7

Corpus Creation

Richard Xiao
Edge Hill University

7.1 Introduction

A corpus can be defined as a collection of machine-readable authentic texts (including transcripts of spoken data) that is sampled to be representative of a particular natural language or language variety (McEnery et al. 2006: 5), though "representativeness" is a fluid concept (see Section 7.3). Corpora play an essential role in natural language processing (NLP) research as well as a wide range of linguistic investigations. They provide a material basis and a test bed for building NLP systems. On the other hand, NLP research has contributed substantially to corpus development (see Dipper 2008 for a discussion of the relationship between corpus linguistics and computational linguistics), especially in corpus annotation, for example, part-of-speech tagging (see Chapter 10), syntactic parsing (see Chapters 8 and 11), semantic tagging (see Chapters 5 and 14), as well as the alignment of parallel corpora (see Chapter 16).

There are thousands of corpora in the world, but most of them are created for specific research projects and are not publicly available. Xiao (2008) provides a comprehensive survey of a wide range of well-known and influential corpora in English and many other languages, while a survey of corpora for less-studied languages can be found in Ostler (2008). Since corpus creation is an activity that takes time and costs money, it is certainly desirable for readers to use such ready-made corpora to carry out their work. Unfortunately, however, this is not always feasible or possible. As a corpus is always designed for a particular purpose, the usefulness of a ready-made corpus must be judged with regard to the purpose to which a user intends to put it. Consequently, while there are many corpora readily available, it is often the case that readers 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. This chapter covers principal considerations involved in creating such DIY ("do-it-yourself") corpora as well as the issues that come up in major corpus creation projects.

This chapter discusses core issues in corpus creation such as corpus size, representativeness, balance and sampling, data capture and copyright, markup and annotation, as well as peripheral issues such as multilingual and multimodal corpora.

7.2 Corpus Size

One must be clear about one's research question (or questions) when planning to build a DIY corpus. This helps you to determine what material you will need to collect. For example, if you wish to compare British English and American English, you will need to collect spoken and/or written data produced by native speakers of the two regional varieties of English; if you are interested in how Chinese speakers acquire French as a second language, you will then need to collect the French data produced by Chinese learners to create a learner corpus; if you are interested in how the English language has evolved over centuries, you will need to collect samples of English produced in different historical periods to build a historical or diachronic corpus. Readers are reminded, though, that many corpora of these kinds are now already available (see Xiao 2008 for a recent survey). Having developed an understanding of the type of data you need to collect, and having made sure that no ready-made corpus of such material exists, one needs to find a source of data. Assuming that the data can be found, one then has to address the question of corpus size.

How large a corpus do you need? There is no easy answer to this question. The size of the corpus needed depends upon the purpose for which it is intended as well as a number of practical considerations. In the early 1960s, when the processing power and storage capacity of computers were quite limited, a one-million-word corpus such as the Brown corpus (i.e., the Brown University Standard Corpus of Present-day American English, see Kučera and Francis 1967) appeared to be as large a corpus as one could reasonably build. With the increase in computer power and the availability of machine-readable texts, however, a corpus of this size is no longer considered large, and in comparison with today's giant corpora like the 100-million-word British National Corpus (BNC, see Aston and Burnard 1998) and the 524-million-word Bank of English (BoE, Collins 2007) it appears somewhat small. An interesting discussion of corpus size and design can be found in Keller and Lapata (2003), who compare similarities and differences in the frequencies for bigrams (i.e., two-word clusters) obtained from the BNC and the Web.

The availability of suitable data, especially in machine-readable form, seriously affects corpus size. In building a *balanced* corpus according to fixed proportions (see Section 7.3), for example, the lack of data for one text type may accordingly restrict the size of the samples of other text types taken. This is especially the case for parallel corpora, as it is common for the availability of translations to be unbalanced across text types for many languages. For example, it will be much easier to find Chinese translations of English news stories than English translations of Chinese literary texts. While it is often possible to transfer paper-based texts into electronic form using OCR (optical character recognition) software, the process costs time and money and is error-prone. Hence, the availability of machine-readable data is often the main limiting factor in corpus creation.

Another factor that potentially limits the size of a DIY corpus is copyright (see Section 7.4 for further discussion). Unless the proposed corpus contains entirely out-of-date or copyright-free data, simply gathering available data and using it in a freely available corpus may expose the corpus creator to legal action. When one seeks copyright clearance, one can face frustration—the construction of the corpus is your priority, not the copyright holder's. They may simply ignore you. Their silence cannot be taken as consent. Copyright clearance in building a large corpus necessitates much effort, trouble, and frustration.

No matter how important legal considerations may seem, one should not lose sight of the paramount importance of the research question. This question controls all of your corpus-building decisions, including the decision regarding corpus size. Even if the conditions discussed above allow for a large corpus, it does not mean that a large corpus is what you want. First, the size of the corpus needed to explore a research question is dependent on the frequency and distribution of the linguistic features under consideration in that corpus (cf. McEnery and Wilson 2001: 80). As Leech (1991: 8–29) observes, size is not all-important. Small corpora may contain sufficient examples of frequent linguistic features. To study features such as the number of present and past tense verbs in English, for example, a sample of 1000 words may prove

sufficient (Biber 1993). Second, small specialized corpora serve a very different yet important purpose from large multi-million-word corpora (Shimazumi and Berber-Sardinha 1996). It is understandable that corpora for lexical studies are much larger than those for grammatical studies, because when studying lexis one is interested in the frequency of the distribution of a word (see Baroni 2009 for a discussion of distributions in text), which can be modeled as contrasting with all others of the same category (cf. Santos 1996:11). In contrast, corpora employed in quantitative studies of grammatical devices can be relatively small (cf. Biber 1988; Givon 1995), because the syntactic freezing point is fairly low (Hakulinen et al. 1980: 104). Third, corpora that need extensive manual annotation (e.g., pragmatic annotation) are necessarily small. Fourth, many corpus tools set a ceiling on the number of concordances that can be extracted, for example, WordSmith version 3.0 can extract a maximum of 16,868 concordances (versions 4.0 and 5.0 do not have this limit). This makes it inconvenient for a frequent linguistic feature to be extracted from a very large corpus. Even if this can be done, few researchers can obtain useful information from hundreds of thousands of concordances (cf. Hunston 2002: 25). The data extracted defies manual analysis by a sole researcher by virtue of the sheer volume of examples discovered. Of course, I do not mean that DIY corpora must necessarily be small. A corpus small enough to produce only a dozen concordances of a linguistic feature under consideration will not be able to provide a reliable basis for quantification, though it may act as a spur to qualitative research.

It is important to note, however, that corpus size is an issue of ongoing debate in corpus creation. Some corpus linguists have argued that size matters (e.g., Krishnamurthy 2000; Sinclair 2004; Granath 2007). Large corpora are certainly of advantage in lexicography and in the study of infrequent linguistic structures (e.g., Keller and Lapata 2003). Also, NLP and language engineering can have different requirements for corpora from those used in linguistic research as discussed above. Corpora used in NLP and language engineering tend to be domain- or genre-specific specialized corpora (e.g., those composed of newspapers or telephone-based transactional dialogues), data for which are often easier to collect in large amounts than for balanced corpora. Furthermore, larger corpora are more reliable in statistical modeling, which is essential in natural language processing and language engineering. In a word, the point I wish to make is that the optimum size of a corpus is determined by the research question the corpus is intended to address as well as practical considerations.

7.3 Balance, Representativeness, and Sampling

One of the commonly accepted defining features of a corpus, which distinguishes a corpus from an archive (i.e., a random collection of texts), is *representativeness*. A corpus is designed to represent a particular language or language variety whereas an archive is not. What does representativeness mean in corpus linguistics? According to Leech (1991: 27), a corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety. Biber (1993: 243) defines representativeness from the viewpoint of how this quality is achieved: "Representativeness refers to the extent to which a sample includes the full range of variability in a population." A corpus is essentially a sample of a language or language variety (i.e., population). Sampling is entailed in the creation of virtually any corpus of a living language. In this respect, the representativeness of most corpora is to a great extent determined by two factors: the range of *genres*, *domains*, and *media* included in a corpus (i.e., *balance*) and how the text chunks for each genre are selected (i.e., *sampling*).

