Unit 7 Using available corpora

7.1 Introduction

Many readers will wish to use ‘off the peg’ corpora to carry out their work. In this unit we will introduce some of the major publicly available corpus resources and explore the pros and cons of using ready-made corpora. A corpus is always designed for a particular purpose (cf. Hunston 2002: 14), hence the usefulness of a ready-made corpus must be judged with regard to the purpose to which a user intends to put it. In this respect, the research question once again plays a crucial role.

There are thousands of corpora in the world, but many of them are created for specific research projects and are thus not publicly available. While abundant corpus resources for languages other than English are now available, this unit will focus upon major English corpora, which are classified in terms of their potential use: general vs. specialized corpus, written vs. spoken corpus, synchronic vs. diachronic corpus, learner corpus, and monitor corpus. Note that there is some overlap in the above classification (cf. unit 7.10). It is used in this unit simply to give a better account of the primary use of the relevant corpora.

7.2 General corpora

A general corpus is balanced with regard to the variety of a given language. While the term balance is relative and closely related to a particular research question (cf. unit 2.4), if the corpus in question claims to be general in nature, then it will typically be balanced with regard to genres and domains that typically represent the language under consideration. The corpus may contain written data, spoken data or both.

A well-known general corpus is the British National Corpus (BNC). The BNC comprises 100,106,008 words, organized in 4,124 written texts and transcripts of speech in modern British English. The corpus is designed to represent as wide a range of modern British English as possible. The written section (90%) includes, among many others kinds of text, samples from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, as well as school and university essays. The spoken section (10%) includes 863 transcripts of a large amount of informal conversation, selected from respondents of different ages, from various regions and from all social classes in a demographically balanced way, together with spoken language collected in all kinds of different contexts, ranging from formal business or government meetings to radio shows and phone-ins (see unit 2.4).

In addition to POS information, the BNC is annotated with rich metadata encoded according to the TEI guidelines, using ISO standard 8879 (SGML, see unit 3.3). Because of its generality, as well as the use of internationally agreed standards for its encoding, the corpus is a useful resource for a very wide variety of research purposes, in fields as distinct as lexicography, artificial intelligence, speech recognition and synthesis, literary studies, and, of course, linguistics. There are a number of ways one may access the BNC corpus. It can be accessed online remotely using the SARA client program (see the Appendix for the Internet address) or using the BNCWeb interface (see case study 1). Alternatively, if a local copy of the corpus is available, the corpus can be explored using stand-alone corpus exploration tools such as WordSmith Tools.

In addition to its usefulness in studying modern British English, the BNC, in combination with corpora of other languages adopting a similar sampling frame (see unit 2.5 for a discussion of sampling frame), can provide a reliable basis for contrastive language study. A number of corpora are designed as matches for the BNC. The American National Corpus (ANC), for example, is an American match for the BNC. The ANC will contain a core corpus of at least 100 million words, comparable across genres with the BNC. Beyond this, the corpus will include an additional component of potentially several hundreds of millions of words, chosen to provide both the broadest and largest selection of texts possible. The Korean National Corpus also adopts the BNC model (see unit 5.2), as does the Polish National Corpus (PNC) (see Lewandowska-Tomaszczyk 2003: 106).

7.3 Specialized corpora

A specialized corpus is specialized relative to a general corpus. It can be domain or genre specific and is designed to represent a sublanguage (see unit 2.2). For example, the Guangzhou Petroleum English Corpus contains 411,612 words of written English from the petrochemical domain. The HKUST Computer Science Corpus is a one-million-word corpus of written English sampled from undergraduate textbooks in computer science. Both corpora are domain specific.

