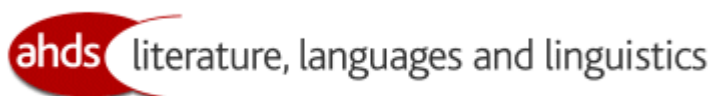




AHDS Guides to Good Practice

Developing Linguistic Corpora: a Guide to Good Practice

Edited by Martin Wynne



Produced by

ISSN 1463 5194

<http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm>
(12.10.2005)

Preface (Martin Wynne)

Chapter 1: Corpus and Text — Basic Principles (John Sinclair, Tuscan Word Centre © John Sinclair 2004)

1. Who builds a corpus?
2. What is a corpus for?
3. How do we sample a language for a corpus?

Acknowledgements

Notes

Chapter 2: Adding Linguistic Annotation (Geoffrey Leech, Lancaster University © Geoffrey Leech 2004)

1. What is corpus annotation?
 2. What different kinds of annotation are there?
 3. Why annotate?
 4. Useful standards for corpus annotation
- Detailed and explicit documentation should be provided
5. The encoding of annotations
 6. Annotation manual
 7. Some 'provisional standards' of best practice for different linguistics levels
 8. Evaluation of annotation: realism, accuracy and consistency
 9. Getting down to the practical task of annotation

Chapter 3: Metadata for corpus work (Lou Burnard, University of Oxford © Lou Burnard 2004)

1. What is metadata and why do you need it?
2. Scope and representation of metadata
3. Editorial metadata
4. Analytic metadata
5. Descriptive metadata
6. Metadata categories for language corpora: a summary
7. Conclusions

Notes

Chapter 4: Character encoding in corpus construction (Anthony McEnery and Richard Xiao, Lancaster University © Anthony McEnery and Richard Xiao 2004)

1. Introduction
2. Shift in: what is character encoding about?
3. Legacy encoding: complementary and competing character codes
4. Globalisation: efforts to unify character codes
5. Unicode Transformation Formats (UTFS)
6. Shift out: conclusions and recommendations

Notes

Chapter 5: Spoken language corpora (Paul Thompson, University of Reading © Paul Thompson 2004)

1. Introduction
2. Data collection
3. Transcription
4. Representation and annotation
5. Access

Chapter 6: Archiving, distribution and preservation (Martin Wynne, University of Oxford © Martin Wynne 2004)

1. Introduction
 2. Planning for the future
- How will users find the corpus?
What file format should my corpus text files be in for archiving?
3. Conclusion

Appendix: How to build a corpus (John Sinclair, Tuscan Word Centre © John Sinclair 2004)

Introduction

The World Wide Web

Perfectionism

Indicative, not definitive

Corpus-building software

Procedure

Notes

Bibliography

Preface (Martin Wynne)

A linguistic corpus is a collection of texts which have been selected and brought together so that language can be studied on the computer. Today, corpus linguistics offers some of the most powerful new procedures for the analysis of language, and the impact of this dynamic and expanding sub-discipline is making itself felt in many areas of language study.

In this volume, a selection of leading experts in various key areas of corpus construction offer advice in a readable and largely non-technical style to help the reader to ensure that their corpus is well designed and fit for the intended purpose.

This Guide is aimed at those who are at some stage of building a linguistic corpus. Little or no knowledge of corpus linguistics or computational procedures is assumed, although it is hoped that more advanced users will also find the guidelines here useful. It also has relevance for those who are not building a corpus, but who need to know something about the issues involved in the design of corpora in order to choose between available resources and to help draw conclusions from their analysis.

Increasing numbers of researchers are seeing the potential benefits of the use of an electronic corpus as a source of empirical language data for their research. Until now, where did they find out about how to build a corpus? There is a great deal of useful information available which covers principles of corpus design and development, but it is dispersed in handbooks, reports, monographs, journal articles and sometimes only in the heads of experienced practitioners. This Guide is an attempt to draw together the experience of corpus builders into a single source, as a starting point for obtaining advice and guidance on good practice in this field. It aims to bring together some key elements of the experience learned, over many decades, by leading practitioners in the field and to make it available to those developing corpora today.

The modest aim of this Guide is to take readers through the basic first steps involved in creating a corpus of language data in electronic form for the purpose of linguistic research. While some technical issues are covered, this Guide does not aim to offer the latest information on digitisation techniques. Rather, the emphasis is on the principles, and readers are invited to refer to other sources, such as the latest AHDS information papers, for the latest advice on technologies. In addition to the first chapter on the principles of corpus design, Professor Sinclair has also provided a more practical guide to building a corpus, which is added as an appendix to the Guide. This should help guide the user through some of the more specific decisions that are likely to be involved in building a corpus.

Alert readers will see that there are areas where the authors are not in accord with each other. It is for the reader to weigh up the advantages of each approach for his own particular

project, and to decide which course to follow. This Guide not aim to synthesize the advice offered by the various practitioners into a single approach to creating corpora. The information on good practice which is sampled here comes from a variety of sources, reflecting different research goals, intellectual traditions and theoretical orientations. The individual authors were asked to state their opinion on what they think is the best way to deal with the relevant aspects of developing a corpus, and neither the authors nor the editor have tried to hide the differences in approaches which inevitably exist. It is anticipated that readers of this document will have differing backgrounds, will have very diverse aims and objectives, will be dealing with a variety of different languages and varieties, and that one single approach would not fit them all.

I would like to thank the authors of this volume for their goodwill and support to this venture, and for their patience through the long period it has taken to bring the Guide to publication. I would like to acknowledge the extremely helpful advice and editorial work from my colleague Ylva Berglund, which has improved many aspects of this guide.

Chapter 1: Corpus and Text — Basic Principles (John Sinclair, Tuscan Word Centre © John Sinclair 2004)

A corpus is a remarkable thing, not so much because it is a collection of language text, but because of the properties that it acquires if it is well-designed and carefully-constructed.

The guiding principles that relate corpus and text are concepts that are not strictly definable, but rely heavily on the good sense and clear thinking of the people involved, and feedback from a consensus of users. However unsteady is the notion of *representativeness*, it is an unavoidable one in corpus design, and others such as *sample* and *balance* need to be faced as well. It is probably time for linguists to be less squeamish about matters which most scientists take completely for granted.

I propose to defer offering a definition of a corpus until after these issues have been aired, so that the definition, when it comes, rests on as stable foundations as possible. For this reason, the definition of a corpus will come at the end of this paper, rather than at the beginning.

1. Who builds a corpus?

Experts in corpus analysis are not necessarily good at building the corpora they analyse — in fact there is a danger of a vicious circle arising if they construct a corpus to reflect what they already know or can guess about its linguistic detail. Ideally a corpus should be designed and built by an expert in the communicative patterns of the communities who use the language that the corpus will mirror. Quite regardless of what is inside the documents and speech events, they should be selected as the sorts of documents that people are writing and reading, and the sorts of conversations they are having. Factual evidence such as audience size or circulation size can refine such sampling. The corpus analyst then accepts whatever is selected.

This could be stated as a principle:

1. The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.

Obviously if it is already known that certain text types contain large numbers of a microlinguistic feature such as proper nouns or passive verb phrases, it becomes a futile activity to "discover" this by assembling a corpus of such texts.

Selection criteria that are derived from an examination of the communicative function of a text are called *external criteria*, and those that reflect details of the language of the text are called *internal criteria*. Corpora should be designed and constructed exclusively on external criteria ([Clear 1992](#))¹.

2. What is a corpus for?

A corpus is made for the study of language; other collections of language are made for other purposes. So a well-designed corpus will reflect this purpose. The contents of the corpus should be chosen to support the purpose, and therefore in some sense represent the language from which they are chosen.

Since electronic corpora became possible, linguists have been overburdened by truisms about the relation between a corpus and a language, arguments which are as irrelevant as they are undeniably correct. Everyone seems to accept that no limits can be placed on a natural language, as to the size of its vocabulary, the range of its meaningful structures, the variety of its realisations and the evolutionary processes within it and outside it that cause it to develop continuously. Therefore no corpus, no matter how large, how carefully designed, can have exactly the same characteristics as the language itself.

Fine. So we sample, like all the other scholars who study unlimitable phenomena. We remain, as they do, aware that the corpus may not capture all the patterns of the language, nor represent them in precisely the correct proportions. In fact there are no such things as "correct proportions" of components of an unlimited population.

2. Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.

However hard we strive, a corpus will occasionally show features which we suspect not to be characteristic of the language under study, or fail to show features which are expected. Following our first principle above, we should not feel under pressure to use the patterns of the language to influence the design of the corpus, but we should review the design criteria to check that they are adequate.

To optimise the application of this principle we can make use of an important resource within ourselves, which is not available to most scientific researchers in other disciplines. As sophisticated users of at least one language, we have an inbuilt awareness of language structure, often called intuition, that gives a personal, independent and non-negotiable assessment of language pattern. Intuition can help in many ways in language research, in conjunction with other criteria of a more examinable nature. The drawbacks to intuition are

(a) that we cannot justify its use beyond personal testimony, and (b) that people differ notoriously in their intuitive judgements. In this context we should also be aware that an incautious use of intuition in the selection of texts for a corpus would undermine the first principle².

3. How do we sample a language for a corpus?

There are three considerations that we must attend to in deciding a sampling policy:

1. The orientation to the language or variety to be sampled.
2. The criteria on which we will choose samples.
3. The nature and dimensions of the samples.

1. Orientation

This is not a crisply delineated topic, and has largely been taken for granted so far in corpus building. The early corpora, for example the Brown corpus and those made on its model ([Hofland and Johansson 1982](#)), were *normative* in their aims, in that their designers wanted to find out about something close to a standard language. The word "standard" appears in the original Brown title; by choosing published work only, they automatically deselected most marked varieties. Most of the large reference corpora of more recent times adopt a similar policy; they are all constructed so that the different components are like facets of a central, unified whole. Such corpora avoid extremes of variation as far as possible, so that most of the examples of usage that can be taken from them can be used as models for other users.

Some corpora have a major variable already as part of the design — a historical corpus, for example, is deliberately constructed to be internally contrastive, not to present a unified picture of the language over time (though that could be an interesting project). Another kind of corpus that incorporates a time dimension is the *monitor* corpus ([Sinclair 1982](#)); a monitor corpus gathers the same kind of language at regular intervals and its software records changes of vocabulary and phraseology. *Parallel* corpora, or any involving more than one language, are of the same kind — with inbuilt contrasting components; so also is the small corpus used in [Biber et. al. \(1999\)](#) to demonstrate varietal differences among four externally-identified varieties of contemporary English. These corpora could be called *contrastive* corpora because the essential motivation for building them is to contrast the principal components.

