lister basically allows you to perform some simple statistical analyses on your corpus. For instance, it will calculate the total number of words in your corpus, which is referred to as the total number of 'tokens'. It will also count how many times each individual word form appears; each different word in your corpus is known as a 'type'. The words in the list can be sorted in different ways (e.g. in alphabetical order, in order of frequency) to help you find information more easily. Figure 1.1 shows two extracts from a word frequency list taken from a 5000-word corpus of newspaper articles about fast foods. The list on the left is sorted alphabetically while the one on the right is sorted in order of frequency.

A concordancer allows the user to see all the occurrences of a particular word in its immediate contexts. This information is typically displayed using a format known as keyword in context (KWIC), as illustrated in Figure 1.2. In a KWIC display, all the occurrences of the search pattern are lined up in the centre of the screen with a certain amount of context showing on either side. As with word lists, it is possible to sort concordances so that it becomes easier to identify patterns.

Why use corpora to investigate language?

Now that we understand what corpus linguistics is and how we can generally go about it, let us explore some of the reasons why we might want to use corpora to investigate language use. There are, of course, other types of resources that you can use to help you learn more about LSPs.

a	104	the	224
abdomen	2	and	132
abdominal	1	of	129
about	12	to	112
absolutely	1	a	104
academy	1	in	80
accelerating	1	food	71
accommodated	1	is	58
according	4	for	54
acid	1	far	51
acids	1	fast	50
across	3	that	47
action	1	are	42
actively	1	you	41
activists	1	MG	40
activity	1	more	39
adapt	1	than	39
add	2	calories	38
added	5	with	37
adding	1	at	35

Figure 1.1 Extracts from a word frequency list sorted alphabetically and in order of frequency

For example, you have no doubt consulted numerous dictionaries, and you may also have consulted printed texts, or asked a subject field expert for help. You may even have relied on your intuition to guide you when choosing a term or putting together a sentence. All of these types of resources may provide you with some information, but as you will see, corpora can offer a number of benefits over other types of resources. This is not to say, of course, that corpora are perfect or that they contain all the answers. Nevertheless, we think you will find that a corpus can be a valuable resource and a useful complement to other types of resources, such as dictionaries, printed texts, subject field experts and intuition. In the following sections, we will outline some of the shortcomings of these resources for LSP learners and then look at some of the ways in which corpora can help to overcome these drawbacks.

sser, who believes promoting fast food to kids should be ba
ys Eric Schlosser, author of *Fast Food Nation*, out this mon
he inexpensive appearance of fast food is an illusion." Sti
The simple answer is, yes, a fast food meal can be okay. Th
s, you pull into the nearest fast food restaurant for a qui
u or your kids are eating at fast food restaurants more tha
d to the commercial kitchen, fast food was born, and with I
althy diet. To help you make fast food choices and be an in
said is not new. Though the fast-food giant has been sayin
time before consumers force fast-food chains to tame their

Figure 1.2 A concordance for 'fast'

Dictionaries

One of the main types of resource used by LSP learners is a printed dictionary, and although dictionaries can be invaluable for solving some types of language learning problems, they are not always sophisticated enough to meet all the needs of an LSP learner.

One of the biggest problems associated with dictionaries is their inherent incompleteness. The world around us and the language used to describe it are evolving all the time, which means that printed dictionaries go out of date very quickly. It takes a long time to compile and publish a dictionary, and fields such as science and technology, for example, frequently evolve so quickly that by the time a dictionary is printed, it no longer reflects the current state of knowledge or language.

