5
Building a written corpus
What are the basics?

Mike Nelson

1. Introduction: what does building a written corpus entail?

It is clear at the outset that there are a huge number of variables involved in building a written corpus that need to be considered before beginning what can be a very daunting task. Some of these issues are related to all corpus creation, and others can be seen to be specific to written corpora. In general terms, the very first question to ask oneself is 'Do I have to do this?' The exponential rise of available electronic corpora over the last twenty years has provided the academic community with an enormous amount of ready-made data that can be accessed easily on-line. Thus the question of 'Why build a new corpus?' must be very seriously considered. If you do decide to continue, then the purpose of the corpus must be very clear – what do you want to achieve by creating the corpus and what will it be used for? Depending on your answer to these questions, there then come crucial decisions to make regarding the size of the corpus, of how it should be balanced, the sampling methods to use, the kinds of texts that should be used, the use of full, or samples of, text and how representativeness could be achieved. Further issues concern whether you want to create a corpus for a specialist purpose or for more general purposes. When you have planned your corpus and the content it will contain, will you stick to this plan rigidly, or will you take texts from where you can readily get them? Finally, there are psychological elements attached to corpus creation: planning, finding, gathering and formatting data over a long period of time can be mentally very taxing and requires a project planner's approach to the task.

There are also further issues that are specific to written corpora. There is a commonly held belief that creating a written corpus is easier than creating a spoken corpus. While this may to some extent be true, there are still many difficulties involved in the creation of a written corpus: for example, in the choosing of texts, gaining access to the texts you require and dealing with the process of turning text into a computer-readable format, followed by storage and analysis. This chapter will attempt to provide assistance and advice on these issues and hopefully provide an insight into the whole process of building a written corpus from inception to completion. In the first part of the chapter the general areas mentioned above will be considered with reference to the literature and also with specific reference to how some well-known written corpora have been designed. The next section deals

with the gathering, computerising and organisation of written data, including a brief review of optical character recognition (OCR) software. There is then a mention of storage and analysis of data. In my own research, I have created three different written corpora that each presented different challenges. As the chapter progresses, these corpora are referred to so that the theory behind corpus building can be given a practical application.

2. Planning a written corpus

There is common consensus in the world of corpora that a corpus is 'not simply a collection of texts' (Biber *et al.* 1998: 246). By implication, therefore, a clear and detailed plan needs to be created long before anything else can be done. This notion should also be tempered with the realities of data collection (Kilgariff *et al.* 2006 discuss detailed corpus design and the comparison made between the design document and final corpus composition). Whatever the difficulties, however, the initial planning document is of great importance and the following section will now address the elements that need to be considered when designing a written corpus. The first design element, however, is whether a design is needed at all.

Why create a new written corpus?

Reading the chapter by David Lee in this book will give you an idea of the vast number of corpora available today. His website presents, among other things, an overview of corpora currently available (see also the *CALPER* website of the Pennsylvania State University, USA). Thus, before any decision is made regarding corpus creation, a great deal of thought needs to be given to whether or not an existing corpus would serve the purposes of your research. It may be that a given corpus may have a small section within it that would be suitable, or possibly could be used to supplement a corpus that you may later create. Essentially, you have to ask yourself whether creating a new corpus will bring something new to both your research and the research community. One further factor that must also be taken into consideration is that although many corpora do already exist, access to them may not always be possible, leaving you in the unfortunate position of having to re-invent the wheel. Contact with the creators at an early stage of your work should tell you whether this will be a problem or not.

If you decide to go ahead with creating a written corpus, there must follow several decisions that need to be made in the planning process. All these decisions depend on the purpose to which you will put your corpus. The different types of corpora are discussed elsewhere in this book and so will not be discussed in detail here, but, whatever the purpose of your corpus, one of the first considerations is how big it should be.

How big should the written corpus be?

The question of the size of corpora has been central to recent corpus development, and there has been the overriding belief among many corpus creators that 'biggest is best'. Briefly, the discussion can be seen in terms of creating corpora to be as large as possible, for example for lexicographic projects, as opposed to creating smaller, more specialist corpora, often for pedagogical purposes. Thus, the purpose to which the corpus is ultimately put is a critical factor in deciding its size. If you are reading this book and considering corpus creation, it

is probable that a smaller corpus is more likely to be your aim, but some issues of what is considered to be an adequate size for a corpus need to be discussed (see Sinclair 2002 for a succinct summary of this issue, which he refers to as the 'incredible shrinking corpora'.)