There is a guiding principle here of great importance, and one which is commonly ignored.

3. Only those components of corpora which have been designed to be independently contrastive should be contrasted.

That is to say, the existence of components differentiated according to the criteria discussed below, or identified by archival information, does not confer representative status on them, and so it is unsafe to use them in contrast with other components. Now that with many corpus management systems it is possible to "dial-a-corpus" to your own requirements, it is important to note that the burden of demonstrating representativeness lies with the user of such selections and not with the original corpus builder. It is perfectly possible, and indeed very likely, that a corpus component can be adequate for representing its variety within a large normative corpus, but inadequate to represent its variety when freestanding.

This point cannot be overstated; a lot of research claims authenticity by using selections from corpora of recognised standing, such as the Helsinki Corpus, which is a notable reference corpus covering the language of almost a millennium in a mere 1,572,820 words. Each small individual component of such a corpus makes its contribution to the whole and its contrasts with other segments, but was never intended to be a freestanding representative of a particular state of the language. See the detailed description at <http://helmer.aksis.uib.no/icame/hc/>. Normative, historical, monitor and varietal corpora are not the only kinds; demographic sampling has been used a little, and there are all sorts of specialised corpora. For an outline typology of corpus and text see [Sinclair \(2003\)](#), which is a summary and an update of a report made for the European Commission (for that report see the EAGLES server at <http://www.ilc.pi.cnr.it>).

2. Criteria

Any selection must be made on some criteria and the first major step in corpus building is the determination of the criteria on which the texts that form the corpus will be selected.

Common criteria include:

1. the mode of the text; whether the language originates in speech or writing, or perhaps nowadays in electronic mode;
2. the type of text; for example if written, whether a book, a journal, a notice or a letter;
3. the domain of the text; for example whether academic or popular;
4. the language or languages or language varieties of the corpus;
5. the location of the texts; for example (the English of) UK or Australia;
6. the date of the texts.

Often some of these large-scale criteria are pre-determined by constraints on the corpus design — for example a corpus called MICASE stands for the Michigan Corpus of Academic Spoken English, and the corpus consists of speech events recorded on the Ann Arbor campus of the University of Michigan on either side of the millennium; it follows that the language in the corpus will mainly be of the large variety called American English. All the

above criteria are pre-determined, and all but the date are built into the name of this corpus, so its own structural criteria will be set at a more detailed level³.

All but the most comprehensive corpora are likely to use one or more criteria which are specific to the kind of language that is being gathered, and it is not possible to anticipate what these are going to be. The corpus designer should choose criteria that are easy to establish, to avoid a lot of labour at the selection stage, and they should be of a fairly simple kind, so that the margin of error is likely to be small. If they are difficult to establish, complex or overlapping they should be rejected, because errors in classification can invalidate even large research projects and important findings.

Now that there are a number of corpora of all kinds available, it is helpful to look at the criteria that have been used, and to evaluate them in three ways — as themselves, how useful and valuable a variety of the language they depict; as a set of criteria, how they interact with each other and avoid ambiguity and overlap; and the results that they give when applied to the corpus.

4. Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.

Beyond these criteria it is possible to envisage an unlimited categorisation of people, places and events, any of which are potentially valuable for one study or another (see the typology mentioned above). The gender of the originator of a text has been a popular criterion in recent years, though few texts have a single originator whose gender is known, and hoaxes are not unknown (for example it was recently alleged that the works of a famous crime writer noted for rough-and-tough stories were in fact composed by his wife). It is essential in practice to distinguish structural criteria from useful information about a text.

For a corpus to be trusted, the structural criteria must be chosen with care, because the concerns of balance and representativeness depend on these choices. Other information about a text can, of course, be stored for future reference, and scholars can make up their own collections of texts to suit the objectives of their study. The question arises as to how and where this information should be stored, and how it should be made available. Because it is quite commonly added to the texts themselves, it is an issue of good practice, especially since in some cases the additions can be much larger than the original texts.

In the early days of archiving text material, the limitations of the computers and their software required a structurally simple model; also before there was an abundance of language in electronic form, and before the internet made it possible for corpora to be accessed remotely, it was necessary to agree protocols and practices so that data could be made available to

the research community. The model that gained widest acceptance was one where additional material was interspersed in the running text, but enclosed in diamond brackets so that it could — at least in theory — be found quickly, and ignored if the text was required without the additions.

Nowadays there is no need to maintain a single data stream; modern computers have no difficulty storing the plain text without any additions, and relating it token by token to any other information set that is available, whether "mark-up", which is information about the provenance, typography and layout of a printed document, or "annotation", which is analytic information usually about the language⁴. It is also possible nowadays to store facsimiles of documents and digitised recordings of speech, and have the computer link these, item by item, to plain text, thus removing even the need to have mark-up at all.

5. Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.

3. Sampling

Looking down from the totality of the corpus, the major criteria will define several *components*, while at the other end are the individual *texts*, which will be such things as written or printed documents, and transcripts of spoken events. *Cells* are the groupings formed from the intersection of criteria.

The first-level components will be small in number, for practical reasons, because if there are too many then either each component will be very small or the corpus will be very large. The simplest classification is binary, so that if a corpus of spoken language is first divided into "private" and "public", then each of these types will have to be represented by a sufficiently large amount of text for its characteristics to become evident. If the next criterion is "three or fewer active participants", as against "more than three active participants", then each of the original categories is divided into two, and the theoretical size of the corpus doubles.

Each criterion divides the corpus into smaller cells; if we assume that the criteria are binary and cross-cutting then (as we have just seen) two criteria divide the corpus into four cells, three into eight, four into sixteen etc. You then have to decide what is the acceptable minimum number of words in a cell; this depends quite a lot on the type of study you are setting out to do, but if it is not substantial then it will not supply enough reliable evidence as part of the overall picture that the corpus gives of the language. This is known as the "scarce data problem". The matter of size is discussed later, and the example in the following paragraph is only illustrative.

If you decide on, say, a million words as the minimum for a cell, then with four criteria you need a corpus with a minimum size of sixteen million words. Each additional binary criterion doubles the minimum size of the corpus, and in addition we find that real life is rarely as tidy as this model suggests; a corpus where the smallest cell contains a million words is likely in practice to have several cells which contain much more. This involves the question of balance, to which we will return. There are also questions of criteria that have more than two options, and of what to do with empty or underfilled cells, all of which complicate the picture.

The matter of balance returns as we approach the smallest item in a corpus, the text. Here arises another issue in sampling that affects, and is affected by, the overall size of the corpus. Language artefacts differ enormously in size, from a few words to millions, and ideally, documents and transcripts of verbal encounters should be included in their entirety. The problem is that long texts in a small corpus could exert an undue influence on the results of queries, and yet it is not good practice to select only part of a complete artefact. However it is an unsafe assumption that any part of a document or conversation is representative of the whole — the result of research for decades of discourse and text analysis make it plain that position in a communicative event affects the local choices.

The best answer to this dilemma is to build a large enough corpus to dilute even the longest texts in it. If this is not practical, and there is a risk that a single long text would have too great an influence on the whole, so recourse has to be made to selecting only a part of it, and this has to be done on "best guess" grounds. But even a very large corpus may find it almost impossible to get round copyright problems if the builders insist on only complete texts. The rights holders of a valuable document may not agree to donate the full text to a corpus, but if it is agreed that occasional passages are omitted, so that the value of the document is seriously diminished, then the rights holders might be persuaded to relent.

These are the issues in text selection, and one point in particular should be made clearly. There is no virtue from a linguistic point of view in selecting samples all of the same size. True, this was the convention in some of the early corpora, and it has been perpetuated in later corpora with a view to simplifying aspects of contrastive research. Apart from this very specialised consideration, it is difficult to justify the continuation of the practice. The integrity and representativeness of complete artefacts is far more important than the difficulty of reconciling texts of different dimensions.

6. Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.

4. Representativeness

It is now possible to approach the notion of representativeness, and to discuss this concept we return to the first principle, and consider the users of the language we wish to represent. What sort of documents do they write and read, and what sort of spoken encounters do they have? How can we allow for the relative popularity of some publications over others, and the difference in attention given to different publications? How do we allow for the unavoidable influence of practicalities such as the relative ease of acquiring public printed language, e-mails and web pages as compared with the labour and expense of recording and transcribing private conversations or acquiring and keying personal handwritten correspondence? How do we identify the instances of language that are influential as models for the population, and therefore might be weighted more heavily than the rest?

The previous paragraph is a set of questions to which there are no definite answers, and yet on which the whole character of the corpus will rest. According to claims, the most likely document that an ordinary English citizen will cast his or her eye over is *The Sun* newspaper; in a corpus of British English should we then include more texts from that paper than from any other source? If this argument is rejected on stylistic grounds — perhaps that the language of *The Sun* is particularly suited to the dramatic presentation of popular news and views and would not be recommended as a general model for written work — then the corpus builder is adopting a prescriptive stance and is risking the vicious circle that could so easily arise, of a corpus constructed in the image of the builder.

The important steps towards achieving as representative a corpus as possible are:

1. decide on the structural criteria that you will use to build the corpus, and apply them to create a framework for the principal corpus components;
2. for each component draw up a comprehensive inventory of text types that are found there, using external criteria only;
3. put the text types in a priority order, taking into account all the factors that you think might increase or decrease the importance of a text type — the kind of factors discussed above;
4. estimate a target size for each text type, relating together (i) the overall target size for the component (ii) the number of text types (iii) the importance of each (iv) the practicality of gathering quantities of it;
5. as the corpus takes shape, maintain comparison between the actual dimensions of the material and the original plan;
6. (most important of all) document these steps so that users can have a reference point if they get unexpected results, and that improvements can be made on the basis of experience.

Let me give one simple example of these precepts in operation. The precursor of The Bank of English contained a substantial proportion of the quality fiction of the day. This came from

a request from one of the sponsors, who felt that a corpus was such an odd thing (in 1980 it was an odd thing) that users of the end products would be reassured if there was quite a lot of "good writing" in it. That is to say, under (a) above it was decided that there should be emphasis on this kind of writing; this decision affected the choice of texts under (b) also. However, one of the main aims of creating the corpus was to retrieve evidence in support of the learning of the English language, and the requirements of this mundane purpose clashed with some of the prominent features of modern fiction. For example, the broad range of verbs used to introduce speech in novels came out rather too strongly — *wail*, *bark* and *grin* are all attested in this grammatical function, and while their occurrence is of interest to students of literary style, they are of limited utility to learners seeking basic fluency in English ([Sinclair et. al. 1990](#) p. 318).