The criteria used to select texts for inclusion in a corpus are principally external to the texts themselves and dependent upon the intended use for the corpus (Aston and Burnard 1998: 23). The distinction between external and internal criteria corresponds to Biber's (1993: 243) situational vs. linguistic perspectives. External criteria are defined situationally irrespective of the distribution of linguistic features whereas internal criteria are defined linguistically, taking into account the distribution of such features. Internal criteria have sometimes been proposed as a measure of corpus representativeness (e.g., Otlogetswe 2004). In my view, it is problematic; indeed it is circular, to use internal criteria such as the

distribution of words or grammatical features as the primary parameters for the selection of corpus data. A corpus is typically designed to study linguistic distributions. If the distribution of linguistic features is predetermined when the corpus is designed, there is no point in analyzing such a corpus to discover naturally occurring linguistic feature distributions. The corpus has been skewed by design. As such, I agree with Sinclair (2005) when he says that the texts or parts of texts to be included in a corpus should be selected according to external criteria so that their linguistic characteristics are, initially at least, independent of the selection process. This view is also shared by many other scholars including Atkins et al. (1992: 5–6) and Biber (1993: 256). Yet, once a corpus is created by using external criteria, the results of corpus analysis can be used as feedback to improve the representativeness of the corpus. In Biber's (1993: 256) words, "the compilation of a representative corpus should proceed in a cyclical fashion."

In addition to text selection criteria, Hunston (2002: 30) suggests that another aspect of representativeness is change over time: "Any corpus that is not regularly updated rapidly becomes unrepresentative." The relevance of permanence in corpus design actually depends on how we view a corpus, that is, whether a corpus should be viewed as a static or dynamic language model. The static view typically applies to a *sample corpus* whereas a dynamic view applies to a *monitor corpus*. A monitor corpus is primarily designed to track changes from different periods (cf. Hunston 2002: 16). It is particularly useful in tracking relatively rapid language change, such as the development and the life cycle of neologisms. Monitor corpora are constantly (e.g., annually, monthly, or even daily) supplemented with fresh material and keep increasing in size. For example, the Bank of English (BoE) has increased in size progressively since its inception in the 1980s (Hunston 2002: 15) and is around 524 million words at present. In contrast, a sample corpus is designed to represent a static snapshot of a particular language variety at a particular time. Static sample corpora, if resampled, may also allow the study of slower paced language change over time. For example, the LOB (Lancaster-Oslo-Bergen Corpus of British English, Johansson et al. 1978) and Brown corpora are supposed to represent written British and American English in the early 1960s; and their recent updates, Freiberg-LOB (FLOB, see Hundt et al. 1998) and Freiberg-Brown (Frown, see Hundt et al. 1999) corpora, represent written British and American English in the early 1990s respectively. Sample corpora such as these make it possible to track language change over the intervening three decades.

In addition to the distinction between sample and monitor corpora, representativeness has different meanings for *general* and *specialized* corpora. Corpora of the first type typically serve as a basis for an overall description of a language or language variety. The BNC corpus, for example, is supposed to represent modern British English as a whole. In contrast, a specialized corpus tends to be specific to a particular domain (e.g., medicine or law) or genre (e.g., newspaper text or academic prose). For a general corpus, it is understandable that it should cover, proportionally, as many text types as possible so that the corpus is maximally representative of the language or language variety it is supposed to represent. Even a specialized corpus, for example, one dealing with telephone calls to an operator service should be balanced by including within it a wide range of types of operator conversations (e.g., line fault, request for an engineer call out, number check, etc.) between a range of operators and customers (cf. McEnery et al. 2001) so that it can be claimed to represent this variety of language.

While both general and specialized corpora should be representative of a language or language variety, they have different criteria for representativeness. The representativeness of a general corpus depends heavily on sampling from a broad range of genres whereas the representativeness of a specialized corpus, at the lexical level at least, can be measured by the degree of *closure* (McEnery and Wilson 2001: 166) or *saturation* (Belica 1996: 61–74) of the corpus. Closure/saturation for a particular linguistic feature (e.g., size of lexicon) of a variety of language (e.g., computer manuals) means that the feature appears to be finite or is subject to very limited variation beyond a certain point. To measure the saturation of a corpus, the corpus is first divided into segments of equal size based on its tokens. The corpus is said to be saturated at the lexical level if each addition of a new segment yields approximately the same number of new lexical items as the previous segment, that is, when the curve of lexical growth is asymptotic, or flattening out. The notion of saturation is claimed to be superior to such concepts as balance for its measurability (Teubert 2000). It should be noted, however, that saturation is only concerned with lexical

features. While it may be possible to adapt saturation to measure features other than lexical growth, there have been few attempts to do this to date (though see McEnery and Wilson 2001: 176–183 for a study of part-of-speech and sentence type closure).

It appears, then, that the representativeness of a corpus, especially a general corpus, depends primarily on how balanced the corpus is; in other words, the range of text categories included in the corpus. As with representativeness, the acceptable balance of a corpus is determined by its intended uses. Hence, a general corpus that contains both written and spoken data (e.g., the BNC) is balanced; so are written corpora such as Brown and LOB, and spoken corpora such as the Cambridge Nottingham Corpus of Discourse in English (CANCODE). A balanced corpus usually covers a wide range of text categories that are supposed to be representative of the language or language variety under consideration. These text categories are typically sampled proportionally for inclusion in a corpus so that "it offers a manageably small scale model of the linguistic material which the corpus builders wish to study" (Atkins et al. 1992: 6).

Balance appears to be a more important issue for a static sample corpus than for a dynamic monitor corpus. As corpora of the latter type are updated frequently, it is usually "impossible to maintain a corpus that also includes text of many different types, as some of them are just too expensive or time consuming to collect on a regular basis" (Hunston 2002: 30–31). The builders of monitor corpora appear to feel that balance has become less of a priority—sheer size seems to have become the basis of the corpus's authority, under the implicit and arguably unwarranted assumption that a corpus will in effect balance itself when it reaches a substantial size.

While balance and representativeness are important considerations in corpus design, they depend on the research question and the ease with which data can be captured and thus must be interpreted in relative terms. In other words, a corpus should only be as representative as possible of the language variety under consideration. For example, if one wants a corpus that is representative of general English, a corpus representative of newspapers will not do; if one wants a corpus representative of newspapers, a corpus representative of *The Times* will not do. Corpus balance and representativeness are fluid concepts that link directly to research questions. The research question one has in mind when building (or thinking of using) a corpus defines the required balance and representativeness. Any claim of corpus balance is largely an act of faith rather than a statement of fact as, at present, there is no reliable scientific measure of corpus balance. Rather the notion relies heavily on intuition and best estimates. Another argument supporting a loose interpretation of balance and representativeness is that these notions *per se* are open to question (cf. Hunston 2002: 28–30). To achieve corpus representativeness along the lines of the Brown corpus model one must know how often each genre is used by the language community in the *sampling period*. Yet it is unrealistic to determine the correlation of language production and reception in various genres (cf. Hausser 1999: 291; Hunston 2002: 29). The only solution to this problem is to treat corpus-based findings with caution. It is advisable to base your claims on your corpus and avoid unreasonable generalizations. Likewise, conclusions drawn from a particular corpus must be treated as deductions rather than facts (cf. also Hunston 2002: 23). With that said, however, I entirely agree with Atkins et al. (1992: 6), who comment that:

> It would be short-sighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as "unreliable" or "irrelevant" because the corpus used cannot be proved to be 'balanced.'

Given that language is infinite whereas a corpus is finite in size, *sampling* is unavoidable in corpus creation. Unsurprisingly, corpus representativeness and balance are closely associated with sampling. Given that we cannot exhaustively describe natural language, we need to sample it in order to achieve a level of balance and representativeness that matches our research question. Having decided that sampling is inevitable, there are important decisions that must be made about how to sample so that the resulting corpus is as balanced and representative as practically possible.

As noted earlier, with few exceptions, a corpus is typically a sample of a much larger *population*. A sample is assumed to be representative if what we find for the sample also holds for the general

population (cf. Manning and Schütze 1999: 119). In the statistical sense, samples are scaled down versions of a larger population (cf. Váradi 2000). The aim of sampling theory "is to secure a sample which, subject to limitations of size, will reproduce the characteristics of the population, especially those of immediate interest, as closely as possible" (Yates 1965: 9).