It is interesting to note that there has recently much interest in the creation and exploitation of specialized corpora in academic or professional settings. For example, the Corpus of Professional Spoken American English (CPSA) has been constructed from a selection of transcripts of interactions in professional settings, containing two main subcorpora of one million words each. One subcorpus consists mainly of academic discussions such as faculty council meetings and committee meetings related to testing. The second subcorpus contains transcripts of White House press conferences, which are almost exclusively question-and-answer sessions (cf. Barlow 1998). The Michigan Corpus of Academic Spoken English (MICASE) is a corpus of approximately 1.7 million words (nearly 200 hours of recordings) focusing on contemporary university speech within the domain of the University of Michigan (cf. MICASE Manual). The entire corpus can be accessed now online at the corpus website. A much more ambitious project has been initiated by the Professional English Research Consortium (PERC), which aims to create a 100-million-word Corpus of Professional English (CPE), consisting of both spoken and written discourse used by working professionals and professionals-in-training and covering a wide range of domains such as science, engineering, technology, law, medicine, finance and other professions.

As language may vary considerably across domain (see case studies 4 and 6) and genre (see case study 5), specialized corpora such as those introduced above provide valuable resources for investigations in the relevant domains and genres. It is important to note that specialized corpora can also be extracted from general corpora (see unit 7.10).

7.4 Written corpora

The first modern corpus of English was a corpus of written American English, the Brown University Standard Corpus of Present-day American English (i.e. the Brown corpus, see Kucĕra and Francis 1967). The corpus was compiled using 500 chunks of approximately 2,000 words of written texts. These texts were sampled from 15 categories. All were produced in 1961. The components of the Brown corpus are given in Table 7.1.

Table 7.1 Text categories in the Brown corpus

Code	Text category	No. of samples	Proportion
A	Press reportage	44	8.8%
B	Press editorials	27	5.4%
C	Press reviews	17	3.4%
D	Religion	17	3.4%
E	Skills, trades and hobbies	38	7.6%
F	Popular lore	44	8.8%
G	Biographies and essays	77	15.4%
H	Miscellaneous (reports, official documents)	30	6.0%
J	Science (academic prose)	80	16.0%
K	General fiction	29	5.8%
L	Mystery and detective fiction	24	4.8%
M	Science fiction	6	1.2%
N	Western and adventure fiction	29	5.8%
P	Romantic fiction	29	5.8%
R	Humour	9	1.8%
Total		500	100%

There are a number of corpora which follow the Brown model. The Lancaster-Oslo-Bergen Corpus of British English (i.e. LOB, see Johansson, Leech and Goodluck 1978) is a British match for the Brown corpus. The corpus was created using the same sampling techniques with the exception that LOB aims to represent written British English used in 1961. The two corpora provide an ideal basis for the comparison of the two major varieties of English as used in the early 1960s. Both Brown and LOB are POS tagged. Sub-samples from both corpora have also been parsed (see unit 4.4.3 for a discussion of parsing). The Lancaster Parsed Corpus (LPC) is a sub-sample of approximately 133,000 words taken from the LOB corpus that has been parsed. The Susanne corpus is a 128,000 word sub-sample taken from the Brown corpus that has been parsed.

Two Freiburg corpora are available to mirror the Brown/LOB relationship in the early 1990s rather than 1960s. The Freiburg-LOB Corpus of British English (i.e. FLOB, see Hundt, Sand and Siemund 1998) and the Freiburg-Brown Corpus of American English (i.e. Frown, see Hundt, Sand and Skandera 1999) represent written British and American English as used in 1991. In addition to providing a basis for comparing the two major varieties of English in the early 1990s, the two Freiburg corpora also enable users to track language changes in British and American English over the intervening three decades between Brown/LOB and FLOB/Frown. The POS tagged versions of both FLOB and Frown are available. Brown/LOB and FLOB/Frown will be used in case study 2 of Section C.

In addition to British and American English, a couple of corpora have been created for varieties of English using the Brown sampling model. For example, the Australian Corpus of English (i.e. ACE, also known as the Macquarie Corpus) represents written Australian English from 1986 and after; the Wellington Corpus (WWC) represents written New Zealand English, covering the years between 1986 and 1990; and the Kolhapur corpus represents Indian English dating from 1978. Yet not all Brown matches focus on varieties of English. As noted in unit 7.2, a sampling frame may cross languages as well as language varieties. An example of this using the Brown sampling frame is the Lancaster Corpus of Mandarin Chinese (i.e. LCMC, see

McEnery, Xiao and Mo 2003), a Chinese match for the FLOB and Frown corpora. Such a corpus makes it possible to contrast Chinese with two major English varieties.