This clash between the design of the corpus and its function became clear as soon as work started on the first Cobuild grammar ([1990](#)). Because the corpus had been carefully designed and fully documented, it was possible to determine — and therefore roughly counterbalance — the bias that had been introduced. In fairness to the original designers, it should be emphasised that there were no previous models to turn to at that time, and no way of assessing the effects of different varieties of a language.

A corpus that sets out to represent a language or a variety of a language cannot predict what queries will be made of it, so users must be able to refer to its make-up in order to interpret results accurately. In everything to do with criteria, this point about documentation is crucial. So many of our decisions are subjective that it is essential that a user can inspect not only the contents of a corpus but the reasons that the contents are as they are. Sociolinguistics is extremely fashion-conscious, and arguments that are acceptable criteria during one decade may look very old-fashioned in the next.

Also at any time a researcher may get strange results, counter-intuitive and conflicting with established descriptions. Neither of these factors proves that there is something wrong with the corpus, because corpora are full of surprises, but they do cast doubt on the interpretation of the findings, and one of the researcher's first moves on encountering unexpected results will be to check that there is not something in the corpus architecture or the selection of texts that might account for it.

7. The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.

5. Balance

The notion of balance is even more vague than representativeness, but the word is frequently used, and clearly for many people it is meaningful and useful. Roughly, for a corpus to be pronounced balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgements.

Most general corpora of today are badly balanced because they do not have nearly enough spoken language in them; estimates of the optimal proportion of spoken language range from 50% — the neutral option — to 90%, following a guess that most people experience many times as much speech as writing. Another factor that affects balance is the degree of specialisation of the text, because a specialised text in a general corpus can give the impression of imbalance.

This is a problem in the area of popular magazines in English, because there are a large number of them and most use a highly specialised language that non-devotees just do not understand. So as a text type it is a very important one, but it is almost impossible to select a few texts which can claim to be representative. How are magazines for fly fishermen, personal computers and popular music going to represent the whole variety of popular magazines (as is the case in The Bank of English)?

It was mentioned above that not all cells need to be filled; for example the written component of a corpus may subdivide into newspapers, magazines, books etc., for which there are no exact equivalents in the spoken language, which might divide into broadcasts, speaker-led events, organised meetings and conversations. The idea of maintaining a balance prompts the corpus builder to try to align these categories, however roughly, so that there is not too much very formal or very informal language in the corpus as a whole. If — as is frequently reported — many users value informal and impromptu language as revealing most clearly how meaning is made, a deliberate imbalance may be created by selection in favour of this variety, and this should be documented so that users are aware of the bias that has been knowingly introduced into the corpus.

Specialised corpora are constructed after some initial selectional criteria have been applied, for example the MICASE corpus cited above. More delicate criteria are used to partition them, but the issues of balance and representativeness remain cogent and central in the design.

8. The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.

6. Topic

The point above concerning a text type where most of the exemplars are highly specialised, raises the matter of topic, which most corpus builders have a strong urge to control. Many corpus projects are so determined about this that they conduct a semantic analysis of the language on abstract principles like those of Dewey or Roget, and then search for texts that match their framework. Three problems doom this kind of enterprise to failure. One is that the corpus will not conform to the classification, the second (also faced by library cataloguers) is that no two people agree on any such analysis, and the third is that topic classification turns out to be much more sociolinguistic than semantic, and therefore dependent on the culture and not on the meanings of the words. This last point emerges strongly when we try to make corpora in more than one language but sharing the same topic classification.

As well as these practical points, our first principle rules out topic as a source of corpus criteria. The most obvious manifestation of topic is certainly found in the vocabulary, and the notion of vocabulary as a defining characteristic of a corpus is strong; hence it seems strange to many people that it is essential that the vocabulary should not be directly controlled. But vocabulary choice is clearly an internal criterion.

There are external correlates, and these will indirectly control the vocabulary of the selected texts. For example many social institutions, like professional bodies and educational establishments, do show the kind of vocabulary consistency at times that we associate with topic, and they can be used as external criteria, but topic is most definitely a matter of language patterns, mainly of vocabulary selection and discourse focusing.

9. Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.

7. Size

The minimum size of a corpus depends on two main factors:

1. the kind of query that is anticipated from users,
2. the methodology they use to study the data.

There is no maximum size. We will begin with the kind of figures found in general reference corpora, but the principles are the same, no matter how large or small the corpus happens to be. To relate the kind of query to the size of the corpus, it is best to start with a list of the "objects" that you intend to study; the usual objects are the physical word forms or objects created by tags, such as lemmas. Then try them out on one of the corpora that is easy to interrogate, such as the million-word corpora on the ICAME CD-ROM ([Hofland 1999](#)). The

Brown-group of corpora are helpful here, because they have been proof-read and tagged and edited over many years, and with a million words the sums are easy.

To illustrate how this can be done, let us take the simple case of a researcher wishing to investigate the vocabulary of a corpus. For any corpus one of the first and simplest queries is a list of word forms, which can be organised in frequency order. (NB word forms are not lemmas, where the various inflections of a "word" in the everyday sense are gathered together, but the message would not be much different with lemmas⁵).

The frequencies follow Zipf's Law ([1935](#)), which basically means that about half of them occur once only, a quarter twice only, and so on. So for the first million-word corpus of general written American English (the Brown corpus), there was a vocabulary of different word forms of 69002, of which 35065 occurred once only. At the other end of the frequency scale, the commonest word, *the* has a frequency of 69970, which is almost twice as common as the next one, *of*, at 36410.

There is very little point in studying words with one occurrence, except in specialised research, for example authorship studies ([Morton 1986](#)). Recurrence — a frequency of two or more — is the minimum to establish a case for being an independent unit of the language; but only two occurrences will tell us very little indeed about the word. At this point the researcher must fix a minimum frequency below which the word form will not be the object of study. Let us suggest some outline figures that may guide practice. A word which is not specially ambiguous will require at least twenty instances for even an outline description of its behaviour to be compiled by trained lexicographers. But there are other factors to consider, the consequences of what seems to be a general point that alternatives — members of a set or system — are often not equally likely. The same tendency that we see in Zipf's Law is found in many other places in the numerical analysis of a corpus. Very often the main meaning or use or grammatical choice of a word is many times as frequent as the next one, and so on, so that twenty occurrences may be sufficient for the principal meaning of a word, while some quite familiar senses may occur only seldom. This applies also to frequent words which can have some important meanings or uses which are much less common than the principal ones. Word classes occur in very different proportions, so if the word can be both noun and verb, the verb uses are likely to be swamped by the noun ones, and for the verb uses researchers often have recourse to a tagged corpus. In many grammatical systems one choice is nine times as common as the other ([Halliday 1993](#)), so that for every negative there are nine positives.

So some additional leeway will have to be built in to cope with such contingencies. If the objects of study are lemmas rather than word forms, the picture is not very different. The

minimum number of instances needed for a rough outline of usage will rise to an average of about fifty for English (but many more for highly inflected languages).

If the research is about events which are more complicated than just word occurrence, then the estimate of a suitable corpus size will also get more complicated. For example if the research is about multi-word phrases, it must be remembered that the occurrence of two or more words together is inherently far rarer than either on its own. So if each of the two words in a minimal phrase occur 20 times in a million word corpus, for 20 instances of the two together the arithmetic suggests a corpus of approximately 5 billion words will be needed. For three words together of this frequency the size of the corpus could be beyond our imaginings.

However, words do not occur according to the laws of chance, and if the phrases chosen are normal ones in the language, they will occur many times more often than the arithmetic projection above; so a much smaller corpus is likely to contain sufficient instances. To estimate roughly the size of a corpus for retrieval of a combination of two objects, first estimate the size you will need for the less common object on its own and then raise that figure by an order of magnitude. If there are 20 instances per million words for each of two words in a phrase, then twenty million words is likely to provide 20 instances of the pair (rather than the 5 billion projected by the arithmetic); if there are three of this frequency than 200 million words will probably be enough.

These are the kinds of figures that you will need to use in estimates of your optimal corpus size. Now we must build in the considerations of the methodology that you intend to use, because this can have a dramatic effect on the size.

The main methodological point is whether, having examined directly the initial results of corpus searches you intend to return to indirect methods and use the computer for further stages, recycling and refining early results⁶. If the latter, you will have to increase the minimum number of occurrences of your object quite substantially. This is because the regularities of occurrence that the machine will search for are not on the surface, and the way the computer works is to examine the *cotexts* minutely searching for frequently repeated patterns. Having found these it can then isolate instances of unusual and particular co-occurrences, which can either be discarded or studied separately after the main patterns have been described. For example, if the computer searches for the adjectives that come between *in* and *trouble*, in text sequence (Bank of English 17/10/04) these are:

unspecified, terrible, deep, serious, deep, Cuba, serious, serious, great...

It is reasonable already to anticipate that *deep* and *serious* are likely to be important recurrent collocates, but single instances of the others do not offer useful evidence. In fact

unspecified does not recur, *terrible* is a good collocate, with 33 instances out of 1729. *Deep* is an important collocate with 251 instances, 14.5%, while *Cuba* is unique. *Serious* is slightly greater than *deep* at 271. *Great*, on the other hand, scores merely 8. The next in sequence is *big*, which at 235 instances is up with *deep* and *serious*. As we examine more and more instances, these three adjectives gradually separate themselves from all the others because of the number of times they appear — in total (757), almost half of all the instances. The nearest contender is *real*, at 142 quite considerably less common, and after that *financial* at 113. The computer also records as significant collocates *terrible* (35), *dire* (31) and *desperate* (28); *deeper* (14), *double* (14), *foul* (11), *bad* (14), *such* (28), *enough* (17) and *worse* (11).

The pure frequency picks out the three or four collocates that are closely associated with the phrase *in trouble*, and reference to the statistical test (here the t-score) adds another dozen or so adjectives which, while less common in the pattern are still significantly associated and add to the general gloom that surrounds the phrase. Single occurrences like *unspecified* and *Cuba* drop into obscurity, as do *terminal* (2) and *severe* (4), which occur among the first 30 instances.

