2 First capture your data

1. THE BASIC PROBLEM: WHAT'S ALREADY AVAILABLE?

When you consider using a computer to carry out language research, you are likely to become aware very quickly that the availability of data for manual research is very different from the availability of data for computer-based research. This problem was mentioned in section 3.1.1 of Chapter 1, and because it can have a fundamental effect on project viability it is worth considering it now in some detail.

With some important exceptions, most types of language are available in much greater quantity and variety in manually readable forms than in computer-readable forms. The main exceptions, rapidly becoming more significant, are the texts which have been generated by computer, such as e-mail and its analogues and word-processed documents. This second group includes most recent printed publications, although access to them may still be problematic for the general researcher because of copyright and confidentiality problems, which are discussed in more detail in section 3.3 below. Manually readable data of the same type may suffer from corresponding restrictions, but determined researchers can generally get satisfactory access to the contents of even the rarest of books and the most precious of manuscripts. Computer-readable versions of these and other older documents may simply not exist and would therefore need to be created before the project can begin.

As has already been pointed out in Chapter 1, the availability of data can affect the scope of the research, and the extra problems involved in obtaining computer-readable data may restrict your choices to an unacceptable extent. This problem needs to be considered in the initial analysis of the project as part of the benefits and drawbacks of the use of the computer in the research. This chapter describes the points to be considered in selecting data for examination and the main methods of acquiring it and making it suitable for use in the research work.

The collection of computer-readable language that you assemble for your

project, selected on the basis of your research criteria, is usually referred to as a **corpus** to distinguish it from the more random collections of texts held in text archives. The next section describes the main factors that you will need to consider in designing your own research corpus.

2. WHAT DO YOU WANT TO LOOK AT? THE PROBLEMS OF CORPUS DESIGN

Whatever your research objectives are you need to ensure that your corpus will be capable of fulfilling them. You may decide to use an existing corpus, in which case some of the points raised in this section may be less relevant, but they will still be worth considering to ensure that you select your data properly.

2.1. Contents: the corpus as a sample

Before you can consider the availability of data for your research project you will need to decide exactly what you need to examine. This is not as simple or straightforward a question as it may seem to be. In most cases, if not all, the corpus is a sample of a larger collection of language, and is intended to allow conclusions to be drawn about this larger body rather than about itself. This is often true even in cases where the corpus contains everything that is available from within a specific manifestation of language.

As a typical example, consider the Chaucer project described in Chapter 1. The corpus consisted of the entire text of Chaucer's *Canterbury Tales*, but this fact does not prevent it from being merely a sample. Even if the point of the research had been only to identify characteristics of spelling variation within this work, the existence of many other manuscript versions of part or all of the text prevents it from being an exhaustive collection of the entire body of language under review. A set of conclusions based only on the Ellesmere manuscript would naturally provide some evidence of the variations found within that manuscript, but the *Canterbury Tales* manuscript presumably forms only a part of the output of the scribe or scribes responsible, and so again could only be considered as a sample of the language being examined. The fact that other manuscripts produced by the same scribe or scribes may not be identifiable with the necessary degree of accuracy, or may not even have survived, does not affect the conceptual basis of the research: it simply makes the accumulation of a properly representative sample more difficult.

The question of representativeness underlies the whole question of corpus design. If useful conclusions are to be drawn from a sample, the sample must have similar characteristics to the population from which it is drawn. In many cases, these characteristics are themselves the subject of the investigation, and cannot be known in advance, so that some other basis for selection needs to be found. The results of the investigation may be the only opportunity for testing the selection method which has been adopted, and it may be that you will need to carry out a pilot study of an exploratory corpus to refine the basis

of selection for the final corpus which will be used in the project itself. The details of this process will vary enormously from one project to another, and will need to be considered thoroughly before work begins.

2.2. Size: how much is enough?

Once the contents of your ideal corpus have been specified, you will need to decide how much of it you need. The representativeness of the corpus as a sample of the area of language under investigation also depends on its size in relation to the needs of the research project. The most common features of the language will be well represented even in relatively small quantities of text, and if these are the main subject of the work you may only need a relatively small corpus. Unfortunately, it is generally difficult, if not actually impossible, to know in advance what size of corpus will meet the requirements of any given research project. The most effective way of fulfilling the investigative needs of a particular project may well be, again, to carry out a pilot survey based on a relatively small corpus and to extend it if necessary.

3. WHERE DID YOU GET THOSE TEXTS?

There are several ways of acquiring computer-readable texts. In some cases the text may already be available in computer-readable form, and can be bought from commercial or academic organisations, or may even be available free of charge. Some extra work may be involved in preparing the data to make it compatible with the computer hardware and software being used, and this is considered in more detail in section 4 below. In other cases, the text may only be available in printed or manuscript form, and will need to be converted to make it computer-readable before it can be used.