Another problem with dictionaries is their size. Although it is possible to compile large, multi-volume dictionaries that attempt to cover a specialized subject field in its entirety, not many people will be able to afford such dictionaries, and they certainly would not want to carry them around! Most users would prefer to have a dictionary that will fit in their rucksack, which means that the lexicographers who create the dictionaries have to choose which information to include and which to leave out. Unfortunately, their choices do not always correspond with the needs of LSP users. For example, acronyms and other abbreviated forms (e.g. 'ISP' for 'Internet Service Provider') are a common feature in many LSPs, but these are frequently omitted from dictionaries. We saw above that some new terms do not make it into the dictionary, but a related problem is that out-of-date terms are not always taken out. Dictionaries often contain 'linguistic deadwood' – terms that remain in the dictionary even though

they have dropped out of current usage. This means that an LSP learner who consults a dictionary for help may find and use terms that are no longer appropriate.

Another of the most common criticisms of dictionaries is that they do not provide enough in the way of contextual or usage information. LSP learners must pay attention to how terms are used, which means that in addition to information about what a term means, they also need information about how to use that term in a sentence (e.g. what other words 'go with' the term in question). This information can be provided by presenting terms in context instead of in isolation. Although most lexicographers and users alike would agree that the inclusion of contextual fragments in a dictionary can go some way towards meeting these needs, most dictionaries still do not provide this type of information. As we have seen, one reason for this is because people do not want to have large dictionaries. Given the space restrictions imposed on printed dictionaries, even those dictionaries which do contain usage examples only have room for a few per entry.

An additional point of interest with regard to the limitations of most dictionaries is that they cannot easily provide information about how frequently a given term is used. Texts are better guides to naturalness, as determined by frequency or lexical variety, than either dictionaries or intuition. Of course, for LSP users, decisions regarding the appropriateness of a given lexical choice will be made on the basis of more than just frequency; nevertheless, frequency data can help you to make informed decisions.

Finally, even if the relevant information is contained in the dictionary, users sometimes have trouble knowing where to find it. For example, in a single dictionary, users may find that the entries for some terms are listed under the acronym (e.g. under 'ISP' instead of 'Internet service provider'; under 'HTML' instead of 'hypertext markup language'), whereas the entries for other terms are listed under the expanded term, even though such terms might be better known by the acronym (e.g. under 'random access memory' instead of 'RAM'; under 'central processing unit' instead of 'CPU').

Printed texts

Given some of the shortcomings of dictionaries described above, many LSP users turn to printed texts to find up-to-date details about the meaning and use of specialized terms. You may have done this yourself, for instance. You may have read books or articles that have been written about the specialized field in question, and such material has probably provided you with a number of examples of the terminology and style that are appropriate to the LSP you are studying.

Nevertheless, consulting texts in their conventional printed form presents a number of pitfalls. As previously mentioned, in order to physically gather

together a printed corpus, you may need to spend hours at the library and/or photocopier. Once the printed corpus is gathered, further hours must be spent consulting the texts, which often means reading lots of irrelevant material before stumbling upon a discussion of a pertinent point. Furthermore, in order to find a variety of examples and to make sure that the style or terms you choose are generally accepted by experts in the field and not simply the idiosyncratic usage of a single author, it is necessary to consult a selection of texts, not just one or two. Unfortunately, it can sometimes be difficult to detect patterns of linguistic and stylistic generality when they are spread over several documents. Therefore, acquiring and consulting texts in printed form has two major disadvantages. The first is that you typically cannot gather and consult a wide enough range of documents to ensure that all relevant concepts, terms and linguistic patterns will be present. Second, the analysis of printed texts is inherently error-prone: the unaided human mind simply cannot discover all the significant patterns, let alone organize them in meaningful ways.

Subject field experts

As noted above, one of the first things that users learn about dictionaries is that even the best ones do not contain the answers to all their questions. With luck and good research skills, users may be able to find the necessary answers in printed texts, but failing that, they may eventually decide to turn to a subject field expert for help. Subject field experts are people who have received training in and/or work in a specialized subject field, and who are therefore familiar with the LSP used to communicate in that field. Subject field experts can be extremely valuable resources; however, they are typically very busy people, and they may not be able to drop everything to answer your questions in time for you to meet your deadlines.