While the early corpora of the 1960s were modest in size by the standards of today, there was already a pervasive attitude that bigger corpora would be better. Halliday and Sinclair (1966) proposed the necessity of a corpus of around twenty million words. This was unrealistic at the time, but between the 1960s and 1980s corpora rapidly grew in size, encompassing the 'three generations' of Leech (1991) from several hundred thousand words to several hundred million (British National Corpus [BNC], Bank of English, Cambridge International Corpus [CIC], which stands at one billion words). This view of the need for large corpora was summed up by Sinclair when he said that 'The only guidance I would give is that a corpus should be as large as possible and keep on growing' (1991: 18). Sinclair based this need for large corpora on the fact that words are unevenly distributed in texts and that most words occur only once. Thus 'In order to study the behaviour of words in texts, we need to have available quite a large number of occurrences' (Sinclair 1991: 18). While this view of corpora was the prevailing one, it did not go unchallenged.

A movement then grew in the 1990s that was more concerned with corpus exploitation than corpus exploration (Ma 1993; Tribble 1997, 1998; Flowerdew 1998). This movement saw the value of smaller corpora and stressed their pedagogical purpose over their lexicographical potential. Small corpora, it was held, can be very useful, providing they can offer a 'balanced' and 'representative' picture of a specific area of the language. This recognition of a need for smaller, more specialised corpora increased. Ma (1993) noted that the division of corpora based on size is between corpora that are used for examining 'general' English, and those that are used for examining more specific areas of language use. The usefulness of smaller corpora was seen to be a *pedagogical* usefulness, as opposed to a *general explorative* usefulness – Ma related this utility of smaller corpora to 'groups of learners' (1993: 17). He also listed a number of 'pedagogic' corpora ranging in size from a corpus of philosophy texts at 6,854 words to a corpus of over one million words (Ma 1993: 17). Tribble's use of 'exemplar texts' to exemplify genres, while keeping the overall size of the corpus down to manageable levels, continued this trend and he noted that 'If you are involved in language teaching rather than lexicography, single word lists from small selective corpora can be seriously useful' (Tribble 1997). The further pedagogical usefulness of small corpora was suggested by Howarth (1998) in relation to teaching non-native speakers and de Beaugrande (2001) distinguished between small specialist corpora and what he terms 'learnable' corpora that are built using examples of language that match the fluency levels of specific groups of learners.

To summarise, corpora that have been used for lexicographical purposes – looking at the whole language – have, perhaps by necessity, always been created to be as large as possible. However, the need for smaller corpora – looking at specific areas of the language – has been recognised, especially in relation to teaching and use in the language classroom. Kennedy (1998: 68) noted that researchers should therefore 'bear in mind that the quality of the data they work with is at least as important [as the size]'.

Case study: Determining the size of a corpus

The Business English Corpus (BEC) was created between 1998 and 2000 to represent the language used in business by native speakers (see website for full details.) The corpus