3.1. Texts already available elsewhere

Three main developments have greatly increased the availability of computer-readable text over the last few years:

- the expansion of the large institutional text archive sites
- the expansion of the world-wide computer network system, often called the **internet**
- the increasing importance of the domestic CD-ROM market
- the use of word processing for the origination of new published material.

The detailed implications of these developments for computer-based language research are dealt with below.

3.1.1. Text archives

Collections of computer-readable texts have been accumulated at large academic sites for a number of years. Both the number of text archive sites and

the range of texts held are increasing at a significant rate. The major sites from which English language texts are currently available include the Oxford Text Archive and Project Gutenberg, both of which have large numbers of texts freely available over the internet. It is fairly easy, using the internet directory and search facilities, to find the full range of sites currently available and the texts held at them.

Obviously, if you decide to build your corpus entirely from the holdings of text archives, its composition will be restricted to the range of texts held at them, and this could prove to be a very serious limitation. There seems to be a preponderance of literary texts in the major archives, often skewed towards particular periods. In the English language section of the Oxford Text Archive, for example, there is a relatively large number of Old English and Medieval texts, and a similar abundance of nineteenth-century literature. This could be very useful for the historical researcher engaged with these periods, but rather less so for general modern language work.

3.1.2. The internet

A system of connection between major international academic and commercial computer sites has existed for many years and has allowed the exchange of information through e-mail and other forms of file transfer. This international networking system is now generally known as the internet (though other, more romantic metaphors are also used, 'information superhighway' seeming particularly attractive to politicians and journalists). During the last few years this system has expanded enormously and has become both more generally available and much easier to use. Access to the internet is available from almost any computer. The only additional hardware required is a communications accessory known as a **modem**, which allows connection to telephone lines, together with appropriate communications software.

If you have access to an academic computer system there will almost certainly be an internet connection already available, and you should get local advice on the best way of using it. If not, you will need to explore the hardware and software options to find the best possible way of getting access, and then take account of the costs of the chosen solution in appraising the project.

One of the main advantages of access to the internet is the fact that many of the text archives described above are connected to it, and at least some of the texts held in them may be available through it. In some cases they can be transferred directly over the internet without the need for prior arrangement or payment of fees, and this can provide a very useful basis for the construction of corpora from the wide selection of texts held in the archives, often with minimal effort and at little or no initial cost. There are usually some conditions affecting the use of texts obtained in this way, but these tend to relate to commercial exploitation and would rarely affect their use for normal academic research purposes.

Facilities for file transfer form an integral part of the internet connection system, and once an appropriate site has been located, the required texts can usually be copied to your own machine using standard file transfer software. The exact nature of the software in use varies greatly from one installation to another, but the international computer-using community is eagerly pursuing the twin goals of standardisation and simplification, so that at most academic sites' files can be transferred using a program which may actually be easier to use than your normal word processor.

The text of *Canterbury Tales* which was used for the project described in Chapter 1 was obtained from the Oxford Text Archive before file transfer through the internet became available. It was supplied on a tape which was readable only by a large mainframe computer. The data on the tape required the extensive involvement of specialist computer operating staff, and significant manipulation using text editors on the mainframe, before it could be transferred to a storage medium which was accessible to the machine used for the project. The tape, of course, could only be sent physically through the postal system, so that the text took some days to reach me and then needed several more days work before it was usable. Texts obtained recently from the same site have been transferred over the internet in a matter of minutes, using a few very simple instructions which also allowed interactive browsing through the site catalogue.

In some cases the texts held at a particular site may not be available for transfer to other machines, but may be directly accessible for research purposes from other sites. In these cases, generally involving public access to established large scale corpora, only the software available at the host site can be used to access and analyse the texts, and restrictions may be placed on the nature of the work that can be carried out. The disadvantages inherent in these restrictions may well be outweighed by the advantages of access to such large scale reference corpora. COBUILD's Bank of English is currently available on this basis. In some cases corpora may also be available on CD-ROM, as described in the next section, and you may be able to analyse them using your own software.

In addition to texts, a wide range of software is available from some internet sites, which act as national or international software archives. The material available relates to the whole range of computing needs, and some of it is potentially useful for language analysis purposes. The conditions attached to the use of the software varies significantly from one item to another. In many cases it is freely available as public domain software and can be used without payment. In other cases the software is deposited in the archive as **shareware**, freely available for you to test to check whether it is appropriate for your needs, but subject to a registration fee if you wish to use it in your work. The conditions of use will be made clear in the documentation accompanying the software.

The standard of both the software available at these sites and its documentation vary enormously, and you should take great care in assessing its

suitability for use in your research before relying on its operation. It is also important to ensure that it is free of **viruses** or other malignant interference before installing it on your system, and you may need to obtain expert local help. Despite the potential problems, the software available at these sites can provide a very useful source of analysis tools, which may be capable of fulfilling at least part of your research requirements. You can use the usual internet directory facilities to find details of the archives and their currently available range of software.