Even when experts do make themselves available, LSP learners are then faced with the delicate task of eliciting the necessary information from these experts. On the one hand, LSP learners are looking for expert advice because they themselves are not specialists in the subject field, but on the other hand, if the LSP learners do not formulate their questions well, then the experts may unintentionally give advice that is inappropriate to the text in question. When consulting experts, LSP learners must be very careful not to ask leading questions or they risk getting distorted information. Sometimes experts may find it very difficult to spontaneously suggest terms or expressions; however, they are often able to confidently verify or reject suggestions that are put to them.

One final difficulty that LSP learners face when consulting a subject field expert is that they are only getting one person's opinion. Different experts may give conflicting advice on concepts and terms (e.g. they may subscribe to different schools of thought on a subject). Therefore, ideally,

LSP learners would benefit from consulting multiple experts in a given subject field, though in practice, this is not always a realistic aim.

Intuition

An additional resource that you may use as an LSP learner is your own intuition, particularly with regard to language and style. However, this can sometimes lead to problems. As we will see in Chapter 2, some people are learning an LSP in a foreign language, and they may experience some **interference** from their native language. For example, a native French speaker who is learning the LSP of optical scanning technology in English may know that the term for '*tête*' is 'head' and the term for '*numériseur*' is 'scanner', but when it comes to putting a sentence together, s/he may rely on intuition to come up with a structure that uses English words but is based on French syntax (e.g. '*tête de numériseur*' translated as 'head of the scanner' instead of 'scan head'). Even people who are learning an LSP in their native language may experience interference because the structures used in an LSP may not be the same as those used in general language. For example, in legal documents, such as wills or contracts, ideas are often expressed using structures that are longer and more complex than those used in general language. Whereas in general language you might say, 'When I die, I would like my children to inherit my house', this same idea might be expressed in legal LSP in the following way: 'I give my entire interest in the real property which was my residence at the time of my death, together with any insurance on such real property, but subject to any encumbrances on said real property, equally to my two children.' In short, relying on intuition that may be useful in another language or in LGP may lead to errors when attempting to communicate in the LSP in question.

Corpora

We will now look at how corpora can help to overcome some of the drawbacks of the resources mentioned above. The first thing to note is that the physical constraint imposed on printed media does not apply to electronic media such as corpora: hundreds of thousands of words of running text can be stored on a diskette and millions can fit easily on to a hard drive or optical disk. Therefore, corpora have the potential to be more extensive than other resources. In addition, their electronic form means that they are easier to update than printed resources, and they are also easier to consult. As we will see in Chapter 7, using specially designed software, data can be searched much more comprehensively in electronic form than in printed form. Searching for a word or phrase in a printed text is a labour-intensive and time-consuming task. In contrast, conducting a full-text search of a corpus can be done in seconds. Moreover, features

such as wildcard searches (e.g. using the search string *print** to retrieve *print*, *printed*, *printer*, *printers*, *printing*, *prints*, etc.) make it possible to conduct exhaustive searches without exhausting the researcher.

Another strength of corpora is that they contain a wealth of authentic usage information. Since corpora are comprised of texts that have been written by subject field experts, LSP learners have before them a body of evidence pertaining to the function and usage of words and expressions in the LSP of the field. Moreover, with the help of corpus analysis tools, you can sort these contexts so that meaningful patterns are revealed. In addition, a corpus can give an LSP learner a good idea of how a term or expression *cannot* be used. The discovery that certain words or uses of words do not occur in a corpus of authentic texts written by subject field experts can be invaluable for helping learners to establish that even though words may appear in dictionaries, they cannot be used in certain contexts, and that even if a sentence is grammatical, it may not be idiomatic in the LSP in question.