In order to obtain a representative sample from a population, the first concern to be addressed is to define *the sampling unit* and the boundaries of the population. For written text, for example, a sampling unit may be a book, a periodical, or a newspaper. The population is the assembly of all sampling units while the list of sampling units is referred to as a *sampling frame*. The population from which samples for the pioneering Brown corpus were drawn, for instance, was all written English text published in the United States in 1961 while its sampling frame was a list of the collection of books and periodicals in the Brown University Library and the Providence Athenaeum. For the LOB corpus, the target population was all written English text published in the United Kingdom in 1961 while its sampling frame included the *British National Bibliography Cumulated Subject Index* 1960–1964 for books and *Willing's Press Guide* 1961 for periodicals.

In corpus design, a population can be defined in terms of language production, language reception, or language as a product. The first two designs are basically demographically oriented as they use the demographic distribution (e.g., age, sex, social class) of the individuals who produce/receive language data to define the population while the last design is organized around text category/genre of language data. As noted earlier, the Brown and LOB corpora were created using the criterion of language as a product while the BNC defines the population primarily on the basis of both language production and reception. However, it can be notoriously difficult to define a population or construct a sampling frame, particularly for spoken language, for which there are no ready-made sampling frames in the form of catalogues or bibliographies.

Once the target population and the sampling frame are defined, different sampling techniques can be applied to choose a sample that is as representative as possible of the population. A basic sampling method is *simple random sampling*. With this method, all sampling units within the sampling frame are numbered and the sample is chosen by use of a table of random numbers. As the chance of an item being chosen correlates positively with its frequency in the population, simple random sampling may generate a sample that does not include relatively rare items in the population, even though they can be of interest to researchers. One solution to this problem is *stratified random sampling*, which first divides the whole population into relatively homogeneous groups (so-called strata) and then samples each stratum at random (see Evert 2006 for a discussion of random sampling in corpus creation). In the Brown and LOB corpora, for example, the target population for each corpus was first grouped into 15 text categories such as news reportage, academic prose, and different types of fiction; samples were then drawn from each text category. Demographic sampling, which first categorizes sampling units in the population on the basis of speaker/writer age, sex and social class, is also a type of stratified sampling. Biber (1993) observes that a stratified sample is never less representative than a simple random sample.

A further decision to be made in sampling relates to *sample size*. For example, with written language, should we sample full texts (i.e., whole documents) or text chunks? If text chunks are to be sampled, should we sample text initial, middle, or end chunks? Full text samples are certainly useful in text linguistics, yet they may potentially constitute a challenge in dealing with vexatious copyright issues. Also, given its finite overall size, the coverage of a corpus including full texts may not be as balanced as a corpus including text segments of constant size. As a result, "the peculiarity of an individual style or topic may occasionally show through into the generalities" (Sinclair 1991: 19). Aston and Burnard (1998: 22) argue that the notion of "completeness" may sometimes be "inappropriate or problematic." As such, unless a corpus is created to study such features as textual organization, or copyright holders have granted you permission to use full texts, it is advisable to sample text segments. According to Biber (1993: 252), frequent linguistic features are quite stable in their distributions and hence short text chunks (e.g., 2000 running words) are usually sufficient for the study of such features while rare features are more varied in their distribution

and thus require larger samples (Baroni 2009). In selecting samples to be included in a corpus, however, attention must also be paid to ensure that text initial, middle, and end samples are balanced.

Another sampling issue, which particularly relates to stratified sampling, is the proportion and the number of samples for each text category. The numbers of samples across text categories should be proportional to their frequencies and/or weights in the target population in order for the resulting corpus to be considered as representative. Nevertheless, it has been observed that, as with defining a target population, such proportions can be difficult to determine objectively (cf. Hunston 2002: 28–30). Furthermore, the criteria used to classify texts into different categories or genres are often dependent on intuitions. As such, the representativeness of a corpus, as noted, should be viewed as a statement of belief rather than fact. In the Brown corpus, for example, a panel of experts determined the ratios between the 15 text categories. As for the number of samples required for each category, Biber (1993) demonstrates that ten 2000-word samples are typically sufficient.

The above discussion suggests that in creating a balanced, representative corpus, stratified random sampling is to be preferred over simple random sampling while different sampling methods should be used to select different types of data. For written texts, a text typology established on the basis of external criteria is highly relevant while for spoken data demographic sampling is appropriate. However, context-governed sampling must complement samples obtained from demographic sampling so that some contextually governed linguistic variations can be included in the resulting corpus.

7.4 Data Capture and Copyright

For pragmatic reasons noted in Section 7.2, electronic data is preferred over paper-based material in building DIY corpora. The World Wide Web (WWW) is an important source of machine-readable data for many languages. For example, digital text archives mounted on the Web such as Oxford Text Archive (http://ota.ahds.ac.uk/) and Project Gutenberg (http://www.gutenberg.org/catalog/) as well as the digital collections of some university libraries (e.g., http://lib.virginia.edu/digital/collections/text/, http://onlinebooks.library.upenn.edu/) provide large amounts of publicly accessible electronic texts.

The web pages on the Internet normally use Hypertext Markup Language (i.e., HTML) to enable browsers like Internet Explorer or Netscape to display them properly. While the tags (included in angled brackets) are typically hidden when a text is displayed in a browser, they do exist in the source file of a web page. Hence, an important step in building DIY corpora using web pages is tidying up the downloaded data by converting web pages to plain text, or to some desired format, for example, XML (see Section 7.5). In this section, I will introduce some useful tools to help readers to download data from the Internet and clean up the downloaded data by removing or converting HTML tags. These tools are either freeware or commercial products available at affordable prices.

While it is possible to download data page by page, which is rather time consuming, there are a number of tools that facilitate downloading all of the web pages on a selected Web site in one go (e.g., Grab-a-Site or HTTrack), or more usefully, downloading related web pages (e.g., containing certain key words) at one go. The WordSmith Tools (versions 4.0 and 5.0), for example, incorporates the WebGetter function that helps users to build DIY corpora. WebGetter downloads related Web pages with the help of a search engine (Scott 2003: 87). Users can specify the minimum file length or word number (small files may contain only links to a couple of pictures and nothing much else), required language and, optionally, required words. Web pages that satisfy the requirements are downloaded simultaneously (cf. Scott 2003: 88–89). The WebGetter function, however, does not remove the HTML markup or convert it to XML. The downloaded data needs to be tidied up using other tools before they can be loaded into a concordancer or further annotated.

Another tool worth mentioning is the freeware Multilingual Corpus Toolkit (MLCT, see Piao et al. 2002). The MLCT runs in Java Runtime Environment (JRE) version 1.4 or above, which is freely available on the Internet. In addition to many other functions needed for multilingual language processing (e.g.,

markup, part-of-speech tagging, and concordancing), the system can be used to extract texts from the Internet. Once a web page is downloaded, it is cleaned up. One weakness of the program is that it can only download one web page at a time. Yet this weakness is compensated for by another utility that converts all of the web pages in a file folder (e.g., the web pages downloaded using the Webgetter function of WordSmith version 4.0) to a desired text format in one go. Another attraction of the MLCT is that it can mark up textual structure (e.g., paragraphs and sentences) automatically.

Finally, the BootCaT Toolkit provides a suite of utilities that allow the user to bootstrap specialized corpora and terms from the Web on the basis of a small set of terms as input (Baroni and Bernardini 2004). Readers interested in the Web as corpus can refer to Kilgarriff and Grefenstette (2003), Baroni and Bernardini (2006), and Hundt et al. (2007), and refer to Keller and Lapata (2003) for a comparison of the frequencies obtained from the Web and a balanced corpus such as the BNC.

A major issue in data collection is copyright. While it is possible to use copyright-free material in corpus creation, such data are usually old and a corpus consisting entirely of such data is not useful if one wishes to study contemporary English, for example. Such corpora are even less useful in NLP research, which tends to focus on current language use. Simply using copyrighted material in a corpus without the permission of the copyright holders may cause unnecessary trouble. In terms of purposes, corpora are typically of two types: for commercial purposes or for non-profit-making academic research. It is clearly unethical and illegal to use the data of other copyright holders to make money solely for oneself. Creators of commercial corpora usually reach an agreement with copyright holders as to how the profit will be shared. Publishers as copyright holders are also usually willing to contribute their data to a corpus-building project if they can benefit from the resulting corpus (e.g., the British National Corpus, the Longman Corpus Network, and the Cambridge International Corpus).