3.1.3. CD-ROM

The availability of the CD-ROM as an extremely inexpensive mass-storage medium has led to its adoption for storage-hungry multimedia applications, such as complex computer games or editions of encyclopaedias and other reference works. These typically include visual images, short video clips and sound, often alongside large quantities of text. When the storage capacity of the CD-ROM, usually around 650 megabytes, is used exclusively for text, it can accommodate enormous quantities of it. The increased market penetration of the CD-ROM drive in the domestic market, initially driven by the desire for more complex computer games, has accelerated the development of electronic publishing, and this has led to the release of many text collections on CD-ROM. Some of them could be useful for your language research project.

The range of text available in this form suffers from problems similar to those described in section 3.1.1 in relation to the text archives. Many of the texts found in the archives turn up in these collections, often in the same editions, simply because they have been compiled from the same original public domain sources. This is changing, however, and publishers are becoming more aware of the potential of electronic publishing and of the market for computer-readable texts. In particular, publishers of periodicals have begun to release their text on CD-ROM at regular intervals, so that the complete texts of newspapers and magazines can be easily purchased and used for analysis. Major literary collections like the Chadwyck-Healey English Poetry database are also being developed, and are often available through academic libraries.

At the more specialised end of the market, several complete or part corpora are also available in this form. These include:

- the ICAME Collection of English Language Corpora, including the Brown, Lancaster-Oslo/Bergen (LOB), Kolhapur and London-Lund corpora and the diachronic part of the Helsinki corpus, in versions for the Apple Macintosh, MS-DOS and Unix
- the COBUILD word-bank
- the British National Corpus.

The main disadvantage of texts made available on CD-ROM is likely to be the

physical format in which they are stored. Despite the relatively large storage capacity of the individual CD-ROM, many text collections are so enormous that they still need significant data compression to allow them to be fitted into the available space. The software provided with the CD-ROM allows the text to be decompressed, but it may not be possible to transfer readable text files to another storage medium to allow analysis using other, more standard software. Even where it is physically possible to do this, the licensing agreement may prohibit it to protect intellectual property rights. Any such restrictions will need to be considered very carefully when assessing the project.

3.1.4. Word-processed texts

Because most publications are now prepared using word processing or desktop publishing software a computer-readable version of the text automatically exists. In some cases you may be able to get access to it, although publishers will almost certainly wish to impose restrictions on your use of the text and may well charge for it. The best approach is probably direct application to the publisher explaining the nature of the research, and emphasising the safeguards that would be applied to protect copyright. From then on the negotiations will probably depend on the nature of the research, any benefit (such as good publicity) which would accrue to the publishers, and any potential for commercial exploitation of the data which will be obtained form the text.

In some cases the copyright owner will not be a commercial publisher. If, for example, an investigation is to be carried out into academic writing, large quantities of suitable word-processed example texts are likely to be available from within academic institutions, and a simple request for donations of copies of these text is likely to be reasonably productive. Even in these circumstances the originators of the texts are likely to seek reassurance on the protection of their rights over the contents and the uses to which the data will be put.

The sources of word-processed texts are becoming more varied as the use of the computer for text initiation increases. Business communications, e-mail messages and the writing of children from the earliest stages of education are all likely to be available in fairly large quantities. They each present slightly different problems for the researcher in terms of methods of collection and the formal safeguards that will be demanded, but as with commercial publishers a direct approach to the individuals or organisations involved, explaining the basis of the research, will usually be the best way of dealing with them.

3.2. Making texts computer-readable

If, after searching the sources described in section 3.1, it becomes apparent that the texts needed for your research are not already available, it will be necessary to consider how they are to be converted into a form suitable for use by the computer. The main methods available are described in the following sections, but before deciding which you will use you must consider the position of the

copyright holder of the text, if there is one. The law of intellectual property is complex and varies significantly from one country to another, but the basic consideration is that the permission of the copyright holder will almost certainly be needed before texts can be converted to computer-readable form. Some of the considerations surrounding the legal implications of the use of texts in research projects are dealt with in section 3.3 below.

3.2.1. Scanning

Before you can analyse a text it needs to be in a format in which the computer can recognise it, usually in the form of a standard text file on a storage medium such as a floppy disk or a hard disk. If the text is only currently available in printed or written form this file must be created, since it is not normally convenient, or indeed possible, for the computer to read the data directly from the paper copy. This is not to say that computers are completely incapable of recognising printed text, but the process of scanning printed text, storing it as a visual image and then recognising the visual patterns as a collection of characters is extremely complex, relatively slow and rather prone to error. Because of this text scanning is normally carried out as a separate process of data input, and the results are stored in a file and checked for correctness before any analysis is performed.