has both written and spoken elements, but the principles used to determine size can be applied to written corpora; three main criteria were used: *pragmatic*, *historical* and *pedagogical*. Practical considerations must always play a part in corpus creation and the larger lexicographical corpora such as the BNC and COBUILD that run into hundreds of millions of words were not a feasible option for one lone researcher to undertake. Meyer (2002: 32–3) notes that for the American component of ICE it was calculated that eight hours of work was needed to process a 2,000-word sample of written text. Thus resources must always be weighed against projected corpus size. Once the decision had been made that a smaller corpus was to be created, the figure of one million words was arrived at as a result of the two remaining criteria. The second criterion used was that of historical precedent. The figure of one million seems to be a 'magic' number in terms of older corpora size. Many influential older corpora – what Leech (1991) called the 'first generation' – were around the one-million-word mark or often much smaller in size. Examples of this are the Survey of English Usage (SEU) at University College London at one million words, and the Brown Corpus. There was, therefore, a historical reason for the one-million-word target size of the BEC. In addition to this tradition, smaller, specialist corpora, of which the BEC is one, have often used the one-million-word mark (or smaller) as a target number of running words. Comparative specialist corpora to the BEC would be the Guangzhou Petroleum English Corpus of 411,612 words, the Hong Kong University of Science and Technology (HKUST) Computer Science Corpus at one million words and the Århus Corpus of Contract Law, also at one million words. Fang (1993) in describing the creation of the HKUST corpus, specifically referred to the older generation of corpora, giving their size as one of the reasons for their choice of size. Additionally, he added that 'one million words represent a reasonably large proportion of the finite subset of the language under study' (Fang 1993: 74). As the BEC is not meant as a general English corpus, and in line with the specialist corpora noted above, one million words was deemed a reasonable sample size in order to achieve a representative picture of Business English. The final reason for the one-million-word size of the BEC was pedagogical. Smaller corpora enable easier access to the data found in them. This in turn leads to easier transferral of results to the classroom.

One final point can be mentioned. In some cases, the overall size of a corpus can be secondary to the need for adequate sampling. Thus, a second, written corpus of Business English textbooks that was created for the same project (Nelson 2000) had no predetermined goals for overall size. In this case, the sampling procedures, to be described later in this chapter, set the final number of books to be included at thirty-three, which in turn affected the final size of the corpus, which came to 593,294 running words (Meyer 2002: 33 refers to Biber's 1993 use of statistical procedures to determine corpus size by calculating the frequency of occurrences of linguistic features in a text).

Once the size of the corpus has been considered, the issues of sampling, balance and representativeness are then the next matters to be dealt with.

3. Sampling, balancing and making your written corpus representative

These three issues cannot be seen in isolation: in order to achieve an acceptable level of representativeness, the problems of sample size and balance must also be addressed. However, it should also be remembered that the corpus itself is a sample and needs to be

representative of a given aspect or aspects of language so that 'The first step towards achieving this aim is to define the whole of which the corpus is to be a sample'(Renouf 1987: 2).

Defining the sample base

Any corpus creator is faced with a 'chicken and egg' situation. In order to study language, be it general or specific, one must first decide what that language is, what defines it and where it can be found. As a result of this 'chicken and egg' situation, sampling and representativeness are difficult problems. These problems have dogged corpus linguists since the beginning and still do today. Clear (1992) gave three main reasons why sampling can be problematic for the corpus linguist. First, there is the problem noted above, that the population from which the sample is to be drawn is poorly defined. Second, 'there is no obvious unit of language which is to be sampled and which can be used to define the population' (Clear 1992: 21). Finally, considering the size of any aspect of language, the researcher can never be sure that all instances have been accounted for satisfactorily, and Clear on the 'Corpora' bulletin board noted as follows:

> I have a favourite analogy for corpus linguistics: it's like studying the sea. The output of a language like English has much in common with the sea; e.g. – both are very very large … – and difficult to define precisely, – subject to constant flux, currents, influences, never constant, – part of everyday human and social reality. Our corpus building is analogous to collecting bucketfuls of sea water and carrying them back to the lab. It is not physically possible to take measurements and make observations about all the aspects of the sea we are interested in *in vivo*, so we collect samples to study *in vitro*.
>
> (Clear 1997, personal communication)

More recently, Kilgariff *et al.* noted that 'There are no generally agreed objective criteria that can be applied to this task: at best, corpus designers strive for a reasonable representation of the full repertoire of available text types' (Kilgariff *et al.* 2006: 129). A corpus, virtually by definition, is therefore biased to a greater or lesser extent. Yet despite the difficulties, sampling is still necessary. We need to determine how many samples will be representative, how big the samples should be, and what kind of samples to use (full text or extracts).

How many samples?