In creating DIY corpora for use in non-profit-making research, you might think that you need not worry about copyright if you are not selling your corpus to make a profit. Sadly, this is not the case. Copyright holders may still take you to the court. They may, for example, suffer a loss of profit because your use of their material diminishes their ability to sell it: why buy a book when you can read it for free in a corpus (cf. also Amsler 2002)? Copyright issues in corpus creation are complex and unavoidable. While corpus linguists have brought them up periodically for discussion, there is as yet no satisfactory solution to the issue of copyright in corpus creation.

The situation is complicated further by variation in copyright law internationally. According to the copyright law of EU countries, the term of copyright for published works in which the author owns the copyright is the author's lifetime plus 70 years. Under U.S. law, the term of copyright is the author's lifetime plus 50 years; but for works published before 1978, the copyright term is 75 years if the author renewed the copyright after 28 years.

One is able to make some use of copyrighted text without getting clearance, however. Under the convention of "fair dealing" in copyright law, permission need not be sought for short extracts not exceeding 400 words from prose (or a total of 800 words in a series of extracts, none exceeding 300 words); a citation from a poem should not exceed 40 lines or one quarter of the poem. So one can resort to using small samples to build perfectly legal DIY corpora on the grounds of fair usage. But the sizes of such samples are so small as to jeopardize any claim of balance or representativeness.

I maintain that the fair use doctrine as it applies to citations in published works should operate differently when it applies to corpus creation so as to allow corpus creators to build corpora quickly and legally. The limited reproduction of copyrighted works, for instance, in chunks of 3000 words or one-third of the whole text (whichever is shorter) should be protected under fair use for non-profit-making research and educational purposes. A position statement along these lines has been proposed by the corpus using community articulating the point of view that distributing minimal citations of copyrighted texts and allowing the public indirect access to privately held collections of copyrighted texts for statistical purposes are a necessary part of corpus linguistics research and should be inherently protected as fair use, particularly in non-profit-making research contexts (see Cooper 2003). This aim is not a legal reality yet, however. It will undoubtedly take time for a balance between copyright and fair use for corpus building to develop.

So, what does one do about copyright? My general advice is: if you are in doubt, seek permission. It is usually easier to obtain permission for samples than for full texts, and easier for smaller samples than for larger ones. If you show that you are acting in good faith, and only small samples will be used in non-profit-making research, copyright holders are typically pleased to grant you permission. If some do refuse, you remember it is their right to do so and move on to try other copyright holders until you have enough data.

It appears easier to seek copyright clearance for Web pages on the Internet than for material collected from printed publications. It has been claimed (Spoor 1996: 67) that a vast majority of the documents published on the Internet are not protected by copyright, and that many authors of texts are happy to be able to reach as many people as possible. However, readers should bear in mind that this may not be the case. For example, Cornish (1999: 141) argues that probably all material available on the Web is copyrighted, and that digital publications should be treated the same way as printed works.

Copyright law is generally formulated to prevent someone from making money from selling intellectual property belonging to other people. Unless you are making money using the intellectual property of other people, or you are somehow causing a loss of income to them, it is quite unlikely that copyright problems will arise when building a corpus. Yet copyright law is in its infancy. Different countries have different rules, and it has been argued that with reference to corpora and copyright there is very little which is obviously legal or illegal (cf. Kilgarriff 2002). My final word of advice is: proceed with caution.

7.5 Corpus Markup and Annotation

Data collected using a sampling frame as discussed in Section 7.3 forms a raw corpus. Yet such data typically needs to be processed before use. For example, spoken data needs to be transcribed from audio/video recordings; written texts may need to be rendered machine readable, if they are not already, by keyboarding or OCR scanning. Beyond this basic processing, however, lies another form of preparatory work—corpus markup. In addition, in order to extract linguistic information from a corpus, such information must first of all be encoded in the corpus, a process that is technically known as "corpus annotation."

Corpus markup is a system of standard codes inserted into a document stored in electronic form to provide information about the text itself (i.e., text metadata) and govern formatting, printing or other processing (i.e., structural organization). While metadata markup can be embedded in the same document or stored in a separate but linked document (see below for further discussion of embedding vs. stand-alone annotation), structural markup has to be embedded in the text. Both types of markups are important in corpus creation for at least three reasons. First, the corpus data basically consists of samples of used language. This means that these examples of linguistic usage are taken out of the context in which they originally occurred and their contextual information is lost. Burnard (2002) compares such out-of-context examples to a laboratory specimen and argues that contextual information (i.e., metadata or "data about data") is needed to restore the context and to enable us to relate the specimen to its original habitat. In corpus creation, therefore, it is important to recover as much contextual information as practically possible to alleviate or compensate for such a loss. Second, while it is possible to group texts and/or transcripts of similar quality together and name these files consistently (e.g., as happens with the LOB and Brown corpora), filenames can provide only a tiny amount of extra-textual information (e.g., text types for written data and sociolinguistic variables of speakers for spoken data) and no textual information (e.g., paragraph/sentence boundaries and speech turns) at all. Yet such data are of great interest to linguists as well as NLP researchers and thus should be encoded, separately from the corpus data *per se*, in a corpus. Markup adds value to a corpus and allows for a broader range of research questions to be addressed as a result. Finally, preprocessing written texts, and particularly transcribing spoken data, also involves markup. For example, in written data, when graphics/tables are removed from the original texts, placeholders must be inserted to indicate the locations and types of omissions; quotations in foreign languages should also be marked up. In spoken data, pausing and paralinguistic features such as laughter

need to be marked up. Corpus markup is also needed to insert editorial comments, which are sometimes necessary in preprocessing written texts and transcribing spoken data. What is done in corpus markup has a clear parallel in existing linguistic transcription practices. Markup is essential in corpus creation.

Having established that markup is important in corpus creation, we can now move on to discuss markup schemes. It goes without saying that extra-textual and textual information should be kept separate from the corpus data (texts or transcripts) proper. Yet there are different schemes one may use to achieve this goal. One of the earliest markup schemes was COCOA. COCOA references consist of a set of attribute names and values enclosed in angled brackets, as in <A WILLIAM SHAKESPEARE>, where A (author) is the attribute name and WILLIAM SHAKESPEARE is the attribute value. COCOA references, however, only encode a limited set of features such as authors, titles, and dates (cf. McEnery and Wilson 2001: 35). Recently, a number of more ambitious metadata markup schemes have been proposed, including for example, the Dublin Core Metadata Initiative (DCMI, see Dekkers and Weibel 2003), the Open Language Archives Community (OLAC, see Bird and Simons 2000), the ISLE Metadata Initiative (IMDI, see Wittenburg et al. 2002), the Text Encoding Initiative (TEI, see Sperberg-McQueen and Burnard 2002), and the Corpus Encoding Standard (CES, see Ide and Priest-Dorman 2000). DCMI provides 15 elements used primarily to describe authored Web resources. OLAC is an extension of DCMI, which introduces refinements to narrow down the semantic scope of DCMI elements and adds an extra element to describe the language(s) covered by the resource. IMDI applies to multimedia corpora (see Section 7.7) and lexical resources as well. From even this brief review it should be clear that there is currently no widely agreed standard way of representing metadata, though all of the current schemes do share many features and similarities. Possibly the most influential schemes in corpus building are TEI and CES, hence I will discuss both of these in some detail here.

The Text Encoding Initiative (TEI) was sponsored by three major academic associations concerned with humanities computing: the Association for Computational Linguistics (ACL), the Association for Literary and Linguistic Computing (ALLC), and the Association for Computers and the Humanities (ACH). The aim of the TEI guidelines is to facilitate data exchange by standardizing the markup or encoding of information stored in electronic form. In TEI, each individual text (referred to as "document") consists of two parts: header (typically providing text metadata) and body (i.e., the text itself), which are in turn composed of different "elements." In a TEI header (tagged as <teiHeader>), for example, there are four principal elements (see Burnard 2002):

- A file description (tagged as <fileDesc>) containing a full bibliographic description of an electronic file.
- An encoding description (tagged as <encodingDesc>), which describes the relationship between an electronic text and the source or sources from which it was derived.
- A text profile (tagged as <profileDesc>), containing a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting.
- A revision history (tagged as <revisionDesc>), which records the changes that have been made to a file.