The density of the patterns of collocation is one of the determinants of the optimal size of a corpus. Other factors include the range of ambiguity of a word chosen, and sometimes its distribution among the corpus components.

If you intend to continue examining the first results using the computer, you will probably need several hundred instances of the simplest objects, so that the programs can penetrate below the surface variation and isolate the generalities. The more you can gather, the clearer and more accurate will be the picture that you get of the language.

8. Specialised corpora

The proportions suggested above relate to the characteristics of general reference corpora, and they do not necessarily hold good for other kinds of corpus. For example, it is reasonable to suppose that a corpus that is specialised within a certain subject area will have a greater concentration of vocabulary than a broad-ranging corpus, and that is certainly the case of a corpus of the English of Computing Science ([James et al 1994](#)). It is a million words in length, and some comparisons with a general corpus of the same length (the LOB corpus) are given in Table 1 (the corpus of English of Computing Science is designated as 'HK').

	LOB	HK	%
Number of different word-forms (types)	69990	27210	39%
Number that occur once only	36796	11430	31%
Number that occur twice only	9890	3837	39%
Twenty times or more	4750	3811	80%
200 times or more	471	687	(69%)

Table 1. Comparison of frequencies in a general and a specialised corpus.

The number of different word forms, which is a rough estimate of the size of the vocabulary, is far less in the specialised text than it is in the general one — less than 40% of its size. The proportion of single occurrences is another indication of the spread of the vocabulary, and here the proportional difference between the two corpora is even greater, with the specialised corpus having only 31% of the total of the other corpus. Word forms which occur twice are also much less common in the specialised corpus, but the gap closes quite dramatically when we look at the figures for twenty occurrences. At a frequency of 200 and above the proportions are the other way round, and the general corpus has only 69% of the number of such words in the specialised corpus. Assuming that the distribution of the extremely common words is similar in the two corpora, these figures suggest that the specialised corpus highlights a small, probably technical vocabulary.

This is only one example, but it is good news for builders of specialised corpora, in that not only are they likely to contain fewer words in all, but it seems as if the characteristic vocabulary of the special area is prominently featured in the frequency lists, and therefore that a much smaller corpus will be needed for typical studies than is needed for a general view of the language.

9. Homogeneity

The underlying factor is homogeneity. Two general corpora may differ in their frequency profile if one is more homogenous than the other, while specialised corpora, by reducing the variables, offer a substantial gain in homogeneity.

Homogeneity is a useful practical notion in corpus building, but since it is superficially like a bundle of internal criteria we must tread very carefully to avoid the danger of vicious circles. As long as the choice of texts in a corpus still rests ultimately with the expertise and common sense of the linguist, it is appropriate for the linguist to use these skills to reject obviously odd

or unusual texts. In any variety of a language there will be some texts — "rogue" texts — which stand out as radically different from the others in their putative category, and therefore unrepresentative of the variety on intuitive grounds. If they are included because of some high principle of objectivity, they are just wasted storage in the computer⁷. The principle of recurrence (see below) implies that a single occurrence of a feature is unlikely to be accepted as an authentic feature of a language or variety; hence unless texts share a large number of features the corpus will be of little use. There is a balance to be struck between coverage and homogeneity in the attempt to achieve representativeness.

The use of homogeneity as a criterion for acceptance of a text into a corpus is based certainly on the impression given by some features of its language, but is a long way from the use of internal criteria. A corpus builder who feels that this criterion might threaten the status of the corpus can of course simply not make use of it, because it is really just a short cut. Rogue texts are usually easy to identify, and of course they must be genuinely exceptional; if we begin to perceive groups of them then it is our classification that must be re-examined, because we may have inadvertently collapsed two quite distinct text types.

It must be conceded at this point that we have moved in corpus design away from a completely objective stance and a blind reliance on objective external criteria. It is pleasant to envisage a utopia where corpora will be so large that a proportion of unique and strange texts can be included (if selected on objective criteria) without sabotaging the aims of the corpus design; if so this happy state of affairs is still quite a long way off. Such texts would largely disappear because their patterns would never be strong enough to be retrieved, so while corpora are still, in my opinion, very small indeed it is sensible in practical terms not to put "rogue texts" in at all. Provided that designers and builders accept the burden of documenting their decisions, there is little danger of distortion.

10. A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.

10. Character of corpus research

It is necessary to say something here about the "typical studies" mentioned above, because at many points in this chapter there are assumptions made about the nature of the research enquiries that engage a corpus. This section is not intended in any way to limit or circumscribe any use of corpora in research, and we must expect fast development of new methodologies as corpora become more accessible and the software more flexible. But in any resource provision, the provider must have some idea of the use to which the resource will be put, and that is certainly so with corpora.

Corpus research is mainly centred on the recurrence of objects; initially surface entities like word forms, objects can be re-defined after going through a process of generalisation, which means that forms which are not identical can be classified as instances of the same object. As noted above, the lemma is a clear example of this process.

Studies range from (a) close attention to textual interpretation, using only a few instances, through (b) the substantial quantities needed for language description on which the section above on "size" is based, to (c) large-scale statistical processing. All rely on recurrence as their starting point. The opposite of recurrence, uniqueness, cannot be observed with certainty in a corpus, because, as conceded near the beginning of this chapter, uniqueness in a corpus does not entail uniqueness in a language. However, very rare events can be, and are, studied, and of course the arithmetic of combinations means that most stretches of text that are more than a few words long are unlikely to recur, ever.

The use of a corpus adds quite literally another dimension to language research. If you examine a KWIC concordance, which is the standard format for reporting on recurrence, it is clear that the horizontal dimension is the textual one, which you read for understanding the progress of the text and the meaning it makes as a linear string, while the vertical dimension shows the similarities and differences between one line and the lines round about it. The main "added value" of a corpus is this vertical dimension, which allows a researcher to make generalities from the recurrences.

The current dilemma of much corpus linguistics is that the number of occurrences that a researcher can examine at once — in practice a screenful, some 23 lines — is a rather small amount of evidence, given the great variability of the language in use. On the other hand, to hand over the examination to the computer, where almost any number of instances could be processed very quickly, requires programming skills or a thorough knowledge of available software resources and how to make use of them. There is obvious advantage in getting the machine to do as much of the work as possible — in particular the gain in objectivity that results — but it requires much more investment in advance than the simple direct scrutiny of a small sample.

11. What is not a corpus?

As we move towards a definition of a corpus, we remind ourselves of some of the things that a corpus might be confused with, because there are many collections of language text that are nothing like corpora.

The World Wide Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective. At present it is

quite mysterious, because the search engines, through which the retrieval programs operate, are all different, none of them are comprehensive, and it is not at all clear what population is being sampled. Nevertheless, the WWW is a remarkable new resource for any worker in language (see [Appendix](#)), and we will come to understand how to make best use of it.

An archive is not a corpus. Here the main difference is the reason for gathering the texts, which leads to quite different priorities in the gathering of information about the individual texts.

A collection of citations is not a corpus. A citation is a short quotation which contains a word or phrase that is the reason for its selection. Hence it is obviously the result of applying internal criteria. Citations also because lack the textual continuity and anonymity that characterise instances taken from a corpus; the precise location of a quotation is not important information for a corpus researcher.

A collection of quotations is not a corpus for much the same reasons as a collection of citations; a quotation is a short selection from a text, chosen on internal criteria and chosen by human beings and not machines.

These last two collections correspond more closely to a concordance than a corpus. A concordance also consists of short extracts from a corpus, but the extracts are chosen by a computer program, and are not subject to human intervention in the first instance. Also the constituents of a corpus are known, and searches are comprehensive and unbiased. Some collections of citations or quotations may share some or all of these criteria, but there is no requirement for them to adopt such constraints. A corpus researcher has no choice, because he or she is committed to acquire information by indirectly searching the corpus, large or small.

A text is not a corpus. The main difference ([Tognini Bonelli 2001 p.3](#)) is the dimensional one explained above. Considering a short stretch of language as part of a text is to examine its particular contribution to the meaning of the text, including its position in the text and the details of meaning that come from this unique event. If the same stretch of language is considered as part of a corpus, the focus is on its contribution to the generalisations that illuminate the nature and structure of the language as a whole, far removed from the individuality of utterance.

12. Definition

After this discussion we can make a reasonable short definition of a corpus. I use the neutral word "pieces" because some corpora still use sample methods rather than gather complete texts or transcripts of complete speech events. "Represent" is used boldly but qualified. The

primary purpose of corpora is stressed so that they are not confused with other collections of language.

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

Acknowledgements

This chapter relates historically to a paper entitled *Corpus Creation* which was presented to the Council of Europe in February 1987; it was revised for publication in [Sinclair \(1989\)](#) and updated again as a chapter in [Sinclair \(1991\)](#). After a further decade it has been completely rewritten, but covers much the same ground as the earlier papers.

I am grateful to Knut Hofland, Sattar Izwaini and Martin Wynne for help with Table 1 and the other statistics.

Notes

- [1.](#) See the brief discussion on homogeneity later in this chapter.
- [2.](#) For a discussion of the role and limitations of intuition, see [Sinclair \(2004\)](#).
- [3.](#) See the MICASE website, <http://www.hti.umich.edu/m/micase/> under "Speech event and speaker categories", which is a very elaborate classification, leading to small cells and many empty ones.
- [4.](#) For further discussion of this point see [Sinclair 2004](#).
- [5.](#) Knut Hofland reports that in the LOB corpus there are 25,992 instances of tag/word combinations that occur once only, as compared with 36,796 word forms (see [Table 1](#)). While the lemmatisation reduces the number of different objects, the tag assignment increases the number by giving more than one tag to the same form.
- [6.](#) There is a brief treatment of this point in [Sinclair \(2001\)](#).
- [7.](#) One good example of a rogue text appeared in one of the earliest specialised corpora — Roe's corpus of textbooks in physical science ([Roe 1977](#)). This million-word corpus held a dozen or so full-text documents, which showed considerable homogeneity, all except one. The rogue text turned out to be an Open University textbook, and it reflected the innovative style of OU teaching, the new student body attracted to the OU, and the resolve to make learning a part of life. So any generalisation that applied to all the other texts was not supported by the OU text, and virtually none of its features were found in the others. The contrast was so marked that Roe had to make a separate statement for the OU text and one

for all the rest. The text excluded itself, which was ironic because Roe had chosen it deliberately as an example of good current communication; its "rogue" status has nothing to do with its worth as an academic textbook, but shows the sharp difference in approach that is associated with the OU.