Frequency information is another type of data that is much more easily obtainable when using an electronic corpus and corpus analysis tool. Knowledge about frequency allows you to analyse the lexical patterns associated with words in a more objective and consistent way, but such observations are difficult to make when working with printed documents since the human eye may simply not notice a pattern when its occurrences are spread over several pages or documents. As we saw on page 14, the word lister feature of a corpus analysis tool makes it possible to calculate how many times a given word appears in a corpus.

An LSP corpus basically contains thousands of words that have been written by subject field experts and, as such, it can be seen to represent distilled expert knowledge. Obviously, actual subject field experts cannot make themselves constantly available for consultation; however, once compiled, a corpus is constantly available to the LSP learners, and they can consult it as often as they like. The unrestricted availability of the corpus is important because language learning goes on all the time. In addition, the fact that a corpus contains articles written by many subject field experts means that LSP learners have access to more than one expert opinion, which means they are better able to judge whether terms or expressions are generally accepted in the subject field, or whether they are simply the preference of one particular expert. Finally, LSP learners do not need to worry about asking leading questions since all the evidence contained in the corpus is available to the user.

As an LSP learner, you will probably always use your intuition to a certain extent; however, a corpus can provide you with a means of backing up this intuition. One common reason that LSP learners turn to external resources is for reassurance. A corpus can be seen as a testbed that you can use to verify or reject your hypotheses about the LSP that you are learning. This approach is more reliable than assuming that LSP or native-

language norms can be transferred directly into the LSP. Unlike judgements based on intuition, naturally occurring data has the principal benefit of being observable and verifiable by all who examine it. This means that a corpus can act as an objective frame of reference.

An additional advantage of the corpus-based approach is that it is more efficient to consult a single corpus-based resource than multiple types of conventional resources. As a busy LSP learner who is probably working to tight deadlines, you may not have the luxury of being able to spend hours and hours searching for material that will help you with your task. Therefore, the fewer resources that you are required to consult, the better. A corpus is a single yet broad-ranging resource that can meet the majority of your LSP needs: because a corpus consists of naturally occurring running text, you can retrieve information about both lexical and non-lexical (e.g. style, punctuation, grammar, register) elements of language. In other words, a corpus is a 'one-stop shop'. With the help of an LSP corpus, you can spend less time looking for reference material and more time studying it.

Using a corpus can be an enjoyable as well as an informative experience. In our experience, we have found that learners often find it tiresome and frustrating to consult conventional resources and are excited by the possibility of using electronic resources. Corpus-based resources are not only more interesting for learners to use, but they can also be effectively employed to teach research skills and to improve LSP proficiency. Corpora can be used not only to find answers to questions, but also to prompt discussions about students' work or other interesting issues. Corpora have been frequently known to reveal aspects of language that neither the teacher nor the students would ever have thought of investigating!

In conclusion, we should repeat that corpora may not provide the answers to all your questions, and they should not necessarily be seen as a replacement for all other types of resources. Instead, they can be viewed as a complementary resource that can be used in conjunction with other types of resources. For example, intuition or dictionary use might lead you to come up with a hypothesis that can be further investigated in a corpus, or an investigation in a corpus might provide you with information that helps you to formulate some questions for a subject field expert.

Key points

- Corpus linguistics is an approach or a methodology for studying language use.
- A corpus is a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria. These criteria depend on the nature and purpose of the project at hand.
- Corpora have been used in a wide range of disciplines, including lexicography, language teaching and learning, and sociolinguistics.

- Many different types of corpora have been developed for different applications, including written and spoken corpora, general reference corpora, special purpose corpora, monolingual and multilingual corpora, synchronic and diachronic corpora, open and closed corpora, and learner corpora.
- Corpora can be manipulated and analysed with the help of corpus analysis tools, such as word listers and concordancers.
- Corpora offer a number of advantages over other types of resources (e.g. dictionaries, printed texts, subject field experts, intuition) and can therefore be used as useful complements to such resources.
 - Their electronic form means that corpora can be larger and more up-to-date than printed resources, and they can be searched more easily.
 - Corpora consist of authentic texts that can be used to find out what people do and do not say, as well as how often they say it.
 - Corpora can be used to conduct new investigations or to test existing hypotheses.
 - Corpora can be fun and interesting to explore!