Each element may contain embedded sub-elements at different levels. Of these, however, only <fileDesc> is required to be TEI-compliant; all of the others are optional. Hence, a TEI header can be very complex, or it can be very simple, depending upon the document and the degree of bibliographic control sought. The body part of a TEI document is also conceived as being composed of elements. In this case, an element can be any unit of text, for example, chapter, paragraph, sentence, or word. Formal markup in the body (i.e., structural markup) is by far rarer than in the header (for metadata markup). It is primarily used to encode textual structures such as paragraphs and sentences. Note that the TEI scheme applies to both the markup of metadata and textual structure as well as the annotation of interpretative linguistic analysis.

The TEI scheme can be expressed using a number of different formal languages. The first editions used the Standard Generalized Markup Language (SGML); the more recent editions (i.e., TEI P4, 2002 and

TEI P5, 2007) can be expressed in the Extensible Markup Language (XML). SGML and XML are very similar, both defining a representation scheme for texts in electronic form, which is device and system independent. SGML is a very powerful markup language, but associated with this power is complexity. XML is a simplified subset of SGML intended to make SGML easy enough for use on the Web. Hence, while all XML documents are valid SGML documents, the reverse is not true. Nevertheless, there are some important surface differences between the two markup languages. End tags can optionally be left out in SGML but they cannot in XML. An attribute name (i.e., generic identifier) in SGML may or may not be case sensitive, but it is always case sensitive in XML. Unless it contains spaces or digits, an attribute value in SGML may be given without double (or single) quotes whereas quotes are mandatory in XML.

As the TEI guidelines are expressly designed to be applicable across a broad range of applications and disciplines, treating not only textual phenomena, they are designed for maximum generality and flexibility (cf. Ide 1998). As such, about 500 elements are predefined in the TEI guidelines. While these elements make TEI very powerful and suitable for the general purpose encoding of electronic texts, they also add complexity to the scheme. In contrast, the Corpus Encoding Standard (CES) is designed specifically for the encoding of language corpora. CES is described as "simplified" TEI in that it includes only the subset of the TEI tagset relevant to corpus-based work. While it simplifies the TEI specifications, CES also extends the TEI guidelines by adding new elements not covered in TEI, specifying the precise values for some attributes, marking required/recommended/optional elements, and explicating detailed semantics for elements relevant to language engineering (e.g., sentence, word, etc.) (cf. Ide 1998).

CES covers three principal types of markups: (1) document-wide markup, which uses more or less the same tags as for TEI to provide a bibliographic description of the document, encoding description, etc.; (2) gross structural markup, which encodes structural units of text (such as volume, chapter, etc.) down to the level of paragraph (but also including footnotes, titles, headings, tables, figures, etc.) and specifies normalization to recommended character sets and entities; (3) markup for sub-paragraph structures, including sentences, quotations, word abbreviations, names, dates, terms and cited words, etc. (see Ide 1998).

CES specifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation as well as general architecture. Three levels of text standardization are specified in CES: (1) the metalanguage level, (2) the syntactic level, and (3) the semantic level. Standardization at the metalanguage level regulates the form of the syntactic rules and the basic mechanisms of markup schemes. Users can use a TEI-compliant Document Type Definition (DTD) to define tag names as well as "document models" that specify the relations among tags. As texts may still have different document structures and markups even with the same metalanguage specifications, standardization at the syntactic level specifies precise tag names and syntactic rules for using the tags. It also provides constraints on content. However, the data sender and the data receiver can interpret even the same tag names differently. For example, a <title> element may be intended by the data sender to indicate the name of a book while the data receiver is under no obligation to interpret it as such, because the element can also show a person's rank, honor, and occupation, etc. This is why standardization at the semantic level is useful. In CES, the <h.title> element only refers to the name of a document. CES seeks to standardize at the semantic level for those elements most relevant to language engineering applications, in particular, linguistic elements. The three levels of standardization are designed to achieve the goal of universal document interchange. Like the TEI scheme, CES not only applies to corpus markup, it also covers encoding conventions for the linguistic annotation of text and speech, currently including morpho-syntactic tagging (i.e., part-of-speech tagging, see Chapter 10) and parallel text alignment in parallel corpora (see Chapter 16).

CES was developed and recommended by the Expert Advisory Groups on Language Engineering Standards (EAGLES) as a TEI-compliant application of SGML that could serve as a widely accepted set of encoding standards for corpus-based work. CES is available in both SGML and XML versions. The XML version, referred to as XCES, has also developed support for additional types of annotation and resources, including discourse/dialogue, lexicons, and speech (Ide et al. 2000). On the other hand, while metalanguages such as SGML and XML usually follow the system of attribute names laid out in implementation standards such as TEI and CES, this may not be necessarily the case.

Closely related to corpus markup is annotation, but the two are different. As annotation is so important in corpus creation and NLP research that specific types of annotation merit in-depth discussions in separate chapters (e.g., Chapters 8, 10, and 14), here I will only discuss annotation briefly. Corpus annotation can be defined as the process of "adding such interpretative, linguistic information to an electronic corpus of spoken and/or written language data" (Leech 1997: 2). While annotation defined in a broad sense may refer to the encoding of both textual/contextual information and interpretative linguistic analysis, as shown by the conflation of the two often found in the literature, the term is used in a narrow sense here, referring solely to the encoding of linguistic analyses such as part-of-speech tagging and syntactic parsing in a corpus text.

Corpus annotation, as used in a narrow sense, is fundamentally distinct from markup, though the distinction is not accepted by all and the two terms are sometimes used interchangeably in the literature. Corpus markup provides relatively objectively verifiable information regarding the components of a corpus and the textual structure of each text. In contrast, corpus annotation is concerned with interpretative linguistic information. "By calling annotation 'interpretative,' we signal that annotation is, at least in some degree, the product of the human mind's understanding of the text" (Leech 1997: 2). For example, the part of speech of a word may be ambiguous and hence is more readily defined as corpus annotation than corpus markup. On the other hand, the sex of a speaker or writer is normally objectively verifiable and as such is a matter of markup, not annotation.

Corpus annotation can be undertaken at different levels and may take various forms. For example, at the phonological level, corpora can be annotated for syllable boundaries (phonetic/phonemic annotation) or prosodic features (prosodic annotation); at the morphological level corpora can be annotated in terms of prefixes, suffixes and stems (morphological annotation); at the lexical level, corpora can be annotated for parts-of-speech (POS tagging), lemmas (lemmatization), and semantic fields (semantic annotation); at the syntactic level, corpora can be annotated with syntactic analysis (parsing, treebanking, or bracketing); at the discoursal level, corpora can be annotated to show anaphoric relations (coreference annotation), pragmatic information like speech acts (pragmatic annotation) or stylistic features such as speech and thought presentation (stylistic annotation). Of these the most widespread type of annotation is part-of-speech tagging (see Chapter 10), which has been successfully applied to many languages; syntactic parsing is also developing rapidly (see Chapters 8 and 11) while some types of annotation (e.g., discoursal and pragmatic annotations) are presently relatively undeveloped.

I have so far assumed that the process of annotation leads to information being mixed in the original corpus text or so-called base document when it is applied to a corpus (i.e., the annotation becomes so-called embedded annotation). However, the Corpus Encoding Standard recommends the use of "stand-alone annotation," whereby the annotation information is retained in separate SGML/XML documents (with different Document Type Definitions) and linked to the original and other annotation documents in hypertext format. In contrast to embedded annotation, stand-alone annotation has a number of advantages (Ide 1998):

- It provides control over the distribution of base documents for legal purposes.
- It enables annotation to be performed on base documents that cannot easily be altered (e.g., they are read-only).
- It avoids the creation of potentially unwieldy documents.
- It allows multiple overlapping hierarchies.
- It allows for alternative annotation schemes to be applied to the same data (e.g., different POS tagsets).
- It enables new annotation levels to be added without causing problems for existing levels of annotation or search tools.
- It allows annotation at one level to be changed without affecting other levels.