7.5 Spoken corpora

In addition to the spoken part of general corpora such as the BNC and the genre specific spoken corpora introduced in units 7.2 and 7.3, a number of spoken English corpora are available. They include, for example, the London-Lund Corpus (LLC), the Lancaster/IBM Spoken English Corpus (SEC), the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), the Santa Barbara Corpus of Spoken American English (SBCSAE) and the Wellington Corpus of Spoken New Zealand English (WSC).

The LLC is a corpus of spoken British English dating from the 1960s to the mid-1970s. The corpus consists of 100 texts, each of 5,000 words, totalling half a million running words. A distinction is made between dialogue (e.g. face-to-face conversations, telephone conversations and public discussion) and monologue (both spontaneous and prepared) in the organization of the corpus. The corpus is prosodically annotated.

The SEC corpus consists of approximately 53,000 words of spoken British English, mainly taken from radio broadcasts dating from the mid-1980s and covering a range of speech categories such as commentary, news broadcast, lecture and dialogue. The corpus is available in an orthographically transcribed form. A POS tagged, parsed or prosodically annotated version is also available.

CANCODE is part of the Cambridge International Corpus (CIC). The corpus comprises five million words of transcribed spontaneous speech collected in Britain between 1995 and 2000, covering a wide variety of situations: casual conversation, people working together, people shopping, people finding out information, discussions and many other types of interaction. A unique feature of CANCODE is that the corpus has been coded with information pertaining to the relationship between the speakers: whether they are intimates (living together), casual acquaintances, colleagues at work, or strangers. This coding allows users to look more closely at how different levels of familiarity (formality) affect the way in which people speak to each other.

The Santa Barbara Corpus of Spoken American English (SBCSAE) is part of the USA component of the International Corpus of English (i.e. ICE, see unit 7.6). It is based on hundreds of recordings of spontaneous speech from all over the United States, representing a wide variety of people of different regional origins, ages, occupations, and ethnic and social backgrounds. It reflects the many ways that people use language in their lives: conversation, gossip, arguments, on-the-job talk, card games, city council meetings, sales pitches, classroom lectures, political speeches, bedtime stories, sermons, weddings, and so on. The corpus is particularly useful for research into speech recognition as each speech file is accompanied by a transcript in which phrases are time stamped to allow them to be linked with the audio recording from which the transcription was produced. The SBCSAE will be used in case study 5 in Section C.

The WSC corpus comprises one million words of spoken New Zealand English in the form of 551 extracts collected between 1988 and 1994 (99% of the data from 1990–1994, the exception being eight private interviews). Formal speech/monologue accounts for 12 % of the data, semi-formal speech/elicited monologue 13% while informal speech/dialogue accounts for 75%. The unusually high proportion of private material makes the corpus a valuable resource for research into informal spoken registers.

7.6 Synchronic corpora

The written English corpora introduced in unit 7.4 are useful should one wish to compare varieties of English. The typical examples are the corpora of the Brown family: Brown and Frown for American English, LOB and FLOB for British English, ACE for Australian English, WWC for New Zealand English, and the Kolhapur corpus for Indian English (see unit 7.4 for a description of these corpora). While these corpora are generally good for comparing national varieties of English (so-called 'world Englishes', see unit 10.5), the results from such a comparison must be interpreted with caution where the corpora under examination were built to represent English in different time periods (e.g. Brown vs. FLOB) or the Brown model has been modified. A more reliable basis for comparing world Englishes is the International Corpus of English (ICE), which is specifically designed for the synchronic study of world Englishes. The ICE corpus consists of a collection of twenty corpora of one million words each, each composed of written and spoken English produced after 1989 in countries or regions in which English is a first or major language (e.g. Australia, Canada, East Africa, Hong Kong as well as the UK and the USA). As the primary aim of ICE is to facilitate comparative studies of English used worldwide, each component follows a common corpus design as well as a common scheme for grammatical annotation to ensure comparability among the component corpora (cf. Nelson 1996). The ICE corpora are encoded at various levels, including textual markup, POS tagging and syntactic parsing. Readers can visit the ICE website to check the availability of the components corpora in ICE.