Further reading

- Barnbrook, Geoff (1996) *Language and Computers*, Edinburgh: Edinburgh University Press.
- Biber, Douglas, Conrad, Susan and Reppen, Randi (1998) *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge: Cambridge University Press.
- Kennedy, Graeme (1998) *An Introduction to Corpus Linguistics*, London/New York: Longman.
- McEnery, Tony and Wilson, Andrew (1996) *Corpus Linguistics*, Edinburgh: Edinburgh University Press.
- Tognini-Bonelli, Elena (2001) *Corpus Linguistics at Work*, Amsterdam/Philadelphia: John Benjamins.

Exercises

Exercise 1

Write down three sentences containing the word ‘umbrella’. Ask a classmate or friend to do the same. Compare your sentences with those of your classmate. Did you both think of exactly the same sentences? Now compare your sentences to those shown in Figure 1.3, which were taken from the British National Corpus Online (<http://sara.natcorp.ox.ac.uk/>).

The sentences that you thought up are probably not the same as the ones thought up by your classmate, and there is a good chance that none

Angio-Amalgamated company under the EMI umbrella brought to the company under time afterwards making a special metal umbrella.

from the DES (under whose departmental umbrella the group currently resides) ple who take up prescriptions under the umbrella of the NHS are exempt from ch

And fancy borrowing her ghastly old umbrella, when you could have had

Do I have an umbrella?

rella, but it was a very old and ragged umbrella (selflessly, she'd left behind forcing the attacker to back off; an umbrella is excellent for this.

learned to relish shade and arranged my umbrella continually in order to be un ce the 'l' in 'elephant' or the 'm' in 'umbrella'?

was made boss of the party's powerless umbrella body, the Patriotic People's eakfast one of the men returned with an umbrella; everyone else worked with sc Square where I'd been waiting under an umbrella watching a group of Czech sol reatly on just what to assess under the umbrella word 'science' - and how to a ovely girl ran out of the house with an umbrella and held it over my head.

on book-collecting, since the ephemera umbrella seems to cover an extraordina

lank walls and tapped the handle of his umbrella against his chin.

The fact that "I'll return your umbrella" is in the future tense does poked the leaves with the point of his umbrella, a wrinkle of pain on his forehead capabilities which fall under the umbrella term 'learning' are not sufficient. Gothic pew, a rusty suit of armour, an umbrella stand in the form of a bear. with the rain falling on trees, on the umbrella, the only sounds.

iding his female companion with a large umbrella, they moved off toward the horizon shadow cone much like rain falls off an umbrella, as shown in Fig. 1. te spats, a gangster's hat and a rolled umbrella, Gallacher projected the image is to spread the top growth over a wire umbrella to get the hanging stems well branches on the principle of an opening umbrella; Santa Claus masks, red and white locating the briefcase, hat and furled umbrella, and the front door made a clam forward, the hand clutching a furled umbrella by its ferrule, the crook of

The Doctor hooked the handle of his umbrella over his top pocket and pulled

Figure 1.3 Concordance for 'umbrella'

24 *Setting the scene*

of these sentences are the same as the ones that are shown in the concordance. The point of this exercise is to demonstrate that a corpus can be a rich source of data that may provide you with information that you would not have arrived at through introspection alone.

Exercise 2

Now compare the concordance in Figure 1.3 against this dictionary entry for the term 'umbrella' taken from the Oxford Paperback Dictionary (1988 edition):

umbrella n. **1.** a portable protection against rain, consisting of a circular piece of fabric mounted on a foldable frame of spokes attached to a central stick that serves as a handle. **2.** any kind of general protecting force or influence.

What information do you learn from the corpus that was not present in the dictionary, and vice versa? These two types of resource may offer complementary information.