Chapter 2: Adding Linguistic Annotation (Geoffrey Leech, Lancaster University © Geoffrey Leech 2004)

1. What is corpus annotation?

Corpus annotation is the practice of adding interpretative linguistic information to a corpus. For example, one common type of annotation is the addition of *tags*, or labels, indicating the word class to which words in a text belong. This is so-called part-of-speech tagging (or POS tagging), and can be useful, for example, in distinguishing words which have the same spelling, but different meanings or pronunciation. If a word in a text is spelt *present*, it may be a noun (= 'gift'), a verb (= 'give someone a present') or an adjective (= 'not absent'). The meanings of these same-looking words are very different, and also there is a difference of pronunciation, since the verb *present* has stress on the final syllable. Using one simple method of representing the POS tags — attaching tags to words by an underscore symbol — these three words may be annotated as follows:

*present*_NN1 (singular common noun) *present*_VVB (base form of a lexical verb)
*present*_JJ (general adjective)

Some people (notably John Sinclair — see [chapter 1](#)) prefer not to engage in corpus annotation: for them, the unannotated corpus is the 'pure' corpus they want to investigate — the corpus without adulteration with information which is suspect, possibly reflecting the predilections, or even the errors, of the annotator. For others, annotation is a means to make a corpus much more useful — an enrichment of the original **raw corpus**. From this perspective, probably a majority view, adding annotation to a corpus is giving 'added value', which can be used for research by the individual or team that carried out the annotation, but which can also be passed on to others who may find it useful for their own purposes. For example, POS-tagged versions of major English language corpora such as the Brown Corpus, the LOB Corpus and the British National Corpus have been distributed widely throughout the world for those who would like to make use of the tagging, as well as of the original 'raw' corpus. In this chapter, I will assume that such annotation is a benefit, so long as it is done well, with an eye to the standards that ought to apply to such work.

2. What different kinds of annotation are there?

Apart from part-of-speech (POS) tagging, there are other types of annotation, corresponding to different levels of linguistic analysis of a corpus or text — for example:

phonetic annotation

e.g. adding information about how a word in a spoken corpus was pronounced.
prosodic annotation — again in a spoken corpus — adding information about prosodic features such as stress, intonation and pauses.
syntactic annotation — e.g. adding information about how a given sentence is parsed, in terms of syntactic analysis into such units such phrases and clauses

semantic annotation

e.g. adding information about the semantic category of words — the noun cricket as a term for a sport and as a term for an insect belong to different semantic categories, although there is no difference in spelling or pronunciation.

pragmatic annotation

e.g. adding information about the kinds of speech act (or dialogue act) that occur in a spoken dialogue — thus the utterance okay on different occasions may be an acknowledgement, a request for feedback, an acceptance, or a pragmatic marker initiating a new phase of discussion.

discourse annotation

e.g. adding information about anaphoric links in a text, for example connecting the pronoun *them* and its antecedent *the horses* in: *I'll saddle the horses and bring them round.* [an example from the Brown corpus]

stylistic annotation

e.g. adding information about speech and thought presentation (direct speech, indirect speech, free indirect thought, etc.)

lexical annotation

adding the identity of the lemma of each word form in a text — i.e. the base form of the word, such as would occur as its headword in a dictionary (e.g. *lying* has the lemma LIE).

(For further information on such kinds of annotation, see [Garside et al. 1997](#).) In fact, it is possible to think up untold kinds of annotation that might be useful for specific kinds of research. One example is dysfluency annotation: those working on spoken data may wish to annotate a corpus of spontaneous speech for dysfluencies such as false starts, repeats, hesitations, etc. — see Lickley, no date). Another illustration comes from an area of corpus research which has flourished in the last ten years: the creation and study of learner corpora ([Granger 1998](#)). Such corpora, consisting of writing (or speech) produced by learners of a second language, may be annotated with 'error tags' indicating where the learner has produced errors, and what kinds of errors these are ([Granger et al 2002](#)).

3. Why annotate?

As I have already indicated, annotation is undertaken to give 'added value' to the corpus. A glance at some of the advantages of an annotated corpus will help us to think about the standards of good practice these corpora require.

Manual examination of a corpus

What has been built into the corpus in the form of annotations can also be extracted from the corpus again, and used in various ways. For example, one of the main uses of POS tagging is to enhance the use of a corpus in making dictionaries. Thus lexicographers, searching through a corpus by means of a concordancer, will want to be able to distinguish *separate* (verb) from *separate* (adjective), and if this distinction is already signalled in the corpus by tags, the separation can be automatic, without the painstaking search through hundreds or thousands of examples that might otherwise be necessary. Equally, a grammarian wanting to examine the use of progressive aspect in English (is working, has been eating, etc) can simply search, using appropriate search software, for sequences of BE (any form of the lemma) followed — allowing for certain possibilities of intervening words — by the *ing*-form of a verb.

Automatic analysis of a corpus

Similarly, if a corpus has been annotated in advance, this will help in many kinds of automatic processing or analysis. For example, corpora which have been POS-tagged can automatically yield frequency lists or frequency dictionaries with grammatical classification. Such listings will treat *leaves* (verb) and *leaves* (noun) as different words, to be listed and counted separately, as for most purposes they should be. Another important case is automatic parsing, i.e. the automatic syntactic analysis of a text or a corpus: the prior tagging of a text can be seen as a first stage of syntactic analysis from which parsing can proceed with greater success. Thirdly, consider the case of speech synthesis: if a text is to be read aloud by a speech synthesiser, as in the case of the 'talking books' service provided for the blind, the synthesiser needs to have the information that a particular instance of *sow* is a noun (= female pig) rather than a verb (as in *to sow seeds*), because this makes a difference to the word's pronunciation.

Re-usability of annotations

Some people may say that the annotation of a corpus for the above cases is not needed, automatic processing could include the analysis of such features as part of speech: it is unnecessary thereafter to preserve a copy of the corpus with the built-in information about

word class. This argument may work for some cases, but generally the annotation is far more useful if it is preserved for future use. The fact is that linguistic annotation cannot be done accurately and automatically: because of the complex and ambiguous nature of language, even a relatively simple annotation task such as POS-tagging can only be done automatically with up to 95% to 98% accuracy. This is far from ideal, and to obtain an optimally tagged corpus, it is necessary to undertake manual work, often on a large scale. The automatically tagged corpus afterwards has to be post-edited by a team of human beings, who may spend thousands of hours on it. The result of such work, if it makes the corpus more useful, should be built into a tagged version of the corpus, which can then be made available to any people who want to use the tagging as a springboard for their own research. In practice, such corpora as the LOB Corpus and the BNC Sampler Corpus have been manually post-edited and the tagging has been used by thousands of people. The BNC itself — all 100 million words of it — has been automatically tagged but has not been manually post-edited, as the expense of undertaking this task would be prohibitive. But the percentage of error — 2% — is small enough to be discounted for many purposes. So my conclusion is that — as long as the annotation provided is a kind useful to many users — an annotated corpus gives 'value added' because it can be readily shared by others, apart from those who originally added the annotation. In short, an annotated corpus is a sharable resource, an example of the electronic resources increasingly relied on for research and study in the humanities and social sciences.

Multi-functionality

If we take the re-usability argument one step further, we note that annotation often has many different purposes or applications: it is **multi-functional**. This has already been illustrated in the case of POS tagging: the same information about the grammatical class of words can be used for lexicography, for parsing, for frequency lists, for speech synthesis, and for many other applications. People who build corpora are familiar with the idea that no one in their right mind would offer to predict the future uses of a corpus — future uses are always more variable than the originator of the corpus could have imagined! The same is true of an annotated corpus: the annotations themselves spark off a whole new range of uses which would not have been practicable unless the corpus had been annotated.

However, this multi-functionality argument does not always score points for annotated corpora. There is a contrary argument that the annotations are more useful, the more they are designed to be specific to a particular application.

4. Useful standards for corpus annotation

What I have said above about the usefulness of annotated corpora, of course, depends crucially on whether the annotation has been well planned and well carried out. It is important, then, to recommend a set of standards of good practice to be observed by annotators wherever possible.

Annotations should be separable

The annotations are added as an 'optional extra' to the corpus. It should always be easy to separate the annotations from the raw corpus, so that the raw corpus can be retrieved exactly in the form it had before the annotations were added. This is common sense: not all users will find the annotations useful, and annotation should never result in any loss of information about the original corpus data.

Detailed and explicit documentation should be provided

Lou Burnard (in [chapter 3](#)) emphasises the need to provide adequate documentation about the corpus and its constituent texts. For similar reasons, it is important to provide explicit and detailed documentation about the annotations in an annotated corpus. Documentation to be provided about annotations should include the following, so that users will know precisely what they're getting:

How, where, when and by whom were the annotations applied?

Mention any computer tools used, and any phases of revision resulting in new releases, etc.

What annotation scheme was applied?

An annotation scheme is an explanatory system supplying information about the annotation practices followed, and the explicit interpretation, in terms of linguistic terminology and analysis, for the annotation. This is very important — Section 6 below will deal with annotation schemes.

What coding scheme was used for the annotations?

By coding scheme, I mean the set of symbolic conventions employed to represent the annotations themselves, as distinct from the original corpus. Again, I will devote a separate section to this (Section 5).

How good is the annotation?

It might be thought that annotators will always proclaim the excellence of their annotations. However, although some aspects of 'goodness' or quality elude judgement, others can be measured with a degree of objectivity: accuracy and consistency are two such measures. Annotators should supply what information they can on the quality of the annotation. (see further Section 8 below.)

Arguably, the annotation practices should be linguistically consensual

This and the following maxims are more open to debate. Any type of annotation presupposes a typology — a system of classification — for the phenomena being represented. But linguistics, like most academic disciplines, is sadly lacking in agreement about the categories to be used in such description. Different terminologies abound, and even the use of a single term, such as *verb phrase*, is notoriously a prey to competing theories. Even an apparently simple matter, such as defining word classes (POS), is open to considerable disagreement. Against this background, it might be suggested that corpus annotation cannot be usefully attempted: there is no absolute 'God's truth' view of language or 'gold standard' annotation against which the decision to call word *x* as noun and word *y* a verb can be measured.