Stand-alone annotation is in principle ideal and is certainly technically feasible (see Thompson and McKelvie 1997). It may also represent the future standard for certain types of annotation. In addition,

the stand-alone architecture can facilitate multilevel or multilayer annotations as well (see Dipper 2005). Presently, however, there are two problems associated with stand-alone annotation. The first issue is related to the complexity of corpus annotation. As noted earlier, annotation may have multiple forms in a corpus. While some of these readily allow for the separation of annotation codes from base documents (e.g., lemmatization, part-of-speech tagging, and semantic annotation), others may involve much more complexity in establishing links between codes and annotated items (e.g., coreference and stylistic annotations). Even if such links can be established, they are usually prone to error. The second issue is purely practical. As far as I am aware, the currently available corpus exploration tools, including the latest versions of WordSmith (versions 4.0 and 5.0) and Xaira (Burnard and Todd 2003), have all been designed for use with embedded annotation. Stand-alone annotation, while appealing, is only useful when appropriate search tools are available for use on stand-alone annotated corpora.

7.6 Multilingual Corpora

I have so far assumed in this chapter that a corpus only involves one language. Corpora of this kind are monolingual. But there are also corpora that cover more than one language, which are referred to as multilingual corpora. In this section, I will shift my focus to the multilingual dimension of corpus creation.

With ever increasing international exchange and accelerated globalization, translation and contrastive studies are more popular than ever. As part of this new wave of research on translation and contrastive studies, multilingual corpora such as parallel and comparable corpora are playing an increasingly prominent role. As Aijmer and Altenberg (1996: 12) observe, parallel and comparable corpora "offer specific uses and possibilities" for contrastive and translation studies:

- They give new insights into the languages compared—insights that are not likely to be gained from the study of monolingual corpora.
- They can be used for a range of comparative purposes and increase our knowledge of language-specific, typological and cultural differences, as well as of universal features.
- They illuminate differences between source texts and translations, and between native and nonnative texts.
- They can be used for a number of practical applications, for example, in lexicography, language teaching, and translation.

In addition to these benefits of multilingual resources in linguistic research, we can also add to the list the fact that aligned parallel corpora are indispensable to the development of NLP applications such as computer-aided translation and machine translation (see Chapters 17 and 18) and multilingual information retrieval and extraction (see Chapters 19 and 21).

A multilingual corpus involves texts of more than one language. As corpora that cover two languages are conventionally known as "bilingual," multilingual corpora, in a narrow sense, must involve more than two languages, though "multilingual" and "bilingual" are often used interchangeably in the literature, and also in this chapter. A multilingual corpus can be a parallel corpus, or a comparable corpus. Given that corpora involving more than one language are a relatively new phenomenon, with most related research hailing from the early 1990s, it is unsurprising to discover that there is some confusion surrounding the terminology used in relation to these corpora.

It can be said that terminological confusion in multilingual corpora centers on two terms: "parallel" and "comparable." For some scholars (e.g., Aijmer and Altenberg 1996; Granger 1996: 38), corpora composed of source texts in one language and their translations in another language (or other languages) are "translation corpora" while those comprising different components sampled from different native languages using comparable sampling techniques are called "parallel corpora." For others (e.g., Baker 1993: 248, 1995, 1999; Barlow 1995, 2000: 110; Hunston 2002: 15; McEnery and Wilson 1996: 57; McEnery

et al. 2006), corpora of the first type are labeled "parallel" while those of the latter type are comparable corpora. As argued in McEnery and Xiao (2007a: 19–20), while different criteria can be used to define different types of corpora, they must be used consistently and logically. For example, we can say that a corpus is monolingual, bilingual, or multilingual if we take the number of languages involved as the criterion for definition. We can also say that a corpus is a translation or a non-translation corpus if the criterion of corpus content is used. But if we choose to define corpus types by the criterion of corpus form, we must use the terminology consistently. Then we can say a corpus is parallel if the corpus contains source texts and translations in parallel, or it is a comparable corpus if its components or subcorpora are comparable by applying the same sampling frame. It is illogical, however, to refer to corpora of the first type as translation corpora by the criterion of content while referring to corpora of the latter type as comparable corpora by the criterion of form.

Additionally, a parallel corpus, in my terms, can be either unidirectional (e.g., from English into Chinese or from Chinese into English alone), or bidirectional (e.g., containing both English source texts with their Chinese translations as well as Chinese source texts with their English translations), or multidirectional (e.g., the same piece of text with its Chinese, English, French, Russian, and Arabic versions). In this sense, texts that are produced simultaneously in different languages (e.g., UN regulations) also belong to the category of parallel corpora. A parallel corpus must be aligned at a certain level (for instances, at document, paragraph, sentence, or word level) in order to be useful. The automatic alignment of parallel corpora is not a trivial task for some language pairs, though alignment is generally very reliable for many closely related European language pairs (cf. McEnery et al. 2006: 50–51; see Chapter 16 for further discussion).

Another complication in terminology involves a corpus that is composed of different variants of the same language. This is particularly relevant to translation studies because it is a very common practice in this research area to compare a corpus of translated texts—that I call a "translational corpus"—and a corpus consisting of comparably sampled non-translated texts in the same language (see Xiao and Yue 2009). They form a monolingual comparable corpus. To us, a multilingual comparable corpus samples different native languages, with its comparability lying in the matching or comparable sampling techniques, similar balance (i.e., coverage of genres and domains) and representativeness, and similar sampling period (see Section 7.3). By my definition, corpora containing different regional varieties of the same language (e.g., the International Corpus of English, ICE) are not comparable corpora because all corpora, as a resource for linguistic research, have "always been pre-eminently suited for comparative studies" (Aarts 1998: ix), either intralingually or interlingually. The Brown, LOB, Frown, and FLOB corpora are also used typically for comparing language varieties synchronically and diachronically. Corpora such as these can be labeled as "comparative corpora." They are not "comparable corpora" as suggested in the literature (e.g., Hunston 2002: 15).

Having clarified some terminological confusion in multilingual corpus research, it is worth pointing out the distinctions discussed here are purely for the sake of clarification. In reality, there are multilingual corpora that are a mixture of parallel and comparable corpora. For example, in spite of its name, the English–Norwegian Parallel Corpus (ENPC) can be considered as a combination of a parallel and comparable corpus. I will not discuss the state of the art of multilingual corpus research here. Interested readers are advised to refer to McEnery and Xiao (2007b).

Multilingual corpora often involve a writing system that relies heavily on non-ASCII characters. Character encoding is rarely an issue in corpus creation for alphabetical languages (e.g., English) that use ASCII characters. However, even languages that use a small number of accented Latin characters may have encountered encoding problems. For monolingual corpora of many other languages that use different writing systems, especially for multilingual corpora that contain a wide range of writing systems, encoding is all the more important if one wants to display the corpus properly or facilitate data interchange. For example, Chinese can be encoded using GB2312 (Simplified Chinese), Big5 (Traditional Chinese), or Unicode (UTF-8, UTF-7 or UTF-16). Both GB2312 and Big5 are 2-byte encoding systems that require language-specific operating systems or language-support packs if the Chinese characters encoded are to be

displayed properly. Language specific encoding systems such as these make data interchange problematic. It is also quite impossible to display a document containing both simplified and traditional Chinese characters using these encoding systems. As McEnery et al. (2000) note, the main difficulty in building a multilingual corpus of Asian languages is the need to standardize the language data into a single character set. Unicode is recommended as a solution to this problem (see McEnery and Xiao 2005). Unicode is truly multilingual in that it can display characters from a very large number of writing systems. From the Unicode Standard version 1.1 onward, Unicode is fully compatible with ISO 10646-1 (UCS). The combination of Unicode and XML is a general trend in corpus creation (see Xiao et al. 2004). As such, it is to be welcomed.

7.7 Multimodal Corpora

The corpora discussed so far in this chapter, whether spoken or written, have been assumed to be text-based; that is, spoken language is treated as if it is written. In this text-based approach to corpus creation, audio/video recordings of spoken data are transcribed, with the transcript possibly also including varying levels of details of spoken features (e.g., turn overlaps) and paralinguistic features (e.g., laughter). Corpus analysis is then usually undertaken on the textual transcript without reference to the original recording unless one is engaged in prosodic or phonetic research.