In contrast, there are considerably fewer corpora available for regional dialects than national varieties. This is because comparing dialects is assumed to be less meaningful than comparing varieties of a language or two distinct languages. Comparisons of dialects are even claimed by some to be 'vacuous' (Bauer 2002: 108). The spoken component of the BNC does allow users to compare dialects in Britain, but only between broadly sampled groups such as 'South', 'Midlands' and 'North' (see Aston and Burnard 1998: 31). The Longman Spoken American Corpus, which was built to match the demographically sampled component of the spoken BNC, can be used to compare regional dialects in the USA. A corpus that was built specifically for the study of English dialects is the spoken corpus of the Survey of English dialects (i.e. SED; see Beare and Scott 1999). The Survey of English Dialects was started in 1948 by Harold Orton at the University of Leeds. The initial work comprised a questionnaire-based survey of traditional dialects based on extensive interviews from 318 locations all over rural England. During the survey, a number of recordings were made as well as the detailed interviews. The recordings, which were made during 1948-1973, consist of about 60 hours of dialogue of elderly people talking about life, work and recreation. The recordings were transcribed, with sound files linked to transcripts. The corpus is marked up in TEI-compliant SGML (see unit 3.3) and POS tagged using CLAWS (see unit 4.4.1).

There are presently few synchronic corpora suitable for studies of dialects and varieties for languages other than English. The LIVAC (Linguistic Variation in Chinese Speech Communities) corpus is one of the few that exist. It contains texts from representative Chinese newspapers and the electronic media of six Chinese speech communities: Hong Kong, Taiwan, Beijing, Shanghai, Macau and Singapore.

The collection of materials from these diverse communities is synchronized so that all of the components are comparable. The corpus is under construction by the City University of Hong Kong but some samples are already available.

7.7 Diachronic corpora

A diachronic (or historical) corpus contains texts from the same language gathered from different time periods. Typically that period is far more extensive than that covered by Brown/Frown and LOB/FLOB (see unit 7.4) or a monitor corpus (see unit 7.9). Corpora of this type are used to track changes in language evolution. For practical reasons, diachronic corpora typically contain only written language, though corpus builders have tried to construct corpora of speech from earlier periods, e.g., the Helsinki Dialect Corpus (see Peitsara and Vasko 2002) and the Corpus of English Dialogues 1560-1760 (see Culpeper and Kytö 2000). Perhaps the best-known diachronic corpus of English is the Helsinki Diachronic Corpus of English Texts (i.e. the Helsinki corpus), which consists of approximately 1.5 million words of English in the form of 400 text samples, dating from the 8th to 18th centuries. The corpus covers a wide range of genres and sociolinguistic variables and is divided into three periods (Old, Middle, and Early Modern English) and eleven subperiods (see unit 15.4). Another diachronic English corpus of note is the ARCHER corpus (A Representative Corpus of Historical English Registers). The corpus covers both British and American English dating from 1650 to 1990, divided into 50-year periods. Spoken data from the later periods is also included in the corpus. Yet another diachronic corpus, which is more accessible than ARCHER, is the Lampeter Corpus of Early Modern English Tracts. The Lampeter corpus contains 1.1 million words of pamphlet literature dating from 1640 and 1740. As it includes whole texts rather than smaller samples, the corpus is also useful for the study of textual organization in Early Modern English.