However, looking at linguistics more carefully, we can usually observe a certain consensus: examining a text, people can more or less agree which words are nouns, verbs, and so on, although they may disagree on less clear cases. If this is reasonable, then an annotation scheme can be based on a 'consensual' set of categories on which people tend to agree. This is likely to be useful for other users and therefore to fit in with the re-usability goal for annotated corpora. An annotation scheme can additionally make explicit how the annotations apply to the 10% or so of less clear cases, so that users will know how borderline phenomena are handled. Significantly, this consensual approach to categories is found not only in annotated corpora, but also in another key kind of linguistic resource — dictionaries. If, on the other hand, an annotator were to use categories specific to a particular theory and out of line with other theories, the annotated corpus would suffer in being less useful as a sharable resource.

Annotation practices should respect emergent *de facto* standards

This principle of good practice may be seen as complementary to the preceding one. By *de facto* standards, I mean some kind of standardisation that has already begun to take place, due to influential precedents or practical initiatives in the research community. These contrast with *de iure* or 'God's truth' standards, which I have just argued do not exist. 'God's truth' standards, if they existed, would be imposed from on high. *De facto* standards, on the other hand, emerge (often gradually) from the research community in a bottom-up manner.

De facto standards encapsulate what people have found to work in the past, which argues that they should be adopted by people undertaking a new research project, to support a growing consensus in the community. However, often a new project breaks new ground, for example with a different kind of data, a different language, a different purpose those of previous projects. It would clearly be a recipe for stagnation if we were to coerce new projects into the following exactly the practices of earlier ones. Nevertheless it makes sense for new projects to respect the outcomes of earlier projects, and only to depart from their practices where this can be justified. In 8 below, I will refer to some of the incipient standards for different kinds of annotation and mark-up. These can only be presented tentatively, however, as the practice of corpus annotation is continually evolving.

In the early 1990s, the European Union launched an initiative under the name of EAGLES (Expert Advisory Groups on Language Engineering Standards) with the goal of encouraging standardisation of practices for natural language processing in academia and industry, particularly but not exclusively in the EU. One group of 'experts' set to work on corpora, and from this and later initiatives there emerged various documents specifying guidelines (or provisional standards) for corpus annotation. In the following sections, I will refer to the EAGLES documents where appropriate.

5. The encoding of annotations

But before focussing on annotation schemes and the linguistic categories they incorporate, it will be helpful to touch briefly on the encoding of annotations — that is, the actual symbolic representations used. This means we are for the moment concentrating on how annotations are outwardly manifested — for example, what you see when you inspect a corpus file on your computer screen — rather than what their meaning is, in linguistic terms.

As an example, I have already mentioned one very simple device, the underscore symbol, to signal the attachment of a POS tag to a word, as in *Paula*_NP1. The presentation of the tag itself may be complex or simple. Here, for convenience, the category of 'singular proper noun' is represented by a sequence of three characters, N for noun, P for proper (noun), and 1 for singular.

One basic requirement is that the POS tag (or any other annotation device) should be unambiguous in representing what it stands for. Another requirement, useful for everyday purposes such as reading a concordance on a screen, is brevity: the three characters, in this case, concisely signal the three distinguishing grammatical features of the NP1 category. A third requirement, more useful in some contexts than in others, is that the annotation device should be transparent to the human reader rather than opaque. The example NP1 is at least

to some degree intelligible, and is less mystifying than it would be if some arbitrary sequence of symbols, say Q!@, had been chosen.

The type of tag illustrated above originated with the earliest corpus to be POS-tagged (in 1971), the Brown Corpus. More recently, since the early 1990s, there has been a far-reaching trend to standardize the representation of all phenomena of a corpus, including annotations, by the use of a standard mark-up language — normally one of the series of related languages SGML, HTML, and XML (see Lou Burnard, [chapter 3](#)). One advantage of using these languages for encoding features in a text is that they provide a general means of interchange of documents, including corpora, between one user or research site and another. In this sense, SGML/HTML/XML have developed into a world-wide standard which can be applied to any language, to spoken as well as to written language, and to languages of different historical periods. Furthermore, the use of the mark-up language itself can be efficiently parsed or validated, enabling the annotator to check whether there are any ill-formed traits in the markup, which would signal errors or omissions. Yet another advantage is that, as time progresses, tools of various kinds are being developed to facilitate the processing of texts encoded in these languages. One example is the set of tools developed at the Human Communication Research Centre, Edinburgh, for supporting linguistic annotation using XML ([Carletta et al. 2002](#)).

However, one drawback of these mark-up languages is that they tend to be more 'verbose' than the earlier symbolic conventions used, for example, for the Brown and LOB corpora. In this connection we can compare the LOB representation *Paula*_NP1 ([Johansson 1986](#)) with the SGML representation to be found in the BNC (first released in 1995): <w NP1>*Paula*, or the even more verbose version if a closing tag is added, as required by XML: <w type="NP1">*Paula*</w>. In practice, this verbosity can be avoided by a conversion routine which could produce an output, if required, as simple as the LOB one *Paula*_NP1. This, however, would require a further step of processing which may not be easy to manage for the technically less adept user.

Another possible drawback of the SGML/XML type of encoding is that it requires a high-resolution standard of validation which sorts ill with the immensely unpredictable nature of a real-world corpus. This is a particular problem if that corpus contains spontaneous spoken data and data from less 'orderly' varieties of written language — e.g. mediaeval manuscripts, old printed editions, advertisements, handwritten personal letters, collections of children's writing. Attempts have been made to make this type of logical encoding more accessible, by relaxing standards of conformance. Hence there has grown up a practice of encoding corpora using a so-called 'pseudo-SGML', which has the outward characteristics of SGML,

but is not subjected to the same rigorous process of validation (so that errors of well-formedness may remain undetected).

Within the overall framework SGML, different co-existing encoding standards have been proposed or implemented: notably, the CDIF standard used for the mark-up of the BNC (see [Burnard 1995](#)) and the CES recommended as an EAGLES standard ([Ide 1996](#)). One further drawback of the SGML/XML approach to encoding is that it assumes, by default, that annotation has a 'parsable' hierarchical tree structure, which does not allow cross-cutting brackets as in `<x ...> ... <y...> ... <x/> ... <y/>`. Any corpus of spoken data, in particular, is likely to contain such cross-bracketing, for example in the cross-cutting of stretches of speech which need to be marked for different levels of linguistic information — such phenomena as non-fluencies, interruptions, turn overlaps, and grammatical structure are prone to cut across one another in complex ways.

This difficulty can be overcome within SGML/XML, although not without adding considerably to the complexity of the mark-up — for example, by copious use of pointer devices (in the BNC) or by the use of so-called stand-off annotation ([Carletta et al. 2002](#)).

It is fair to say, in conclusion, that the triumph of the more advanced SGML/HTML/XML style of encoding is in the long run assured. But because of the difficulties I have mentioned, many people will find it easier meanwhile to follow the lead of other well-known encoding schemes — such as the simpler styles of mark-up associated with the Brown and ICE families of corpora, or with the CHILDES database of child language data.

CHILDES ('child language data exchange system') is likely to be the first choice not only for those working on child language corpora, but on related fields such as second language acquisition and code-switching. As the name suggests, CHILDES is neither a corpus nor a coding scheme in itself, but it provides both, operating as a service which pools together the data of many researchers all over the world, using a common coding and annotation schemes, and common software including annotation software.

6. Annotation manual

Why do we need an annotation manual? This document is needed to explain the annotation scheme to the users of an annotated corpus. Typically such manuals originate from sets of guidelines which evolve in the process of annotating a corpus — especially if hand editing of the corpus has been undertaken. A most carefully worked-out annotation scheme was published as a weighty book by Geoffrey Sampson ([1995](#)). This explained in detail the parsing scheme of the SUSANNE corpus (a syntactically-annotated part of the Brown corpus). Sampson made an interesting analogy between developing an annotation scheme

and laying down a legal system by the tradition of common law — the 'case law' of annotation evolves, rather as the law evolves over time, through the precedent of earlier cases and the setting of new precedents as need arises.

Although annotation manuals often build up piecemeal in this way, for the present purpose we should see them as completed documents intended for corpus users. They can be thought of as consisting of two sections — (a) a list of annotation devices and (b) a specification of annotation practices — which I will illustrate, as before, using the familiar case of a POS tagging scheme (for an example, see [Johansson, 1986](#), for the LOB Corpus, or Sampson, [1995](#), Ch.3 for the SUSANNE Corpus).

A list of annotation devices with brief explanations

This list acts as a glossary — a convenient first port of call for people trying to make sense of the annotations. For POS tagging, the first thing to list is the tagset — i.e., the list of symbols used for representing different POS categories. Such tagsets vary in size, from about 30 tags to about 270 tags. The tagset can be listed together with a simple definition and exemplification of what the tag means:

NN1 singular common noun (e.g. *book, girl*) NN2 plural common noun (e.g. *books, girls*)
NP1 singular proper noun (e.g. *Susan, Cairo*) etc.

A specification of annotation practices

This gives an account of the various annotation decisions made in:

1. segmentation: e.g. assignment of POS tags assumes a prior segmentation of the corpus into words. This may involve 'grey areas' such as how to deal with hyphenated words, acronyms, enclitic forms such as the *n't* of *don't*.
2. embedding: e.g. in parsing, some units, such as words and phrases, may be included in other units, such as clauses and sentences; certain embeddings, however, may be disallowed. In effect, a grammar of the parsing scheme has to be supplied. Even POS tagging has to involve some embedding when we come to segment examples such as *the New York-Los Angeles flight*.
3. the rules or guidelines for assigning particular annotation devices to particular stretches of text.

The last of these, (c), is the most important: the guidelines on how to annotate particular pieces of text can be elaborated almost *ad infinitum*. Taking again the example of POS tagging, consider what this means with a particular tag such as NP1 (singular proper noun). In the automatic tagging process, a dictionary that matches words to tags can make a large majority of such decisions without human intervention. But problems arise, as always, with 'grey areas' that the manual must attempt to specify. For example, should New York be

tagged as one example of NP1 or two? Should the tag NP1 apply to *[the] Pope*, *[the] Renaissance*, *Auntie*, *Gold* (in *Gold Coast*), *Fifth* (in *Fifth Avenue*), *T* and *S* (in *T S Eliot*), *Microsoft* and *Word* in *Microsoft Word*? If not, what alternative tags should be applied to these cases? The manual should if possible answer such questions in a principled way, so that consistency of annotation practices between different texts and different annotators can be ensured and verified. But inevitably some purely arbitrary distinctions have to be made. Languages suffer to varying extents from ambiguity of word classifications, and in a language like English, a considerable percentage of words have to be tagged variably according to their context of occurrence.