The number of samples to be used in a corpus must be determined by the area under study, the linguistic variation that can be found in that area and the final purpose to which the corpus will be put. In influential work done in the 1980s and 1990s, Biber argued that the internal variation found within a given genre should determine how much of that genre should be included in a corpus (Biber 1988). However, Douglas (2003) reporting on work done with the Scottish Corpus of Texts and Speech (SCOTS Project), while stressing the importance of familiarity with the linguistic idiosyncrasies of the language varieties under analysis, noted that it is difficult to know where one variety ends and another begins, making the problem of linguistic variation more thorny. Yet there is help to be found with the problem of the number of samples to use. Meyer

(2002: 42–4) gives a useful summary of sampling techniques used in the social sciences by establishing a 'sampling frame' which is achieved 'by identifying a specific population that one wishes to make generalizations about' (Meyer 2002: 42). This methodical use of pre-selected samples is known as 'probability sampling'.

A good example of this approach was in the creation of the BNC. In order to determine the number of samples used, various sources of information were utilised to gain an overview of the chosen area. In the written section of the corpus the creators first made the distinction between language that is produced (written) and language that is received (read). They then gathered information on written texts from catalogues of books published per annum, best-seller lists, prize-winners, library lending statistics, lists of current magazines and periodicals and periodical circulation figures. These all dealt with published data and the project faced a problem with unpublished data; therefore, in this area intuition had to be used (see the BNC User Manual, available online).

Another approach to sampling that is commonly used, though sometimes frowned upon, is that of 'non-probability' sampling. This means, in its extreme form, essentially just taking samples from where it is possible to get them; it is often termed 'convenience' or 'opportunistic' sampling. This practice has been widespread in corpus creation and has even been used in large and prestigious projects such as the BNC (see Burnard 2001 for an explanation of what went wrong in the design of the BNC). A further common practice is to use a combination of both: to set out a plan for the corpus and then attempt to fulfil it but then adopt a certain flexibility according to what texts can be obtained.

Whatever the approach used to determine the number of samples, the use of common sense, pragmatics and intuition seem to play a role in even the most carefully planned corpus. Further, the purpose of the corpus exerts a keen influence at this stage of design. The number of samples needs to be adjusted according to what is going to be studied: more samples for more general language issues and fewer for more specific. The next stage is then to determine the size and type of sample that is then gathered.

Sample size and make-up

There has long been a debate in the literature regarding optimal sample sizes in corpora. Early corpora used sample sizes of around 2,000 words randomly taken from carefully selected texts. Other writers criticised the small sample size of the early corpora but have suggested that an increase to around 20,000 words would provide a sample of adequate size to be representative of a genre. Oostdijk (1991) and Kennedy (1998) took this line and Oostdijk suggested that 'A sample size of 20,000 words would yield samples that are large enough to be representative of a given variety' (Oostdijk 1991: 50). In larger projects such as the BNC, target sample sizes of 40,000 words have been used. Once again, the size of your samples will depend on what linguistic features you are attempting to elucidate. In the BEC, which was discussed earlier, a minimum sample size of 20,000 words was decided upon for each genre. Yet this in itself can cause problems. There is, as Biber has pointed out, considerable variation within genre, in that for some genres 20,000 words would provide an adequate sample size. For others this would not. A good example of this dilemma was encountered in creating the BEC. For approximately 20,000 words, 114 faxes were collected from different sources. However, in the category of 'business books', 20,000 words would not cover even one book. For this reason, a larger sample size of 50,000 words was used for books, taking five 10,000 word samples

from five different books. It is clear from this that for the faxes, all the text in the faxes was used: in the books only an extract was taken. This leads us to the next issue in sample make-up: that of the use of whole text or only extracts.

The choice of whether to use extracts or full texts has a significant impact on the kind of data that can be studied. Studies of discourse have shown us that 'few linguistic features of a text are distributed evenly throughout' (Stubbs 1996: 32) with the result that the use of only a small 'sample' of given text will inevitably miss out a great many features present. This is especially important when studying genre. Studies into genre have noted how certain linguistic features are typical of certain parts of a text and an approach to corpus creation that only takes extracts at random will fail to gain a representative sample in this respect. Thus, as with other aspects of corpus design, the purpose to which the corpus will be put is critical in deciding whether to use whole texts or not. The BNC, with 40,000 word extracts, did not use full text (partially for copyright reasons). They used continuous text within a whole, cutting the sample at a logical point such as at the end of a chapter. This approach is well suited to study of general language. In the BEC written section, which was concerned with the specialist language of business, whole texts were used wherever possible. The corpus was first categorised into language used for writing about business (books, journals, magazines) and then into language used for actually doing business. A breakdown of the written section can be seen in Tables 5.1 and 5.2.