The process of scanning and recognising text needs specialised hardware, some form of scanning device, and text recognition software which is compatible with it.

3.2.1(a) Scanning hardware An enormous range of hardware is available, including small hand-held devices, desktop A4-size flatbed scanners and large free-standing units. Hand-held scanners are relatively inexpensive, but can usually only handle relatively narrow blocks of text, often needing two or three passes over the page to scan in a single A4 sheet, and are generally too slow and inaccurate for language research work. At the other end of the market, free-standing units such as the KDEM (Kurzweil Data Entry Machine) are capable of handling large amounts of text rapidly and fairly accurately, but their price and size restricts them to large institutions or major research projects which have significant continuing text scanning requirements. They are often used, for example, in public libraries as part of the automatic reading systems installed for the benefit of visually impaired people. Some large academic institutions may have access to this type of equipment and may provide a scanning service. The middle ground of scanning hardware is occupied by the desktop flatbed scanner, which is relatively easy to use and can handle reasonable amounts of text fairly quickly and accurately.

3.2.1(b) Character recognition software The software needed to convert the scanned image into meaningful text, often referred to as **OCR (optical character recognition)** software, also varies greatly in complexity and ability to deal with different types of text, and one of the more expensive packages will be needed to achieve an acceptable level of accuracy from any text other than perfect quality laser-printed originals, in association with a good quality flatbed scanner.

For a small research project the expenditure that would be involved in obtaining a good quality scanner and adequate recognition software would probably be excessive, especially if there is no likelihood of a continuing need for its use once the project has finished. Luckily, several companies offer a range of scanning services and printed texts can be converted into computer-readable files in a range of formats at relatively low cost.

3.2.1(c) Problems of scanning The main problem associated with the use of scanners is their tendency to error. The following extract shows the uncorrected result of scanning a document. Line numbers have been inserted for ease of reference, and are not present in the scanned text. The original text was a photocopy of a Victorian printed book, and the scanning was carried out on a KDEM:

> That, however, we had some foes, I shall have occasion
> presently to show; but I must rcturn to the scene I was
> describing. I may be pardoned for first giving a slight
> sketch of myself. I hope that I may escape being accused
> of vanity, as I shall not dwell on my personal appearance. 5
> I believe that I inherited some of my parents' good looks;
> but the hardships I have endured have eradicated all traces
> of them. I was well grown for my age (I was barely fifteen),
> but, dressed in my loose shooting costume, my countenance
> ruddy with fresh air and exercise, I looked much 10
> older.
> "What do you suppose would be the lot of a poor man's
> son, if he were to lie discovered acting as you are constantly
> doing in spite of my warnings and commands?" continued
> my father, his voice growing more serious and his look 15
> more grave. "I teil you, boy, that the consequences may
> and will be lamentable; and do not believe, that because
> you are the son of a gentleman, you can escape the punish-
> ment due to the guilty.

Despite the age of the original printed text, and the fact that a photocopy was used to prevent damage to the original, the computer-readable version is fairly accurate. The obvious errors are in lines 2 ('rcturn' instead of 'return') and 16 ('teil' instead of 'tell'), but there is also a less obvious error in line 13. Here the word 'lie' should actually have been 'be', but this is not absolutely obvious even from the surrounding text, and would be much more difficult to detect and correct than the other two errors. However good the scanning process is, some errors will remain, and scanned texts are likely to need signif-

icant proof-reading and manual editing to make them suitable for use in research. Some guidance on automatic and semi-automatic methods of text correction is given in section 4.2 below.

3.2.2. Keyboard entry

Scanning is a useful and fairly efficient method of text conversion where the original text is clearly printed in a form recognisable by the software. Where the format or condition of the text makes it unsuitable for scanning it may need to be keyed directly into the computer. This effectively replaces the scanner and its recognition software with the human keyboard operator. Assuming that the operator has an adequate knowledge of the type of text involved, combined with adequate keyboard skills, the result should be significantly more accurate than input by scanning. On the other hand, the process also tends to be significantly more time-consuming, and unless you have the time to carry out the data entry yourself the costs of keyboarding by commercial operators can be very high indeed.

In some cases, however, there will be no other options. OCR software can cope reasonably well with good quality straightforward printed originals, but if there is any significant degree of complexity in the text format, or if the original is handwritten rather than printed, it is unlikely that scanning will work adequately, or in some cases at all. As an example, a researcher needed to convert the entire text of Johnson's *Dictionary* to computer-readable form to enable the preparation of an electronic edition. The *Dictionary* exists in facsimile editions which photographically reproduce the original eighteenth-century printing. Because of the high cost of direct keyboard entry scanning was attempted, using a sophisticated KDEM unit which is capable of being trained to recognise a wide range of different typefaces. The accuracy rate of the scanning process was rather less than 50 per cent, and the results were completely unacceptable. In the end, despite the high cost, the entire text of the first and fourth editions of the *Dictionary* was entered by keyboard operators employed by a commercial data entry firm.