7.8 Learner corpora

A type of corpus that is immediately related to the language classroom is a *learner corpus*. A learner corpus is a collection of the writing or speech of learners acquiring a second language (L2). The data collected is the L2 productions of such learners. It can be used for either cross-sectional or longitudinal analysis. The term *learner corpus* is used here as opposed to a *developmental corpus*, which consists of data produced by children acquiring their first language (L1). While L2 learner data for longitudinal analysis can also be called 'developmental' data, we use the term *developmental corpus* specifically for L1 data as opposed to *learner corpus*. Well-known developmental corpora include, for example, the Child Language Data Exchange System (CHILDES, see MacWhinney 1992) and the Polytechnic of Wales Corpus (POW, see Souter 1993).

Probably the best-known learner corpus is the International Corpus of Learner English (ICLE, see Granger 2003a). At present the corpus contains approximately three million words of essays written by advanced learners of English (i.e. university students of English as a foreign language in their 3rd or 4th year of study) from fourteen different mother tongue backgrounds (French, German, Dutch, Spanish, Swedish, Finnish, Polish, Czech, Bulgarian, Russian, Italian, Hebrew, Japanese and Chinese). The error and POS tagged version of the corpus will be available in near future. In addition to allowing the comparison of the writing of learners from different L1 backgrounds, the corpus can be used in combination with the Louvain Corpus of

Native English Essays (LOCNESS) to compare native and learner English. The ICLE corpus is available for linguistic research but cannot be used for commercial purposes.

The Longman Learners' Corpus contains ten million words of text written by students of English at a range of levels of proficiency from 20 different L1 backgrounds. The elicitation tasks used to gather the texts varied, ranging from in-class essays with or without the use of a dictionary to exam essays or assignments. Each student essay is coded by L1 background and proficiency level, amongst other things. The corpus is not tagged for part-of-speech, but part of the corpus has been error-tagged manually, although this portion is only for internal use at Longman Dictionaries. Such a corpus offers invaluable information about the mistakes students make and what they already know. The Longman Learners' Corpus is a useful resource, for example, for lexicographers and textbook materials writers who wish to produce dictionaries and course books that address students' specific needs. The corpus is publicly available for commercial purposes. This corpus will be used in case study 3 in Section C of this book.

The Cambridge Learner Corpus (CLC) is a large collection of examples of English writing from learners of English all over the world. It contains over 20 million words and is expanding continually. The English in CLC comes from anonymized exam scripts written by students taking Cambridge ESOL English exams worldwide. The corpus currently contains 50,000 scripts from 150 countries (100 different L1 backgrounds). Each script is coded with information about the student's first language, nationality, level of English, and age, etc. Over eight million words (or about 25,000 scripts) have been coded for errors. Currently, the corpus can only be used by authors and writers working for Cambridge University Press and by members of staff at Cambridge ESOL.

In addition, a number of learner English corpora are available which cover only one L1 background. For example, the HKUST Corpus of Learner English is a 10-million-word corpus composed of written essays and exam scripts of Chinese learners in Hong Kong (see Milton 1998). The Chinese Learner English Corpus (CLEC) contains one million words from writing produced by five types of Chinese learners of English ranging from middle school students to senior English majors (Gui and Yang 2001). The SST (Standard Speaking Test) corpus contains one million words (around 300 hours of recording, or 1,200 samples transcribed from 15-minute oral interview test) of error tagged spoken English produced by Japanese learners (Izumi et al 2003). The JEFLL (Japanese EFL Learner) corpus is a one-million-word corpus containing 10,000 sample essays written by Japanese learners of English from Years 7-12 in secondary schools. The JPU (Janus Pannonius University) learner corpus is a corpus of 400,000 words which contains the essays of advanced level Hungarian university students that were collected from 1992 to 1998 (Horvath 1999). The USE (Uppsala Student English) corpus contains one million words of texts written primarily by Swedish university students who are advanced learners of English with a high level of proficiency. The Polish Learner English Corpus is designed as a half-a-million-word corpus of written learner data produced by Polish learners of English from a range of learner styles at different proficiency levels, from beginning learners to post-advanced learners (cf. Lewandowska-Tomaszczyk 2003: 107). Readers can refer to Pravec (2002) and the Internet links given in the Appendix for a more comprehensive survey of available learner corpora.