Table 5.1 Writing about business

Part of corpus	*Tokens*	*Contents*
Business books	53,470	5 extracts from different books (approx. 10,000 words each)
Business newspapers	64,291	121 articles
Business journals & magazines	78,846	52 articles
Total	**196,607**	

Table 5.2 Writing to do business

Part of corpus	*Tokens*	*Contents*
Annual reports	34,537	3 annual reports
Bus press releases	21,656	29 business press releases
Business contracts	29,602	13 contracts/agreements
Business faxes	23,105	114 faxes
Business letters	26,793	94 letters
Business reports	62,908	17 reports
Company brochures	23,239	13 company brochures
Emails	28,857	202 e-mails
Job advertisements	22,293	87 job advertisements
Manuals	21,160	5 manuals
Memos	12,542	47 memos
Minutes	34,805	15 sets of minutes
Product brochures	26,175	19 product brochures
Quotations	8,997	21 quotations
Miscellaneous	2,427	Oht, job description and agendas
Total	**379,096**	

Balance and representativeness

The discussions on size and sampling above have necessarily touched on questions of representativeness and balance. When attempting to balance a corpus in order to give a representative view of the language chosen it is necessary to ask the question 'Representative of what?' In reality, there are so many variables that the notion of 'representativeness' can almost be seen as a 'non-concept'. Kennedy notes that 'it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre or subject field or topic' (Kennedy 1998: 62). Any attempt at corpus creation is therefore a compromise between the hoped for and the achievable. Yet we have already seen from work done on the BNC how representativeness and balance can be attempted by carefully stratifying the corpus beforehand. In the written component of the BNC, texts were included according to three selection criteria: domain, time and medium. In this way it was felt first that a 'microcosm' of the language could be presented and second that different types of texts could be compared with each other. However, retrospectively, Burnard (2001) admitted that the availability of electronic texts in some areas led to a skewing of data input to the BNC.

Case study: The Published Materials Corpus

In the corpus designed to represent published Business English materials, the Published Materials Corpus (PMC; Nelson 2000), balance and representativeness were achieved by surveying the popularity of use of books in the general market in order to provide an overview of those books actually in use at the time.

For the purpose of the study, the initial population for the PMC was defined as Business English materials published in the UK between 1986 and 1996. In order to gain a representative sample of this population, the PMC presented different problems from the BEC. A representative sample of Business English materials was needed in order to analyse exactly what the language of Business English teaching materials is. The problem was solved in the following way. In 1997, seven major distributors of EFL materials were contacted by phone and were asked to provide a list of their best-selling Business English titles of 1996. Of these, five of the seven responded. Actual sales figures were not available, but the rank order of popularity was obtained from each bookshop. Once the lists were collected, the books were ranked according to their position of popularity at each bookshop and averaged out over the five, so that an overall ranking of popularity was obtained covering all five bookshops. A total of thirty-eight books were obtained for the final list. The main factor, therefore, behind the content of the PMC was popularity of use of the books – it was considered of prime importance to have a corpus that represented the Business English books that teachers and students actually use. Once the sample had been gained in the manner described above, the books included could be broken down in terms of type of book included, gender of author and those books focusing on one or more of the 'four skills'.

Once specifications had been made as to the content of the corpora, the data had to be actually collected and entered into the computer. This stage represented the most difficult of all in this research and, in all, took just over three years to complete.

4. Gathering, computerising and organising written texts

Perhaps the greatest challenge facing corpus developers is that of obtaining the required texts. Where should one get text from? How do copyright issues affect data gathering

and how should the texts be entered, stored, arranged and catalogued once they have been obtained?

At the outset it should be stated that perhaps the best source of texts for corpus usage are other, pre-existing corpora. It was noted earlier in this chapter that careful searching for similar corpora to what you may have in mind is perhaps the best place to start. It is then possible to perhaps collect a certain amount of data yourself and then supplement it with relevant sections of existing corpora. Once this avenue has been explored, it is possible to see that data can be obtained from two basic sources: publicly available texts and privately available sources.