Part of the reason for the high cost of keyboarding lies in the method normally adopted to increase accuracy. To detect and correct typographic errors caused by miskeying the entire set of data is verified by a complete second keying process. During the second data entry stage the computer compares the second set of keystrokes with the data entered during the first run through, and requests clarification of any discrepancies.

Apart from greatly increased accuracy, the main advantage of direct keyboard entry is the facility that it provides for adding other information during the input process. This is discussed in section 4.3 below.

3.2.3. Spoken language

The discussion of text scanning and keyboard entry above have assumed that

the original form of the language involved in your research is printed or written. Spoken language raises an extra set of complications in the shape of the steps needed to convert information which exists only as sound into written form before it can be entered into the computer. Obviously, if your starting point is a transcription of spoken language it is already in written form and can be dealt with as described above. Otherwise, you will need to adopt and apply a transcription convention of some sort to produce the written form which the computer will ultimately analyse.

It would, of course, be much more convenient if the computer could be fed the sound of the spoken language and either analyse it directly or automatically convert it to its written equivalent. This would involve the development of software which would be the audio equivalent of OCR, perhaps 'audio character recognition'. Although some progress has been made in the area of speech recognition (as outlined in Chapter 7) there is not, as yet, an effective method of interpretation which would allow sound to be processed as language.

3.3. Which texts can you use?

All of the methods described above for converting text to computer-readable form assume that you are legally entitled to use the texts you have chosen for your research. With manual research this is rarely a significant problem, since the examination of a text and reasonable quotations from it in an academic publication are covered by well-established and generally recognised conventions. The main problem introduced by the use of the computer is the need to store the text in electronic form. This is often specifically precluded in publishers' copyright notices within publications, and the need to do it may bring you up against the complex laws, varying between different countries, which regulate the concept of intellectual property. There is, however, a basic approach which may help you to avoid most of this complexity.

The first point that you need to establish relates to the precise details of the ownership of any copyright which exists in the text. If the particular edition of a text that you wish to use was published a sufficiently long time ago copyright may well have lapsed. The exact length of time involved varies from one country to another and will need to be checked. It is important to remember that a new edition, even of an old work, may reinstate some form of copyright, and that this also applies to the production of an electronic edition.

Once the copyright owner has been identified, you should write and ask for their formal permission to use the text for your research. You should give them full details of the nature of your research, the medium on which their text will be stored and any form of publication that will be produced from it, and of the acknowledgement that you intend to make if they give you permission. In the case of texts obtained from text archives, the question of copyright and permitted uses will normally have been established when the text was

deposited, and there will generally be a set of standard conditions which you agree to when getting access to the text.

4. WHAT ELSE NEEDS TO BE DONE?

Even when the texts that you need are available or have been specially entered using the most appropriate method, they may still need some adjustment. The physical or logical format of the data may be incompatible with the hardware or software that you intend to use, or there may be an unacceptable level of errors in the text. It is also possible that you may want to add other information to an otherwise perfectly suitable set of texts. This section describes the most likely problem areas and the most useful ways of overcoming them.

4.1. File formats

Computer-readable texts available by transfer from other internet sites, on CD-ROM or in other forms, may still not be usable on the hardware and with the software that you have available. The reasons for this normally relate to the physical format of the text, in other words the medium on which it is stored, or its logical format. Unless both are compatible with your own set-up you will not be able to use the data without carrying out appropriate conversion processes. There are too many possible combinations of formats to deal with the individual conversion methods, but the general principles described in the following sections should cover most eventualities.

4.1.1. Physical formats

Since computers began to be used widely for commercial and academic purposes several forms of physical storage media have been developed. In the earlier days of computing, when the large mainframe computer was the most common type of installation, magnetic tape was the main exchangeable storage medium. With the advent of desktop personal computers the floppy disk largely replaced magnetic tape for everyday storage purposes, but no single standard was adopted for disk size, storage density or data formats. As a result, several different standards still exist. Magnetic tape cartridges or cassettes are still used in some personal computer systems, but mainly for large scale storage, security copies and so on.

Many older computers use 5¼ inch diameter disks, while the modern standard size is 3½ inch. Other sizes have also been used on comparatively recent models, and each different size of disk needs a physically different disk drive. Even within the 3½ inch standard there are two main storage densities, allowing 0.72 or 1.44 megabytes of data to be stored, and a very similar situation exists in 5¼ inch disks. In addition to this, the two main types of personal computer which are currently most widely used – the IBM-compatible PC and the Apple Macintosh – use completely different storage formats on the same physical disk. Disks formatted for use on IBM-compatible machines can only

be used in Apple Macintosh disk drives using special software, and disks formatted for use on the Macintosh range cannot be handled by the disk drives on IBM-compatible PCs. Computers running the Unix operating system often use storage formats which are totally different from either of these, but may be able to access disks written under MS-DOS.