7.9 Monitor corpora

All of the corpora (with the possible exception of CLC) introduced in the previous sections are constant in size. They are sample corpora. In contrast, a *monitor corpus* is constantly (e.g. annually, monthly or even daily) supplemented with fresh material and keeps increasing in size, though the proportion of text types included in the corpus remains constant (see unit 2.3 for further discussion). Corpora of this type are typically much larger than sample corpora. The Bank of English (BoE) is widely acknowledged to be an example of a monitor corpus. It has increased in size progressively since its inception in the 1980s (Hunston 2002: 15) and is around 524 million words at present . The Global English Monitor Corpus, which was started in late 2001 as an electronic archive of the world's leading newspapers in English, is expected to reach billions of words within a few years. The corpus aims at monitoring language use and semantic change in English as reflected in newspapers so as to allow for research into whether the English language discourses in Britain, the United States, Australia, Pakistan and South Africa are convergent or divergent over time. Another example of corpora of this kind is AVIATOR (Analysis of Verbal Interaction and Automated Text Retrieval), developed at the University of Birmingham, which automatically monitors language change, using a series of filters to identify and categorize new word forms, new word pairs or terms, and change in meaning. However, as a monitor corpus does not have a finite size, some corpus linguists have argued that it is an 'ongoing archive' (Leech 1991: 10) rather than a true corpus.

There was an impromptu debate at a joint conference of the Association for Literary and Linguistic Computing (ALLC) and the Association for Computing in the Humanities (ACH) at Christchurch College, Oxford in 1992 between, on the one hand, Quirk and Leech arguing in favour of the balanced sample corpus model and on the other hand Sinclair and Meijs who spoke in favour of the monitor corpus model (cf. Geoffrey Leech, personal communication). Whilst the monitor corpus team won the debate in 1992, it is now clear that the sample corpus model has won the wider debate, as evidenced by it being the dominant tradition in modern corpus building; the majority of corpora which have been built to date are balanced sample corpora, as exemplified by the pioneer English corpora Brown and LOB, and more recently by the British National Corpus. Nonetheless, the idea of the monitor corpus is still important and deserves a review here.

The monitor corpus approach was first developed in Sinclair (1991a: 24-26). Sinclair argued against static sample corpora like Brown and argued in favour of an ever growing dynamic monitor corpus. The ideas expressed in Sinclair (1991a) mirror the way people were thinking about corpora two decades ago. Unsurprisingly, Sinclair has amended his views 'as new advances come on stream' and no longer holds many of the positions held in his 1991a work (Sinclair, personal communication). However, the arguments expressed therein still have some currency, both in the writing of Sinclair (e.g. Sinclair 2004a: 187-191) and others (e.g. Tognini-Bonelli 2001, as noted in unit 1.7). So it is worth reviewing the arguments presented by Sinclair against sample corpora. The major arguments relate to the overall corpus size (one million words) and the sample size (2,000 words). These concerns have largely been neutralized by an increase in both computer power and the availability of electronic texts (in many languages). As Aston and Burnard (1998: 22) comment, 'The continued growth in the size of corpora has generally implied an increase in sample sizes as well as the number of samples.' The overall size of the BNC, for example, is 100 million words. Accordingly, the sample size in the BNC has also increased to 40,000 – 50,000 words while the number of samples has increased to over 4,000. Samples of this size can no longer be said to be 'too brief', and sub-categories composed of such texts can indeed ‘stand as samples themselves.’ Biber (1988, 1993) shows that even in corpora consisting of 2,000-word samples, frequent linguistic features are quite stable both within samples and across samples in different text categories. For relatively rare features and for vocabulary, though, Sinclair’s warning is still valid.

