Unit 8 Going solo: DIY corpora

8.1 Introduction

As noted in unit 7, while there are many ready-made corpora, one may find it necessary to build one's own corpus to address a particular research question. In this unit we will discuss the principal factors one should consider when constructing such corpora. In exploring DIY ('do-it-yourself') corpora, we will revisit the key concepts introduced in units 2 – 4, and explain how they are a useful guide to the creation of DIY corpora. Additionally, we will overview some of the readily available tools which may help in the process of DIY corpus building, especially where web-based material is being used.

8.2 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. 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 like Brown 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 BNC and the Bank of English it appears somewhat small.

The availability of suitable data, especially in machine-readable form, seriously affects corpus size. In building a balanced corpus according to fixed proportions, 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. While it is often possible to transfer paper-based texts into electronic form using OCR 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 building.

Another factor that potentially limits the size of a DIY corpus is copyright. 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 builder 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 (see unit 9 for further discussion of copyright issues relevant to corpus building).

No matter how important legal considerations may seem, however, 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 like the number of present and past tense verbs in English, for example, a sample of 1,000 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, which can be modelled as contrasting with all others of the same category (cf. Santos 1996:11). In contrast, corpora employed in quantitative studies of grammatical devices are 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. semantic annotation and pragmatic annotation, see unit 4) are necessarily small. Fourth, many corpus tools set a ceiling on the number of concordances that can be extracted, e.g. WordSmith version 3 can extract a maximum of 16,868 concordances (version 4 does not has this limit). This makes it inconvenient for a frequent linguistic feature to be extracted from a very large corpus (see case studies 3 and 5 for a solution). 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, we 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. The point we 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.

8.3 Balance and representativeness

As noted in unit 2, representativeness is a qualifying feature of a corpus. To achieve this quality, balance and sampling are important. 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, i.e., a corpus should only be as representative as possible of the language variety under consideration. ICE is a good example of this, as each of its component subcorpora represents one national or regional variety of English; Brown and LOB represent written American and British English in 1961 respectively, while learner corpora such as ICLE represent language varieties used by learners from various L1 backgrounds or at different proficiency levels. Even a general corpus such as the BNC with 100 million words only represents modern British English in a specific time and geographical frame (see unit 7 for a description of these corpora). It is also important to note that the lower proportion of spoken data in corpora such as the BNC does not mean that spoken language is less important or less wide spread than written language. This is so simply because spoken data is more difficult and expensive to capture than written data. Corpus building is of necessity a marriage of perfection and pragmatism.

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 model requires knowledge of which genre is used how often 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). Readers will have an opportunity to explore the validity of the Brown model using the techniques introduced in case study 5 in Section C of this book. 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 (cf. unit 10.15). Likewise, conclusions drawn from a particular corpus must be treated as deductions rather than facts (cf. also Hunston 2002: 23).

8.4 Data capture

For pragmatic reasons, electronic data is preferred over paper-based materials in building DIY corpora. The world-wide-web (WWW) is an important source of machine-readable data for many languages. 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, e.g. XML (see unit 3.3). In this section, we 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 which facilitate downloading all of the web pages on a selected website 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. WordSmith version 4, 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 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 Multilingual Corpus Toolkit (MLCT, see Piao, Wilson and McEnery 2002). This toolkit is available at the website accompanying this book (see the Appendix). 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, POS 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) 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 (see unit 8.5).

8.5 Corpus markup

As noted in unit 3, corpus markup is a basic step in corpus construction. Markup usually provides textual (e.g. paragraph and sentence) and contextual information (e.g. text type, speaker gender and bibliographic source). Textual information is useful for studying textual organization while contextual information is important in recovering the situation in which a particular corpus sample was produced, as corpora usually consist of small isolated samples extracted from larger texts. Markup also helps to organize corpus data in a structured way and enables explorations in language variation (see case study 4).

While markup is clearly essential for corpus construction, the degree of markup needed is closely related to the research question. If a corpus is constructed to compare different genres, markup must show text type information; likewise, if a spoken corpus is built to explore language variation across sociolinguistic variables, then the relevant features such as speaker age, gender and social class must be marked up. Extensive metadata (see unit 3) is encoded in general corpora such as the BNC which can serve for multiple purposes.

However, the markup process is usually time consuming. Excessive markup may also make a corpus less readable to a casual corpus user not viewing the corpus through a markup-aware browsing tool that can hide up such markup. For specialized corpora which use homogeneous data (e.g. articles downloaded from hate newsgroups on the Internet), we suggest that only basic textual information, namely paragraphs and sentences, be marked up, as this type of markup can be conducted relatively easily using available software. As noted, Xaira can pre-process texts in XML format. The MLCT also inserts paragraph and sentence marks automatically.

8.6 Corpus annotation

We noted in unit 4 that corpus annotation is closely related to, but different from markup, and can take many forms such as POS tagging, parsing, semantic annotation and so on.

The form of corpus annotation one should undertake on a corpus is primarily dependent upon one’s research question. For spoken corpora designed for use in speech recognition, the annotation of prosodic features is essential whereas syntactic parsing is less important. Likewise, corpora constructed for grammatical study should be POS tagged, and preferably also parsed while those used in the study of semantics may profitably include semantic annotation. In addition to the research question, a major consideration in corpus annotation is the precision rate with which a form of annotation can be undertaken automatically. As far as the English language is concerned, automatic lemmatization and POS tagging have achieved very high success rates (typically over 97%). Significant progress has also been made in parsing and semantic annotation (see unit 4). In contrast, many other forms of annotation, such as the annotation of coreference, pragmatic features, and speech and thought representation, either cannot be conducted automatically or the output of such automatic processing requires substantial manual correction. Automatic POS tagging can also be successfully undertaken for many other languages such as French, Spanish, Chinese and Korean. Given the current status of automated corpus annotation, it is usually possible for DIY corpora intended for general language study to be annotated for parts-of-speech. Given that errors are inevitable in automatic annotation, corpus size should be taken into account so as to neutralize these errors. With the same precision rate of annotation, a corpus of one hundred thousand words is clearly more reliable than a corpus containing merely a few thousand words, assuming that the errors are relatively random.

8.7 Character encoding

Character encoding is rarely an issue for alphabetical languages (e.g. English) that use ASCII characters. For many other languages that use different writing systems, especially for multilingual corpora that contain a wide range of writing systems, encoding is 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 using these encoding systems.

Unicode solves all of these problems. Unicode is truly multilingual in that it can display the characters of a very large number of writing systems. Hence, a general trend in corpus building is to encode corpora (especially multilingual corpora) using Unicode (e.g. the EMILLE corpora, cf. McEnery, Baker, Gaizauskas and Cunningham 2000). Corpora encoded in Unicode can also take advantage of the latest Unicode-compliant concordancers such as Xaira (Burnard and Todd 2003) and WordSmith version 4 (Scott 2003).

8.8 Unit summary and looking ahead

This unit discussed the principal considerations involved in the creation of DIY corpora, namely, corpus size, balance and representativeness, data capture, corpus markup, annotation and character encoding. Throughout our discussion we have emphasized that almost every decision (with the exception of corpus encoding) is closely related to the research question one wishes to address using the corpus, though pragmatic considerations such as the availability of machine-readable data and the reliability of automatic processing tools may also affect one’s decisions. We have also shown that the key concepts introduced in units 2–4 are a useful guide to the construction of DIY corpora.

A key theme of this unit is the usefulness of the world-wide-web in the construction of DIY corpora. The Internet is an important source of machine-readable data. It also provides many corpus processing tools such as those that facilitate downloading web pages and markup conversion.