As noted in Section 7.5, a corpus is essentially a collection of samples of used language, which have been likened to a laboratory specimen out of its original habitat (Burnard 2005). While corpus markup can help to restore some contextual information, a large part of such information is lost, especially in transcripts of video clips. As Kress and van Leeuwen (2006: 41) observe, "a spoken text is never just verbal, but also visual combining with modes such as facial expressions, gesture, posture and other forms of self-presentation," the latter of which cannot be captured and transcribed easily, if at all. Consequently, "even the most detailed, faithful and sympathetic transcription cannot hope to capture" spoken language (Carter 2004: 26). As such, there has recently been an increasing interest in multimodal corpora. In this kind of corpora, annotated transcripts are aligned with digital audio/video clips with the help of time stamps, which not only renders the corpus searchable with the help of transcripts but also allows the user to access the segments of recordings corresponding to the search results. There are a number of existing multimodal corpora including, for example, the Nottingham Multi-Modal Corpus (NMMC, see Adolphs and Carter 2007), the Singapore Corpus of Research in Education (SCoRE, see Hong 2005), Padova Multimedia English Corpus (see Ackerley and Coccetta 2007), and the Spoken Chinese Corpus of Situated Discourse (SCCSD, see Gu 2002).

Multimodal corpora and multimodal concordancers are still in their infancy (Baldry 2006: 188). They are technically more challenging to develop than purely text-based corpora and corpus tools. However, given the special values of such corpora, and the advances of technologies (e.g., those that help to track and annotate gestures), multimodal corpora will become more common and more widely used in the near future.

7.8 Conclusions

This chapter has focused on corpus creation, covering the major factors that must be taken into account in this process. I have discussed both core issues relating to corpus design (e.g., corpus size, representativeness, and balance) as well as corpus processing (e.g., data collection, markup, and annotation), and peripheral issues such as multilingual and multimodal corpora.

One important reason for using corpora is to extract linguistic information present in those corpora. But it is often the case that in order to extract such information from a corpus, a linguistic analysis must first be encoded in the corpus. Such annotation adds value to a corpus in that it considerably extends the

range of research questions that a corpus can readily address. In this chapter, I have discussed corpus annotation in very general terms. The chapter that follows will explore annotation in greater depth.

References

Aarts, J. (1998) Introduction. In S. Johansson and S. Oksefjell (eds.), *Corpora and Cross-Linguistic Research*, pp. ix–xiv. Amsterdam, the Netherlands: Rodopi.

Ackerley, K. and Coccetta, F. (2007) Enriching language learning through a multimedia corpus. *ReCALL* 19(3): 351–370.

Adolphs, S. and Carter, R. (2007) Beyond the word: New challenges in analyzing corpora of spoken English. *European Journal of English Studies* 11(2): 133–146.

Aijmer, K. and Altenberg, B. (1996) Introduction. In K. Aijmer, B. Altenberg and M. Johansson (eds.), Language in contrast. *Papers from Symposium on Text-Based Cross-Linguistic Studies, Lund, Sweden*, March 1994, pp. 10–16. Lund, Sweden: Lund University Press.

Amsler, R. (2002) Legal aspects of corpora compiling. In *Corpora List Archive on 1st October 2002*. URL: http://helmer.hit.uib.no/corpora/2002-3/0256.html.

Aston, G. and Burnard, L. (1998) *The BNC Handbook*. Edinburgh, U.K.: Edinburgh University Press.

Atkins, S., Clear, J., and Ostler, N. (1992) Corpus design criteria. *Literary and Linguistic Computing* 7(1): 1–16.

Baker, M. (1993) Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, and E. Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair*, pp. 233–352. Amsterdam, the Netherlands: Benjamins.

Baker, M. (1995) Corpora in translation studies: An overview and some suggestions for future research. *Target* 7: 223–243.

Baker, M. (1999) The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics* 4: 281–298.

Baldry, A. P. (2006) The role of multimodal concordancers in multimodal corpus linguistics. In T. D. Royce and W. L. Bowcher (eds.), *New Directions in the Analysis of Multimodal Discourse*, pp. 173–214. London, U.K.: Routledge.

Barlow, M. (1995) *A Guide to ParaConc*. Huston, TX: Athelstan.

Barlow, M. (2000) Parallel texts and language teaching. In S. Botley, A. McEnery, and A. Wilson (eds.), *Multilingual Corpora in Teaching and Research*, pp. 106–115. Amsterdam, the Netherlands: Rodopi.

Baroni, M. (2009) Distributions in text. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics: An International Handbook* (Vol. 2), pp. 803–822. Berlin, Germany: Mouton de Gruyter.

Baroni, M. and Bernardini, S. (2004) BootCaT: Bootstrapping corpora and terms from the Web. In M. Lino, M. Xavier, F. Ferreire, R. Costa, and R. Silva (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*, Lisbon, Portugal, May 24–30, 2004.

Baroni, M. and Bernardini, S. (eds.). (2006) *Wacky! Working Papers on the Web as Corpus*. Bologna, Italy: GEDIT.

Belica, C. (1996) Analysis of temporal change in corpora. *International Journal of Corpus Linguistics* 1(1): 61–74.

Biber, D. (1988) *Variation Across Speech and Writing*. Cambridge, U.K.: Cambridge University Press.

Biber, D. (1993) Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257.

Bird, S. and Simons, G. (2000) *White Paper on Establishing an Infrastructure for Open Language Archiving*. URL: http://www.language-archives.org/docs/white-paper.html.

Burnard, L. (2002) *Validation Manual for Written Language Resources*. URL: http://www.oucs.ox.ac.uk/rts/elra/D1.xml.

Burnard, L. (2005) Metadata for corpus work. In M. Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*, pp. 30–46. Oxford, U.K.: AHDS.

Burnard, L. and Todd, T. (2003) Xara: An XML aware tool for corpus searching. In D. Archer, P. Rayson, A. Wilson, and A. McEnery (eds.), *Proceedings of Corpus Linguistics 2003*, Lancaster, U.K., pp. 142–144. Lancaster, U.K.: Lancaster University.

Carter, R. (2004) Grammar and spoken English. In C. Coffin, A. Hewings, and K. O'Halloran (eds.), *Applying English Grammar: Corpus and Functional Approaches*, pp. 25–39. London, U.K.: Arnold.

Collins (2007) *Collins English Dictionary* (9th ed.). Toronto, Canada: HarperCollins.

Cooper, D. (2003) Legal aspects of corpora compiling. In *Corpora List Archive on 19th June 2003*. URL: http://helmer.aksis.uib.no/corpora/2003-1/0596.html.

Cornish, G. P. (1999) *Copyright: Interpreting the Law for Libraries, Archives and Information Services* (3rd ed.). London, U.K.: Library Association Publishing.

Dekkers, M. and Weibel, S. (2003) State of the Dublin core metadata initiative. *D-Lib Magazine* 9(4). URL: http://www.dlib.org/dlib/april03/weibel/04weibel.html.

Dipper, S. (2005) XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005* (*BXML 2005*), Berlin, Germany, pp. 39–50.

Dipper, S. (2008) Theory-driven and corpus-driven computational linguistics, and the use of corpora. In A. Ludeling and M. Kyto (eds.), *Corpus Linguistics: An International Handbook* (Vol. 1), pp. 68–96. Berlin, Germany: Mouton de Gruyter.

Evert, S. (2006) How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54(2): 177–190.

Givon, T. (1995) *Functionalism and Grammar*. Amsterdam, the Netherlands: John Benjamins.

Granath, S. (2007) Size matters—Or thus can meaningful structures be revealed in large corpora. In R. Facchinetti (ed.), *Corpus Linguistics 25 Years On*, pp. 169–185. Amsterdam, the Netherlands: Rodopi.

Granger, S. (1996) From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, and M. Johansson (eds.), *Language in contrast. Symposium on Text-based Cross-linguistic Studies, Lund, Sweden*, March 1994, pp. 38–51. Lund, Sweden: Lund University Press.

Gu, Y. (2002) Towards an understanding of workplace discourse. In C. N. Candlin (ed.), *Research and Practice in Professional Discourse*, pp. 137–86. Hong Kong: City University of Hong Kong Press.

Hakulinen, A., Karlsson, F., and Vilkuna, M. (1980) *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus*. Department of General Linguistics, University of Helsinki, Helsinki, Finland, Publications No. 6.