This rather complex situation means that you must make sure that the texts that you intend to use for your research are available on an appropriately sized medium in a format which is accessible by your hardware and operating system. If this is not the case, you will need to convert them to the appropriate format. The main conversion methods are dealt with below. Whichever method seems likely to be most appropriate for your individual data problem, a quick check with a local expert will save a great deal of frustration and wasted time and effort.

4.1.1(a) Dual disk drives If the text is available on a different size of disk to the one used by your own computer, but is in the same data format, you need to gain access to a computer compatible with your own which has disk drives catering for both sizes. Storage densities will also be important: drives capable of handling higher storage densities can normally read and write disks with the same format but with a lower density, but you must ensure that you use the appropriate type of disk. If, for instance, you intend to use an IBM-compatible computer which has a low density 3½ inch disk drive, and the data is currently available on high density 5¼ inch disks produced on another IBM-compatible machine, you will need to use an IBM-compatible computer with both a high density 5¼ inch drive and a 3½ inch drive. Both disk drives on such a machine would probably be capable of handling both high and low density disks, and you would need to ensure that the disk used in the 3½ inch drive was a low density disk and was properly formatted as low density before copying the data to it.

In a case like this, the lower storage capacity of the disks to which the data was being copied might mean that more disks would be needed than the original number. If any individual files were larger than the maximum capacity of the new disks they would need to be divided up, using either a text editor or a file-splitting utility, to allow them to fit.

4.1.1(b) Communication software Where you need to convert the file format to one used by a completely different type of computer, or where a machine with dual disk drives of the appropriate types is not available, the data can be transferred directly from one computer to another. This involves the use of a special lead to connect the two computers, or the connection of both of them to the same network or to an appropriate communications link. This is the technique that would be used to collect text files from an archive site through the internet, but it can equally well be used to convert from one format to another with the two computers side by side on the same desktop. Communications software exists for almost all types of computer, and is

generally sufficiently standardised to allow data to be transferred between them, if it is in a suitable form. The main points to consider in information interchange between computers are outlined in section 4.1.2 below.

4.1.1(c) Special disk-formatting software In some cases special software is available which allows computers to handle disks formatted for other operating systems. For example, as already mentioned in section 4.1.1, some Apple Macintosh machines, running the appropriate software, can read and write disks formatted for the MS-DOS operating system which runs on IBM-compatible PCs, and many computers running Unix have similar software to enable MS-DOS file systems to be accessed. More specialised software also exists to allow formats based on older operating systems to be converted, usually to MS-DOS format, but these are becoming less easy to obtain. If local resources are not adequate for your needs, it is still possible to find specialist data transfer firms who can carry out the conversion for you.

4.1.2. Logical formats

Once the text files are in the correct physical format for use on your computer system, it is important to ensure that they are also in an appropriate logical format. Most of the text exploration software described in the following chapters, and most of the programs that you might write yourself, assume that the text is in the approximately standardised **ASCII** (American Standard Code for Information Interchange) format, sometimes referred to as **ANSI** standard. This is a very old standard, whose origins lie in the earliest days of computing, when the composition of the individual computer character meant that only 128 different characters were available.

There are implications of this for encoding any linguistic characters other than the standard unaccented roman alphabet. If characters outside this set occur in your texts they will need to be encoded using some other convention, often requiring two or more separate characters in the computer-readable version of the text to represent a single character in the original. As an example, the diachronic part of the Helsinki Corpus of English Texts contains texts in Old and Middle English which use additional characters no longer used in English. Because these are not provided for in the ASCII system it uses two- or three-character combinations to represent them. As an example, the character *þ* ('thorn') is represented by *+t*. Any software that you use to process the corpus would need to take special conventions like these into account, and if the software is not sufficiently adjustable, the text file may need to be modified so that it will not distort the results of your analysis.

It is also possible that the texts may not be in ASCII format to begin with. If they have been produced from a word processing or desktop publishing program they are likely to contain extra information relating to text formatting, page-layout information, illustrations contained in the text, and so on. Because of this the text encoding is likely to be very different from the simple

ASCII format, and before the file can be processed by the software you have chosen, it will probably need to be converted. Most modern word processing and desktop publishing software packages contain a facility for file format conversion, and ASCII is usually one of the options available. It is important to realise that the process of conversion will remove all the information relating to different typefaces, font sizes, paragraph indentation and so on. If any of this is important in your work, it may be necessary to find or develop software that can handle the file in its original form.