Other languages have different problems: for example, in German the initial capital is used for common nouns as well as for proper nouns, and cannot be used as a criterion for NP1. In Chinese, there is no signal of proper noun status such as capital letters in alphabetic languages. Indeed, more broadly considered, the whole classification of parts of speech in the Western tradition is of doubtful validity for languages like Chinese.

7. Some 'provisional standards' of best practice for different linguistics levels

In this section I will briefly list and comment on some previous work in developing provisional *de facto* standards (see 4 above) of good practice for different levels of linguistic annotation. The main message here is that anyone starting to undertake annotation of a corpus at a particular level should take notice of previous work which might provide a model for new work. There are two caveats, however: (a) these are only a few of the references that might be chased up, and (b) most of these references are for English. If you are thinking of annotating a corpus of another language, especially one which corpus linguistics has neglected up to now, it makes sense to hunt down any work going forward on that language, or on a closely related language. For this purpose, grammars, dictionaries and other linguistic publications on the language should not be neglected, even if they belong to the pre-corpus age.

Part-of-speech (POS) tagging

- The 'Brown Family' of corpora (consisting of the Brown Corpus, the LOB Corpus, the Frown Corpus and the FLOB Corpus) makes use of a family of similar tagging practices, originated at Brown University and further developed at Lancaster. The two tagsets (C5 and C7) used for the tagging of the British National Corpus are well known (see [Garside et al. 1997](#): 254-260).

- An EAGLES document which recommends flexible 'standard' guidelines for EU languages is to be found in Leech and Wilson ([1994](#)), revised and abbreviated in Leech and Wilson ([1999](#)).
- Note that POS tagging schemes are often part of parsing schemes, to be considered under the next heading.

Syntactic annotation

- A well-developed parsing scheme already mentioned is that of the SUSANNE Corpus, Sampson ([1995](#)).
- The Penn Treebank and its accompanying parsing scheme has been the most influential of constituent structure schemes for syntax. (see [Marcus et al 1993](#))
- Other schemes have adopted a dependency model rather than a constituent structure model — particularly the Constraint Grammar model of Karlsson et al. ([1995](#)).
- [Leech, Barnett and Kahrel \(1995\)](#) is another EAGLES 'standards-setting' document, this time focussing on guidelines for syntactic annotation. Because there can be fundamentally different models of syntactic analysis, this document is more tentative (even) than the Leech and Wilson one for POS tagging.

Prosodic annotation

- The standard system for annotating prosody (stress, intonation, etc.) is ToBI (= Tones and Break Indices), which comes with its own speech-processing platform. Its phonological model originated with [Pierrehumbert \(1980\)](#). The system is partially automated, but needs to be substantially adapted for fresh languages and dialects.
- ToBI is well supported by dedicated software and a committed research community. On the other hand, it has met with criticism, and two alternative annotation systems worth examining are INTSINT (see [Hirst 1991](#)) and TSM — tonetic stress marks (see [Knowles et al. 1996](#)).
- For a survey of prosodic annotation of dialogue, see [Grice et al. \(2000: 39-54\)](#).

Pragmatic/Discourse annotation

For corpus annotation, it is difficult to draw a line between pragmatics and discourse analysis.

- An international Discourse Resource Initiative (DRI) came up with some recommendations for the analysis of spoken discourse at the level of dialogue acts (= speech acts) and at higher levels such as dialogue transactions, constituting a kind of 'grammar' of discourse. These were set out in the DAMSL manual (= Dialog Act Markup in Several Layers) ([Allen and Core 1997](#)).
- Other influential schemes are those of TRAINS, VERBMOBIL, the Edinburgh Map Task Corpus, SPAAC ([Leech and Weisser 2003](#)). These all focus on practical task-oriented dialogue. One exceptional case is the Switchboard DAMSL annotation project ([Stolcke et al. 2000](#)), applied to telephone conversational data.
- Discourse can also be analysed at the level of anaphoric relations (e.g. pronouns and their antecedents — see [Garside et al 1997:66-84](#)).

- A survey of pragmatic annotation is provided in [Grice et al. \(2000: 54-67\)](#).
- A European project MATE (= Multi-level annotation, tools engineering) has tackled the issue of standardization in developing tools for corpus annotation, and more specifically for dialogue annotation, developing a workbench and an evaluation of various schemes, investigating their applicability across languages (<http://mate.nis.sdu.dk/>).

Other levels of annotation

There is less to say about other levels of annotation mentioned in 2 above, either because they are less challenging or have been less subject to efforts of standardization. Examples particularly worth notice are:

phonetic annotation

SAMPA (devised by [Wells et al 1992](#)) is a convenient way of representing phonetic (IPA) symbols in 7-bit ASCII characters. It can be useful for any parts of spoken transcriptions where pronunciation has to be represented — but it is now giving way to Unicode.

stylistic annotation

Semino and Short ([2003](#)) have developed a detailed annotation scheme for modes of speech and thought representation — one area of considerable interest in stylistics. This has been applied to a varied corpus of literary and non-literary texts.

8. Evaluation of annotation: realism, accuracy and consistency

In section 4 I mentioned that the **quality** or 'goodness' of annotation was one important — though rather unclear — criterion to be sought for in annotation. Reverting to the POS-tagging example once again, we may distinguish two quite different ideas of quality. The first refers to the linguistic **realism** of the categories. It would be possible to invent tags which were easy to apply automatically with 100% accuracy — e.g. by arbitrarily dividing a dictionary into 100 parts and assigning a set of 100 tags to words in the dictionary according to their alphabetical order — but these tags would be useless for any serious linguistic analysis. Hence we have to make sure that our tagset is well designed to bring together in one category words which are likely to have psychological and linguistic affinity, i.e. are similar in terms of the syntactic distribution, their morphological form, and/or their semantic interpretation.

A second, less abstract, notion of quality refers not to the tagset, but to the accuracy and consistency with which it is applied.

Accuracy refers to the percentage of words (i.e. word tokens) in a corpus which are correctly tagged. Allowing for ambiguity in tag assignment, this is sometimes divided into two categories — precision and recall — see [van Halteren \(1999: 81-86\)](#).

Recall is the extent to which all correct annotations are found in the output of the tagger.

Precision is the extent to which incorrect annotations are rejected from the output.

The obvious question to ask here is: what is meant by 'correct'? The answer is: 'correctness' is defined by what the annotation scheme allows or disallows — and this is an added reason why the annotation scheme has to be specific in detail, and has to correspond as closely as possible with linguistic realities recognized as such..

For example, automatic taggers can achieve tagging as high as 98% correct. However, this is not as good as it could be, so the automatic tagging is often followed by a **post-editing** stage in which human analysts correct any mistakes in the automatic tagging, or resolve any ambiguities.

The first question here is: is it possible for hand-editors to achieve 100% accuracy? Most people will find this unlikely, because of the unpredictable peculiarities of language that crop up in a corpus, and because of the failure of even the most detailed annotation schemes to deal with all eventualities. Perhaps between 99% and 99.5% accuracy might be the best that can be achieved, given that unclear and unprecedented cases are bound to arise. Nevertheless, 99.5% accuracy achieved with the help of a human post-editor would still be preferable to 96% or 97% as the result of just automatic tagging. Accuracy is therefore one criterion of quality in POS-tagging, and indeed in any annotation task.

A second question that may be asked is: how consistently has the annotation task been performed? One way to test this in POS tagging is to have two human annotators post-edit the same piece of automatically-tagged text, and to determine in what percentage of cases they agree with one another. The more this consistency measure (called **inter-rater agreement**) approaches 100%, the higher the quality of the annotation. (Accuracy and consistency are obviously related: if both raters achieve 100% accuracy, it is inevitable that they achieve 100% consistency.)

In the early days of POS-tagging evaluation, it was feared that up to 5% of words would be so uncertain in their word class that a high degree of accuracy and of consistency could not be achieved. However, this is too pessimistic: Baker ([1997](#)) and Voutilainen and Järvinen ([1995](#)) have shown how scores not far short of 100% can be attained for both measures.

A more sophisticated measure of inter-rater consistency is the so-called **kappa coefficient** (*K*). Strictly speaking, it is not enough to compare the output of two manual annotators by counting the percentage of cases where they agree or do not agree. This ignores the fact the

even if the raters assigned the tags totally by chance, in a certain proportion of cases would be expected to be in agreement. This factor is built into the kappa coefficient, which is defined as follows:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

"where $P(A)$ is the proportion of time that the coders agree and $P(E)$ is the proportion of times that we would expect them to agree by chance." ([Carletta 1996](#): 4).

There is no doubt that annotation tends to be highly labour-intensive and time-consuming to carry out well. This is why it is appropriate to admit, as a final observation, that 'best practice' in corpus annotation is something we should all strive for — but which perhaps few of us will achieve.

9. Getting down to the practical task of annotation

To conclude, it is useful to say something about the practicalities of corpus annotation. Assume, say, that you have a text or a corpus you want to work on, and want to 'get the tags into the text'.

- It is not necessary to have special software. You can annotate the text using a general-purpose text editor or word processor. But this means the job has to be done by hand, which risks being slow and prone to error.
- For some purposes, particularly if the corpus is large and is to be made available for general use, it is important to have the annotation validated. That is, the vocabulary of annotation is controlled and is allowed to occur only in syntactically valid ways. A validating tool can be written from scratch, or can use macros for word processors or editors.
- If you decide to use XML-compliant annotation, this means that you have the option to make use of the increasingly available XML editors. An XML editor, in conjunction with a DTD or schema, can do the job of enforcing well-formedness or validity without any programming of the software, although a high degree of expertise with XML will come in useful.
- Special tagging software has been developed for large projects — for example the CLAWS tagger and Template Tagger used for the Brown Family or corpora and the BNC. Such programs or packages can be licensed for your own annotation work. (For CLAWS, see the UCREL website <http://www.comp.lancs.ac.uk/ucrel/>.)
- There are tagsets which come with specific software — e.g. the C5, C7 and C8 tagsets for CLAWS, and CHAT for the CHILDES system, which is the de facto standard for language acquisition data.
- There are more general architectures for handling texts, language data and software systems for building and annotation corpora. The most prominent example of this is GATE ('general architecture for text engineering' <http://gate.ac.uk>) developed at the University of Sheffield.