Publicly available data

Publicly available data can be gathered from a variety of sources – newspapers, journals, magazines and a number of sites on the Internet. In a recent study, Ekbal and Bandyopadhyay (2008) describe the creation of their web-based Bengali news corpus. The data were gathered from the web archive of a popular Bengali newspaper by means of a web crawler that was able to automatically retrieve text in HTML format. The files were then cleaned of HTML formatting to leave just that text related to news items. Modern text analysis programs such as *WordSmith* version 5 (Scott 2007) allow automated web searching and download through its webfile downloader function. (An interesting problem occurred with the NCI project in that on downloading text from newswire services, it was found to contain duplicate text: different sources producing the same or very similar text. This had to be manually eliminated.) In all of the instances, the problem of copyright rears its head. Meyer (2002) notes that US copyright law states that while it is acceptable to copy texts for private/research usage, it is not allowed to put that text into electronic format and distribute it as part of a corpus. Further, laws vary from country to country, so the corpus compiler needs to be aware not only of the laws in their own country, but also of the laws in the country from where the text is being taken. One method to avoid problems with copyright is to use texts if appropriate from one of the many open source text archives available on the internet, such as Project Gutenberg. Using Google to search for 'text archives' results in a number of very useful sites (see especially www.copyright.gov/title17/ for a collection of open source text sites).

Private data

'Private' data here refers to data that are not in the public domain. The best advice that can be given here is that, depending on the type of corpus you are creating, you should use any personal contacts you may have, however tentative, to gain access to the documents you require. If one is personally known to the subjects beforehand, the chances of actually getting data are greatly increased. However, even with people one knows already, it is not always easy to persuade them to help. There has to be a degree of polite ruthlessness, as, depending on their position within a given company or institution, it is sometimes easier for them to refuse to help than to assist in the data gathering process. Thus a certain amount of persistence is needed. A further key issue is that of anonymisation of personal data, for example names and place names. In the BEC, text used a standardised format of replacement such, for example, that any person's name became the word 'personname' and any company became the word 'companyname'.

Data entry

Sinclair (1991: 14) identified three main methods of preparing data for entry into a corpus: *adaptation of data in electronic format, scanning* and *keyboarding*.

Adaptation of material already in electronic form

Text already in electronic format is perhaps the easiest to deal with and to obtain. However, there are some problems arising from the fact that many corpus readers need texts to be in .txt format. Therefore, the original texts need to be stripped of all formatting coding, be it related to word processing (bold, italics) or HTML coding. Once again, corpus programs such as *WordSmith* have text conversion components that can help in this matter.

The simplicity of using electronic text can be seen in the creation of the Medical Anatomy Corpus (MAC) (Nelson 2008). After applying for permission, the whole of *Gray's Anatomy* was downloaded by simply copying and pasting into Word and then storing the files as fourteen .txt files. Each file represented one area of the body, e.g. osteology, veins, embryology. The whole process took two hours. *WordSmith* was then used for both lexical analysis and the development of pedagogical materials (Nelson 2009).

Conversion by optical scanning

For this, two essential items of electronic equipment are needed: a good scanner and efficient OCR (Optical Character Recognition) software. Since the 2000 project of the BEC and PMC, advances have been made in scanner software and there is today a wide variety of very good software available. It is possible to find good reviews of this software on the web and, for example, PCMag.com (see website) regularly reviews software, giving a clear overview of what is available. Wikipedia also has a chart comparing OCR programs produced by the leading companies. You can also Google the phrase 'OCR software review' and retrieve free trial versions of different programs. It should be noted that while scanning often gives very good results, if the quality of the original text has degenerated in any way – for example, if it is a photocopy – then accuracy can go down to 40 or 50 per cent. Thus, every text has to be very carefully manually processed to make sure that the computer text matches the original. This obviously is very time-consuming; for example, 373,011 words were scanned into the BEC over a period of eighteen months and each page required multiple corrections. No matter what the standard of the OCR, this element of manual checking cannot be excluded.