4.2. Text correction

For various reasons it is likely that your text, however it has been obtained, will contain errors. These could be individual typographical errors, caused by incorrect recognition during scanning, miskeying during text input or in the original creation of a document, or they could involve the omission, duplication or misplacing of sections of the text. It is obviously important to minimise the possibility of error during the input process, but however scrupulously it is carried out some mistakes are almost certain to creep in. Error correction is an important stage in the preparation of your text for use in research, and it can be a time-consuming and expensive process. Because of this, it is important to determine an acceptable level of error before you decide how to approach error correction. Errors which are unlikely to distort your results significantly are probably not worth correcting, and in many cases it will be worth carrying out a preliminary investigation to see how accurate the data needs to be. Once you have decided on the level of error detection and correction that you need, it may be possible to make it more efficient by using the computer to assist the process, or even, in some cases, to carry it out for you.

4.2.1. The spell checker

Perhaps the simplest way of using the computer to correct errors is to run the text file through a spell checker. Most word processing packages now contain this facility, and if the language in the text file conforms to the standard embodied in the spell checker a quick run-through could highlight many of the typographical errors. For obvious reasons, this would not have been a suitable method to use for the text of *Canterbury Tales* which formed the basis of the project described at the beginning of Chapter 1, and this will be true of any project which sets out explore spelling variation in texts. Equally obviously, it will only detect the errors that produce word forms which do not exist in the spell checker's dictionary. For example, if the text contains the word *tool* and it is somehow entered as *tolo* it is likely to be detected; if it is entered as *toll* it will almost certainly not be detected. In many cases, the second error could distort the results more seriously and would certainly be less likely to be picked up at a later stage.

4.2.2. Other computer-assisted methods

The computer can be used to carry out automatic error correction, but only where the error is consistent and capable of accurate identification. In the case of words split by hyphenation, for example, it may be possible to write a simple program which can detect the hyphen symbol at the end of a line and reconnect the remaining letters of the split word at the beginning of the next line. Such a program may need to check that the hyphen does not represent a dash, and will need a means of recognising the end of the word, but neither of these should present major problems. Section 2.1 of Chapter 8 provides a more detailed description of a practical example of the problem, and the program written to deal with it is described in Appendix 3.

Similarly, if a word has been systematically wrongly input because of scanning problems or human misunderstanding, all appearances of the incorrect version may be capable of automatic replacement by the correct form. Before this can be done with any confidence, of course, you need to be sure that it genuinely is a consistent error, and that none of the 'errors' could be the correct form of another word. Very often, a standard 'find and replace' utility within a word processor can be used, taking advantage of the opportunity to check the context before carrying out replacement if there is any doubt.

It is much harder to use the computer to detect the omission or duplication of sections of text, although if the number of lines in the original text is available, as it often is in editions of literary texts, a quick line count by the computer and a reconciliation of totals can provide a useful check. If a piece of text has simply been misplaced this check is unlikely to reveal the fact, but this may not distort results significantly so long as the text is complete.

4.2.3. Manual checking

Where the accuracy of the text is sufficiently important, you will almost certainly need to carry out manual checking instead of, or even as a supplement to, the use of the computer. Full proof-reading is probably the only certain way of achieving accuracy, and its usefulness will be entirely dependent on the skill and reliability of the proof-reader. It is also likely to be expensive in terms of time, money or both. In some cases, especially where the checking also needs to cover the addition of other information such as mark-up codes, significant judgement will be needed to determine the accuracy of the text, and the proof-reader will need to understand the project thoroughly and be capable of editorial decisions.

4.3. Mark-up codes

Apart from the need to correct errors, texts may need other information added to them before they are ready for analysis. It may be necessary, for example, to label the different structural levels of a text to ensure that particular forms of analysis are only carried out within an appropriate level. A fairly

complex example of this can be seen in the computer-readable text of Johnson's *Dictionary*, mentioned earlier in section 3.2.2. During the process of keying in the entire text codes were added to it to identify the nature of the text within each dictionary entry. An example, taken from the early stages of the project, is given below for the word *abandon*:

@1 To <<a>>ABA<<'>>NDON.
@2 <cf2>v.a.<cf1>
@3 [Fr. <cf2>abandonner.<cf1> Derived, according to <cf2>Menage, <cf1> from the Italian <cf2>abandonare,<cf1> which signifies to forsake his colours; <cf2>bandum [vexillum] deserere. Pasquier<cf1> thinks it a coalition of <cf2>a ban donner,<cf1> to give up to a proscription; in which sense we, at this day, mention the ban of the empire. Ban, in our own old dialect, signifies a curse; and to <cf2>abandon,<cf1> if considered as compounded between French and Saxon, is exactly equivalent to <cf2>diris devovere.<cf1>]
@4 1. To give up, resign, or quit;
@9 often followed by the particle <cf2>to.<cf1>
@5 The passive gods behold the Greeks defile
Their temples, and <cf2>abandon to<cf1> the spoil
Their own abodes; we, feeble few, conspire
To save a sinking town, involv'd in fire.
@6 <cf2>Dryd.<cf1>
@7 <cf2>\AEneid.<cf1>
@4 2. To desert.
@5 The princes using the passions of fearing evil, and desiring to escape, only to serve the rule of virtue, not to <cf2>abandon<cf1> one's self, leapt to a rib of the ship.
@6 <cf2>Sidney,<cf1>
@8 <cf2>b. ii.<cf1>
@5 Then being alone,
Left and <cf2>abandon'd<cf1> of his velvet friends,
'Tis right, quoth he; thus misery doth part
The flux of company.
@6 <cf2>Shakesp.<cf1>
@7 <cf2>As you like it.<cf1>
@5 What fate a wretched fugitive attends,
Scorn'd by my foes, <cf2>abandon'd<cf1> by my friends.
@6 <cf2>Dryd.<cf1>
@7 <cf2>\AEn.<cf1>
@8 <cf2>2.<cf1>
@4 3. To forsake,
@9 generally with a tendency to an ill sense.