Hausser, H. (1999) *Functions of Computational Linguistics*. Berlin, Germany: Springer-Verlag.

Hong, H. (2005) SCORE: A multimodal corpus database of education discourse in Singapore schools. In *Proceedings of Corpus Linguistics 2005*. http://www.corpus.bham.ac.uk/pclc/ScopeHong.pdf

Hundt, M., Sand, A., and Siemund, R. (1998) *Manual of Information to Accompany the Freiburg-LOB Corpus of British English* (*'FLOB'*). URL: http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM.

Hundt, M., Sand, A., and Skandera, P. (1999) *Manual of Information to Accompany the Freiburg-Brown Corpus of American English* (*'Frown'*). URL: http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM.

Hundt, M., Biewer, C., and Nesselhauf, N. (eds.). (2007) *Corpus Linguistics and the Web*. Amsterdam, the Netherlands: Rodopi.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge, U.K.: Cambridge University Press.

Ide, N. (1998) Corpus encoding standard: SGML guidelines for encoding linguistic corpora. In *LREC-1998 Proceedings*, Granada, Spain, pp. 463–470.

Ide, N. and Priest-Dorman, G. (2000) *Corpus Encoding Standard—Document CES 1*. URL: http://www.cs.vassar.edu/CES/.

Ide, N., Patrice, B., and Romary L. (2000) XCES: An XML-based encoding standard for linguistic corpora. In *LREC-2000 Proceedings*, Athens, Greece, pp. 825–830.

Johansson, S., Leech, G., and Goodluck, H. (1978) *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Oslo, Norway: University of Oslo.

Keller, F. and Lapata, M. (2003) Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29(3): 459–484.

Kilgarriff, A. (2002) Legal aspects of corpora compiling. In *Corpora List Archive on 1st October 2002*. URL: http://helmer.hit.uib.no/corpora/2002-3/0253.html.

Kilgarriff, A. and Grefenstette, G. (eds.). (2003) Special Issue on Web as Corpus. *Computational Linguistics* 29(3): 333–502.

Kress, G. and van Leeuwen, T. (2006) *Reading Images: The Grammar of Visual Design* (2nd ed.). London, U.K.: Routledge.

Krishnamurthy, R. (2000) Size matters: Creating dictionaries from the world's largest corpus. In *Proceedings of KOTESOL 2000: Casting the Net: Diversity in Language Learning*, Taegu, Korea, pp. 169–180.

Kučera, H. and Francis, W. (1967) *Computational Analysis of Present-Day English*. Providence, RI: Brown University Press.

Leech, G. (1991) The state of art in corpus linguistics. In K. Aijmer and B. Altenberg (eds.), *English Corpus Linguistics*, pp. 8–29. London, U.K.: Longman.

Leech, G. (1997) Introducing corpus annotation. In R. Garside, G. Leech, and A. McEnery (eds.), *Corpus Annotation*, pp. 1–18. London, U.K.: Longman.

Manning, C. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McEnery, A. and Wilson, A. (1996/2001) *Corpus Linguistics* (2nd ed. 2001). Edinburgh, U.K.: Edinburgh University Press.

McEnery, A. and Xiao, R. (2005) Character encoding in corpus construction. In M. Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*, pp. 47–58. Oxford, U.K.: AHDS.

McEnery, A. and Xiao, R. (2007a) Parallel and comparable corpora: What is happening? In M. Rogers and G. Anderman (eds.), *Incorporating Corpora: The Linguist and the Translator*, pp. 18–31. Clevedon, U.K.: Multilingual Matters.

McEnery, A. and Xiao, R. (2007b) Parallel and comparable corpora: The state of play. In Y. Kawaguchi, T. Takagaki, N. Tomimori, and Y. Tsuruga (eds.), *Corpus-Based Perspectives in Linguistics*, pp. 131–145. Amsterdam, the Netherlands: John Benjamins.

McEnery, A., Baker, P., Gaizauskas, R., and Cunningham, H. (2000) EMILLE: Building a corpus of South Asian languages. *Vivek: A Quarterly in Artificial Intelligence* 13(3): 23–32.

McEnery, A., Baker, P., and Cheepen, C. (2001) Lexis, indirectness and politeness in operator calls. In C. Meyer and P. Leistyna (eds.), *Corpus Analysis: Language Structure and Language Use*. Amsterdam, the Netherlands: Rodopi.

McEnery, A., Xiao, R., and Tono, Y. (2006) *Corpus-Based Language Studies: An Advanced Resource Book*. London, U.K.: Routledge.

Ostler, N. (2008) Corpora of less studied languages. In A. Ludeling and M. Kyto (eds.), *Corpus Linguistics: An International Handbook* (Vol. 1), pp. 457–484. Berlin, Germany: Mouton de Gruyter.

Otlogetswe, T. (2004) The BNC design as a model for a Setswana language corpus. In *Proceeding of the Seventh Annual CLUK Research Colloquium*, pp. 93–198. University of Birmingham, Edgbaston, U.K., January 6–7, 2004.

Piao, S., Wilson, A., and McEnery, A. (2002) A multilingual corpus toolkit. *Paper Presented at the Fourth North American Symposium on Corpus Linguistics*, Indianapolis, IN, November 1–3, 2002.

Santos, D. (1996) Tense and aspect in English and Portuguese: A contrastive semantical study. PhD thesis, Universidade Tecnica de Lisboa, Lisbon, Portugal.

Scott, M. (2003) *WordSmith Tools Manual*. URL: http://www.lexically.net/wordsmith/version4/.

Shimazumi, M. and Berber-Sardinha, A. (1996) Approaching the assessment of performance unit (APU) archive of schoolchildren's writing from the point of view of corpus linguistics. *Paper Presented at the TALC'96 Conference*, Lancaster University, Lancaster, U.K., August 11, 1996.

Sinclair, J. (1991) *Corpus Concordance Collocation*. Oxford, U.K.: Oxford University Press.

Sinclair, J. (2004) *Trust the Text: Language, Corpus and Discourse*. London, U.K.: Routledge.

Sinclair, J. (2005) Corpus and Text: Basic Principles. In M. Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*, pp. 1–20. Oxford, UK: AHDS.

Sperberg-McQueen, C. M. and Burnard, L. (eds.). (2002) *TEI P4: Guidelines for Electronic Text Encoding and Interchange* (*XML Version*). Oxford, U.K.: Text Encoding Initiative Consortium.

Spoor, J. (1996) The copyright approach to copying on the Internet: (Over)stretching the reproduction right? In H. Hugenholtz (ed.), *The Future of Copyright in a Digital Environment*, pp. 67–80. Dordrecht, the Netherlands: Kluwer Law International.

Teubert, W. (2000) Corpus linguistics—A partisan view. *International Journal of Corpus Linguistics* 4(1):1–16.

Thompson, H. and McKelvie, D. (1997) Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe'97*, Barcelona, Spain, May 1997. URL: http://www.ltg.ed.ac.uk/~ht/sgmleu97.html.

Váradi, T. (2000) Corpus linguistics—linguistics or language engineering? In T. Erjavec and J. Gross (eds.), *Information Society Multi-Conference Proceedings Language Technologies*, pp. 1–5. Ljubljana, Slovenia, October 17–18, 2000.

Wittenburg, P., Peters, W., and Broeder, D. (2002) Metadata proposals for corpora and lexica. In *LREC-2002 Proceedings*, Las Palmas, Spain, pp. 1321–1326.

Xiao, R. (2008) Well-known and influential corpora. In A. Lüdeling and M. Kyto (eds.), *Corpus Linguistics: An International Handbook* (Vol. 1), pp. 383–457. Berlin, Germany: Mouton de Gruyter.

Xiao, R. and Yue, M. (2009) Using corpora in translation studies: The state of the art. In P. Baker (ed.), *Contemporary Approaches to Corpus Linguistics*, pp. 237–262. London, U.K.: Continuum.

Xiao, R., McEnery, A., Baker, P., and Hardie, A. (2004) Developing Asian language corpora: Standards and practice. In *Proceedings of the Fourth Workshop on Asian Language Resources*, Sanya, Hainan Island, pp. 1–8, March 25, 2004.

Yates, F. (1965) *Sampling Methods for Censuses and Surveys* (3rd ed.). London, U.K.: Charles Griffin and Company Limited.