Conversion by keyboarding

When all else fails, the only available option is to enter the text by use of keyboarding. This can be seen as the most time-consuming of all methods. This is necessary, for example, when the original text is in such a degraded condition that it will not scan correctly. It must also be used when original documents are in hand-written format as many scanners/software are not able to work on hand-written text with any degree of accuracy.

Finally, combinations of methods can be used. In the New Corpus for Ireland (NCI), for example, Kilgariff *et al.* (2006) report that a combination of using existing corpora,

contacting publishers and newspapers for permission and collecting data from the web was used.

Confidentiality and ethics

When contacting potential sources of texts, it is essential to ensure both that the data you collect is treated according to the laws of copyright and also that you observe the privacy of the authors, if the texts come from the private domain. It is often both sensible or legally required that you draw up a contract on the usage of the data that you receive from respondents. Once all the data have been gathered, the next step is to store them and make them easily available for retrieval.

Data storage and retrieval

Storage and easy retrieval of data is of central importance in the creation of any corpus that will be used by more than one researcher. The BNC used SGML tagging to provide data on, for example, author's name and recording location. In the BEC a database was set up to allow retrieval of data from the 1,102 texts that form the BEC according to the following criteria: (1) file name; (2) URL (where applicable); (3) text topic; (4) text title; (5) text source; (6) text length; (7) text nationality; (8) gender; (9) text type; (10) date of text origin; and (11) corpus sector. In this way it is possible to search for text according a variety of search criteria.

Annotation

The issue of annotation of corpus data will be dealt with in detail elsewhere in this book, but it is worth noting that as a rule at least one version of your corpus should follow the 'clean-text policy' of Sinclair (1991) who proposed that 'The safest policy is to keep the text as it is, unprocessed and clean of other codes' (Sinclair 1991: 21). Sinclair's reasons for this were two-fold: different researchers impose different priorities in corpus data, and a lack of standardisation in analytical measures would create problems for later research of a different nature. Similarly, there is a lack of agreement on basic linguistic features such as words and morphological division. For the BEC, two versions of the corpus were created. First came a 'clean-text' version, where the corpus consists purely of the texts themselves with no annotation at all. A second version of the corpus of the BEC was then created which was Part-of-Speech (POS) tagged using an automatic tagger – *Autasys* (Fang 1998) – which assigned a grammatical tag to each word. The *LOB* tag-set was used for POS assignation.

5. Concluding comments on written corpora

This chapter has attempted to elucidate the issues that are involved when building a written corpus. Despite the fact that written corpora are purportedly easier to create than spoken, largely because of the problems of spoken language transcription, there are still a wide range of issues that need to be addressed at all stages of the process from planning to data gathering and organisation. In building a written corpus, especially if working alone, one has to some extent to balance academic integrity with practical realities, accuracy

with expediency and size with efficiency. Of the three corpora I have created, one took two years to build (BEC), one took eighteen months (PMC) and one took an afternoon (Medical Anatomy Corpus of 556,000 words). Thus, although it can initially seem a daunting task, careful choice of target linguistic features can facilitate valid research that is not necessarily overwhelming for the lone researcher.

Further reading

BNC User Manual, available online at www.natcorp.ox.ac.uk/docs/userManual/design.xml.ID=wrides (The best way to learn how to do something is to see how others have done it well, in order not to have to re-invent the wheel. You can just focus on the written section, but also get insight into the creation of spoken elements at the same time if you wish.)

Douglas, F. (2003) 'The Scottish Corpus of Texts and Speech: Problems of Corpus Design', *Literary and Linguistic Computing* 18(1): 23–37. (This article focuses not just on size and balance issues, but on establishing norms for good practice in corpus creation. Again, reading how things have been done in practice can give invaluable insight for one's own work.)

Hundt, M., Nesselhauf, N. and Biewer, C. (eds) (2007) *Corpus Linguistics and the Web*. Amsterdam and New York: Rodopi. (This provides a thorough overview of exploiting the web to create corpora and using the web as a corpus itself.)

References

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

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Burnard, L. (2001) 'Where Did We Go Wrong? A Retrospective Look at the Design of the BNC', available at http://users.ox.ac.uk/~lou/wip/silfitalk.html (accessed 28 March 2008).