@5 When he in presence came, to Guyon first
He boldly spake, Sir knight, if knight thou be,
<cf2>Abandon<cf1> this forestalled place at erst,
For fear of further harm, I counsel thee.
@6 <cf2>Spenser's<cf1>
@7 <cf2>Fairy Queen,<cf1>
@8 <cf2>b. ii. cant. 4. stanz. 39.<cf1>
@5 But to the parting goddess thus she pray'd;
Propitious still be present to my aid,
Nor quite <cf2>abandon<cf1> your once favour'd maid.
@6 <cf2>Dryd.<cf1>
@7 <cf2>Fab.<cf1>

In the sample, each element of the original text is set on a separate line and preceded by a code consisting of '@' followed by a number. Each of these codes represents a different type of information contained within the dictionary entry. For example, '@1' signals the headword label, '@2' the grammatical information supplied by Johnson for this headword, '@3' the etymological information, '@4' the individual senses of the headword and so on. This is a very specific coding system developed for a very specific set of needs, but the general needs of those involved in text preparation, storage and investigation have led to the formation of a body to promote the development of a standard system, the Text Encoding Initiative (TEI), which has adopted as its starting-point a mark-up system called SGML (Standard Generalised Mark-up Language). SGML is now widely used in the preparation of texts for deposition in text archives.

The main problem involved in the addition of mark-up codes is the level of judgement that is needed, not only in the design or adoption of a specific coding system for the text, but also in its detailed application. If the codes are to be entered by the keyboard operators while they are keying in the data from the text, which may be the most efficient approach, they will need thorough training before data entry begins, and a high level of supervision during the input process. Even if this is provided, it is likely that the results will need significant manual checking before they can be used in a research project.

5. EXERCISES

1. What types of text would you include in a corpus intended to be representative of the written language read by British schoolchildren?
2. How would you set about designing and gathering a computer-readable corpus of casual conversational English?
3. You want to create a corpus from medieval texts which are not yet available in computer-readable form. Describe the advantages and disadvantages of the various possible methods of entering the data.

FURTHER READING

Kytö, M., Ihalainen, O. and Rissanen, M. (eds) (1988) *Corpus Linguistics, Hard and Soft*, Amsterdam: Rodopi. The following papers are especially relevant: Kytö, M. and Rissanen, M. 'The Helsinki Corpus of English Texts: classifying and coding the diachronic part', pp. 169–79; Renouf, A. 'Coding metalanguage: issues raised in the creation and processing of specialised corpora', pp. 197–206.

Leech, G., Myers, G. and Thomas, J. (eds) (1995) *Spoken English on Computer*, London: Longman.

Leitner, G. (ed.) (1992) *New directions in Corpus Linguistics*, Berlin: Mouton de Gruyter. Part I, 'Corpus design and text encoding' (pp. 73–107) is especially relevant.

Renouf, A. (1987) 'Corpus development', in J. Sinclair (ed.), *Looking Up*, London: Collins ELT, ch. 1 (pp. 1–40).

Sinclair, J. (1991) *Corpus, Concordance, Collocation*, Oxford University Press, ch. 1 (pp. 13–26).

Svartvik, J. (ed.) (1992) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*, Berlin: Mouton de Gruyter. The 'Corpus design and development' section on pp. 129–209 is especially relevant.

Text archives and other sources of corpus data

Internet addresses, site holdings and World-Wide-Web pages vary so rapidly that any listings would be out of date long before publication. The best strategy, once you have connected yourself to the internet in the most convenient way, is to use one of the many search or directory programmes available to find the current addresses of archives and other useful sites. You may need to seek advice from local experts on the use of the software available to you.

The electronic edition of Johnson's *Dictionary*, referred to in sections 3.2.2 and 4.3, is now available on CD-ROM. It is edited by Anne McDermott of the University of Birmingham, and published by Cambridge University Press.