A monitor corpus undergoes a partial self-recycling after reaching some sort of saturation, i.e. the inflow of the new data is subjected to an automatic monitoring process which will only admit new material to the corpus where that material shows some features which differ significantly from the stable part of the corpus (cf. Váradi 2000: 2). There are a number of difficulties with the monitor corpus model, however. First, as this approach rejects any principled attempt to balance a corpus, depending instead upon sheer size to deal with the issue (see unit 1.7), Leech (1991: 10) refers to a monitor corpus as an ‘ongoing archive’, while Váradi (2000: 2) would label it as ‘opportunistic’. As such, monitor corpora are a less reliable basis than balanced sample corpora for quantitative (as opposed to qualitative) analysis. Second, as this approach argues strongly in favour of whole texts, text availability becomes a difficulty in the already sensitive context of copyright. Third, it is quite confusing to indicate a specific corpus version with its word count. Under such circumstances it is only the corpus builders, not the users, who know what is contained in a specific version. Fourth, as a monitor corpus keeps growing in size, results cannot easily be compared. A monitor corpus thus loses its value as ‘a standard reference’ (McEnery and Wilson 2001: 32). Gronqvist (2004) suggests, rightly in our view, that ‘[a] system where it is possible to restore the corpus as it was at a specific time would be worth a lot’ for a monitor corpus. This suggestion would mean that a dynamic monitor corpus should in effect consist of a series of static corpora over time. This is not current practice. One final concern of the dynamic model is that, even if the huge corpus size and required processing speed should not become a problem as the rapid development of computing technology makes this a non-issue, there is no guarantee that the same criteria will be used to filter in new data in the long term (e.g. in 200 years), meaning that even if a diachronic archive of the sort suggested is established, the comparability of the archived version of the corpus would be in doubt.

Monitor corpora are potentially useful, however. A monitor corpus is primarily designed to track changes from different periods (cf. Hunston 2002: 16). In this sense, a monitor corpus is similar to a diachronic corpus. However, a monitor corpus normally covers a shorter span of time but in a much more fine-grained fashion than the diachronic corpora discussed so far. It is possible, using a monitor corpus, to study relatively rapid language change, such as the development and the life cycle of neologisms. In principle, if a monitor corpus is in existence for a very long period of time – 30 years for example – it may also be possible to study change happening at a much slower rate, e.g. grammatical change. At present, however, no monitor corpus has been in existence long enough to enable the type of research undertaken using diachronic corpora like the Helsinki corpus or sample corpora such as LOB and FLOB to be undertaken fruitfully (e.g. Leech 2002; see also unit 15.5).

7.10 Unit summary and looking ahead

In this unit, we introduced some of the major publicly available English corpora which are useful for a range of research purposes, including general vs. specialized corpus, written vs. spoken corpus, synchronic vs. diachronic corpus, learner corpus and monitor corpus. This discussion, however, only covers a very small proportion of the available corpus resources. Readers are advised to refer to the authors' companion website for a more comprehensive survey of well-known and influential corpora for English and other languages.

The distinctions given in this unit have been forced for the purpose of this introduction. It is not unusual to find that any given corpus will be a blend of many of the features introduced here. Consider the BNC, for example. This includes both spoken and written data. While it is a general corpus, one could extract a number of specialized corpora from it (e.g. business English). While presumably intended for use in synchronic research, the written section contains texts spanning from 1960 to 1993 (divided into two periods: 1960-1974 and 1975-1992; or for the BNC World Edition, three periods: 1960-1974, 1975-1984 and 1985-1993.), so the corpus holds some data of interest to diachronic researchers. The spoken section of the BNC even contains some L2 English! So while the distinctions explored here are useful, do not think that they apply in a simple and rigid fashion to all of the corpora you may encounter.

It is also clear that while the sample and monitor corpus models have different focuses in their design criteria, corpora of both types are useful in that they serve different purposes. Specifically, a monitor corpus is better suited for identifying new words and tracking changes in usage/meaning while a representative, balanced sample corpus provides a more reliable language model. Note that the corpora used in our case studies in Section C of this book are all sample corpora.

A recurring theme throughout our discussion of corpora is that the usefulness of a given corpus depends upon the research question a user intends to investigate using that corpus. As such, while there are many corpora readily available, it is often the case that researchers and students will find that they are not able to address their research questions using ready-made corpora. In such circumstances, one must build one's own corpus. In the unit that follows, we will discuss the principal considerations involved in building such DIY corpora.