Clear, J. (1992) 'Corpus Sampling', in G. Leitner (ed.) *New Directions in English Language Corpora*. Berlin and New York: Mouton de Gruyter, pp. 21–31.

de Beaugrande, D. (2001) 'Large Corpora, Small Corpora, and the Learning of Language', in M. Ghadessy (ed.) *Small Corpus Studies and ELT. Theory and Practice*. Philadelphia, PA: John Benjamins, pp. 3–28.

Douglas, F. (2003) 'The Scottish Corpus of Texts and Speech: Problems of Corpus Design', *Literary and Linguistic Computing* 18(1): 23–37.

Ekbal, A. and Bandyopadhyay, S. (2008) 'A Web-based Bengali News Corpus for Named Entity Recognition', *Language Resources and Evaluation* 42(2): 173–82.

Fang, A. (1993) 'Building a Corpus of the English of Computer Science', in J. Aarts, P. de Haan and N. Oostdijk (eds) *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam and Atlanta, GA: Rodopi, pp. 73–8.

——(1998) Autasys Version 1. Tagging Program on view at www.phon.ucl.ac.uk/home/alex/home.htm

Flowerdew, L. (1998) 'Corpus Linguistic Techniques Applied to Textlinguistics', *System* 26: 541–52.

Halliday, M. A. K. and Sinclair, J. (1966) 'Lexis as a Linguistic Level', in C. E. Bazell, J. C. Catford, M.A. K. Halliday and R. H. Robins (eds) *In Memory of J. R. Firth*. London: Longman, pp. 148–62.

Howarth, P. (1998) 'Phraseology and Second Language Proficiency', *Applied Linguistics* 19(1): 24–44.

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

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. Harlow: Addison Wesley Longman.

Kilgariff, A., Rundell, M. and Uí Dhonnchadha, E. (2006) 'Efficient Corpus Development for Lexicography: Building the New Corpus for Ireland', *Language Resources and Evaluation* 40: 127–52.

Leech, G. (1991) 'The State-of-the-Art in Corpus Linguistics', in K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics, Studies in Honour of Jan Svartvik*. London and New York: Longman, pp. 8–29.

Ma, K. C. (1993) 'Small-corpora Concordancing in ESL Teaching and Learning', *Hong Kong Papers in Linguistics and Language Teaching* 16: 11–30.

Meyer, C. (2002) *English Corpus Linguistics*. Cambridge: Cambridge University Press

Nelson, M. (2000) 'A Corpus-based Study of the Lexis of Business English and Business English Teaching Materials', unpublished thesis. University of Manchester, available at http://users.utu.fi/micnel/thesis.html

——(2008) 'The Medical Anatomy Corpus' (unpublished).

——(2009) *Using Key Words in Corpus-based Teaching and Research in Perspectives on Language Learning in Practice*. University of Turku Language centre 30-years Celebration. Turku: University of Turku.

Oostdijk, N. (1991) *Corpus Linguistics and the Automatic Analysis of English*. Amsterdam and Atlanta, GA: Rodopi.

Renouf, A. (1987) 'Corpus Development', in J. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Birmingham: HarperCollins, pp. 1–41.

Scott, M. (2007) *WordSmith Tools 5*. Oxford: Oxford University Press.

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

——(2002) 'Introduction', in M. Ghadessy (ed.) *Small Corpus Studies and ELT. Theory and Practice*. Philadelphia, PA: John Benjamins, pp. xvii–xxiii.

Stubbs, M. (1996) *Text and Corpus Analysis*. Oxford: Blackwell.

Svartvik, J. (1992) 'Corpus Linguistics Comes of Age', in J. Svartvik (ed.) *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991*. Berlin and New York: Mouton de Gruyter, pp. 7–17.

Tribble, C. (1997) E-mail to Corpora Discussion Group, corpora@huib.no

——(1998) 'Genres, Keywords, Teaching: Towards a Pedagogic Account of the Language of Project Proposals', in L. Burnard (ed.) *Teaching and Language Corpora 98 – Proceedings of the 1998 TALC Conference*. Oxford: Oxford University Press, pp. 188–98.