Chapter 3: Metadata for corpus work (Lou Burnard, University of Oxford © Lou Burnard 2004)

1. What is metadata and why do you need it?

Metadata is usually defined as 'data about data'. The word appears only six times in the 100 million word British National Corpus (BNC), in each case as a technical term from the domain of information processing. However, all of the material making up the British National Corpus predates the whole-hearted adoption of this word by the library and information science communities for one very specific kind of data about data: the kind of data that is needed to describe a digital resource in sufficient detail and with sufficient accuracy for some agent to determine whether or not that digital resource is of relevance to a particular enquiry. This so-called discovery metadata has become a major area of concern with the expansion of the World Wide Web and other distributed digital resources, and there have been a number of attempts to define standard sets of metadata for specific subject domains, for specific kinds of activity (for example, digital preservation) and more generally for resource discovery. The most influential of the generic metadata schemes has been the Dublin Core Metadata Initiative (DCMI), which (in the year after the BNC was first published), proposed 15 metadata categories which it was felt would suffice to describe any digital resource well enough for resource discovery purposes. For the linguistics community, more specific and structured proposals include those of the Text Encoding Initiative (TEI), the Open Language Archive Community (OLAC), and the ISLE Metadata Initiative (IMDI).

These and other initiatives have as a common goal the definition of agreed sets of metadata categories which can be applied across many different resources, so that potential users can assess the usefulness of those resources for their own purposes. The theory is that in much the same way that domestic consumers expect to find standardized labelling on their grocery items (net weight in standard units, calorific value per 100 grams, indication of country of origin, etc.), so the user of digital resources will expect to find a standard set of descriptors on their data items. While there can be no doubt that any kind of metadata is better than none, and that some metadata categories are of more general interest than others, it is far less clear on what basis or authority the definition of a standard set of metadata descriptors should proceed. Digital resources, particularly linguistic corpora, are designed to serve many different applications, and their usefulness must thus be evaluated against many different criteria. A corpus designed for use in one context may not be suited to another, even though its description suggests that it will be.

Nevertheless, it is no exaggeration to say that without metadata, corpus linguistics would be virtually impossible. Why? Because corpus linguistics is an empirical science, in which the investigator seeks to identify patterns of linguistic behaviour by inspection and analysis of naturally occurring samples of language. A typical corpus analysis will therefore gather together many examples of linguistic usage, each taken out of the context in which it originally occurred, like a laboratory specimen. Metadata restores and specifies that context, thus enabling us to relate the specimen to its original habitat. Furthermore, since language corpora are constructed from pre-existing pieces of language, questions of accuracy and authenticity are all but inevitable when using them: without metadata, the investigator has no way of answering such questions. Without metadata, the investigator has nothing but disconnected words of unknowable provenance or authenticity.

In many kinds of corpus analysis, the objective is to detect patterns of linguistic behaviour which are common to particular groups of texts. Sometimes, the analyst examines occurrences of particular linguistic phenomena across a broad range of language samples, to see whether certain phenomena are more characteristic of some categories of text than others. Alternatively, the analyst may attempt to characterize the linguistic properties or regularities of a particular pre-defined category of texts. In either case, it is the metadata which defines the category of text; without it, we have no way of distinguishing or grouping the component texts which make up a large heterogeneous corpus, nor even of talking about the properties of a homogeneous one.

2. Scope and representation of metadata

Many different kinds of metadata are of use when working with language corpora. In addition to the simplest descriptive metadata already mentioned, which serves to identify and characterize a corpus regarded as a digital resource like any other, we discuss below the following categories of metadata, which are of particular significance or use in language work:

- editorial metadata, providing information about the relationship between corpus components and their original source ([3. Editorial metadata below](#))
- analytic metadata, providing information about the way in which corpus components have been interpreted and analysed ([4. Analytic metadata below](#))
- descriptive metadata, providing classificatory information derived from internal or external properties of the corpus components ([5. Descriptive metadata below](#))
- administrative metadata, providing documentary information about the corpus itself, such as its title, its availability, its revision status, etc. (this section).

In earlier times, it was customary to provide corpus metadata in a free standing reference manual, if at all. It is now more usual to present all metadata in an integrated form, together

with the corpus itself, often using the same encoding principles or markup language. This greatly facilitates both automatic validation of the accuracy and consistency with which such documentation is provided, and also facilitates the development of more human-readable and informative software access to the contents of a corpus.

A major influence in this respect has been the Text Encoding Initiative (TEI), which in 1994 first published an extensive set of Guidelines for the Encoding of Machine Readable Data (TEI P1). These recommendations have been widely adopted, and form the basis of most current language resource standardization efforts. A key feature of the TEI recommendations was the definition of a specific metadata component known as the TEI Header. This has four major parts, derived originally from the International Standard Bibliographic Description (ISBD), which sought to extend the well-understood principles of print bibliography to the (then!) new world of digital resources:

- a *file description*, identifying the computer file¹ itself and those responsible for its authorship, dissemination or publication etc., together with (in the case of a derived text such as a corpus) similar bibliographic identification for its source;
- an *encoding description*, specifying the kinds of encoding used within the file, for example, what tags have been used, what editorial procedures applied, how the original material was sampled, and so forth;
- a *profile description*, supplying additional descriptive material about the file not covered elsewhere, such as its situational parameters, topic keywords, descriptions of participants in a spoken text etc.
- a *revision description*, listing all modifications made to the file during the course of its development as a distinct object.

The TEI scheme expressed its recommendations initially as an application of the Standard Generalized Markup Language (SGML: ISO 8879). More recently, it has been re-expressed as an application of the current de facto standard language of the internet: the W3C's extensible markup language (XML), information on which is readily available elsewhere.

The scope of this article does not permit exhaustive discussion of all features of the TEI Header likely to be of relevance to corpus builders or users, but some indication of the range of metadata it supports is provided by the summary below. For full information, consult the online version of the TEI Guidelines (<http://www.tei-c.org/Guidelines/HD.html>), or the Corpus Encoding Standard (<http://www.cs.vassar.edu/CES>)².

3. Editorial metadata

Because electronic versions of a non-electronic original are inevitably subject to some form of distortion or translation, it is important to document clearly the editorial procedures and conventions adopted. In creating and tagging corpora, particularly large ones assembled

from many sources, many editorial and encoding compromises are necessary. The kind of detailed text-critical attention possible for a smaller literary text may be inappropriate, whether for methodological or financial reasons. Nevertheless, users of a tagged corpus will not thank the encoder if arbitrary editorial changes have been silently introduced, with no indication of where, or with what regularity. Such corpora can actively mislead the unwary or partially informed user.

A conscientious corpus builder should therefore take care to consider making explicit in the corpus markup at least the following kinds of intervention:

addition or omission

where the encoder has supplied material not present in the source, or (more frequently in corpus work) where material has been omitted from a transcription or encoding.

correction

where the source material is judged erroneous (for example, misprints) but the encoder wishes to preserve the original error, or simply to indicate that it has been corrected.

normalization

where, although not considered erroneous, the source material exhibits a variant form which the encoder wishes to replace by a standardized form, either retaining the original, or silently.

The explicit marking of material missing from an encoded text may be of considerable importance as a means of indicating where non-linguistic (or linguistically intractable) items such as symbols or diagrams or tables have been omitted:

```
<gap desc="diagram"/>
```

Such markup is useful where the effort involved in a more detailed transcription (using more specific elements such as <figure> or <table>, or even detailed markup such as SVG or mathml) is not considered worthwhile. It is also useful where material has been omitted for sampling reasons, so as to alert the user to the dangers of using such partial transcriptions for analysis of text-grammar features:

```
<div type="chapter"> <gap extent="100 sentences" cause="sampling strategy"/>
<s>This is not the first sentence in this chapter.</s>
```

As these examples demonstrate, the tagging of a corpus text encoded in XML is itself a special and powerful form of metadata, instructing the user how to interpret and reliably use the data. As a further example, consider the following hypothetical case. In transcribing a spoken English text, a word that sounds like 'skuzzy' is encountered by a transcriber who does not recognize this as one way of pronouncing the common abbreviation 'SCSI' (small

computer system interface). The transcriber T1 might simply encode his or her uncertainty by a tag such as

```
<unclear extent="two syllables" resp="T1" desc="sounds like skuzzy"/>
```

or even

```
<gap extent="two syllables" cause="unrecognizable word"/>
```

Alternatively, the transcriber might wish to allow for the possibility of "skuzzy" as a lexical item while registering doubts as to its correctness, to propose a "correct" spelling for it, or simply to record that the spelling has been corrected from an unstated deviant form. This range of possibilities might be represented in a number of ways, some of which are shown here:

```
<sic>skuzzy</sic>
```

```
<corr>SCSI</corr>
```

```
<choice> <sic>skuzzy</sic> <corr>SCSI</corr> </choice>
```

The first of these encodings enables the encoder to signal some doubt about the authenticity of the word. The second enables the encoder to signal that the word has been corrected, without bothering to record its original form. The third provides both the dubiously authentic form and its correction, indicating that a choice must be made between them.

This same method might be applied to the treatment of apparent typographic error in printed originals, or (with slightly different tagging since normalization is not generally regarded as the same kind of thing as correction) to the handling of regional or other variant forms. For example, in modern British English, contracted forms such as 'isn't' exhibit considerable regional variation, with forms such as 'isnae', 'int' or 'ain't' being quite orthographically acceptable in certain contexts. An encoder might thus choose any of the following to represent the Scots form 'isnae':

```
<reg>isn't</reg>
```

```
<orig>isnae</orig>
```

```
<choice> <reg>isn't</reg> <orig>isnae</orig> </choice>
```

Which choice amongst these variant encodings will be appropriate is a function of the intentions and policies of the encoder: these, and other aspects of the encoding policy, should be stated explicitly in the corpus documentation, or the appropriate section of the encoding description section of a TEI Header.

4. Analytic metadata

A corpus may consist of nothing but sequences of orthographic words and punctuation, sometime known as *plain text*. But, as we have seen, even deciding on which words make up a text is not entirely unproblematic. Texts have many other features worthy of attention

and analysis. Some of these are structural features such as text, text subdivision, paragraph or utterance divisions, which it is the function of a markup system to make explicit, and concerning which there is generally little controversy. Other features are however (in principle at least) recognizable only by human intelligence, since they result from an understanding of the text.

Corpus-builders do not in general have the leisure to read and manually tag the majority of their materials; detailed distinctions must therefore be made either automatically or not at all (and the markup should make explicit which was the case!). In the simplest case, a corpus builder may be able reliably to encode only the visually salient features of a written text such as its use of italic font or emphasis, or by applying probabilistic rules derived from other surface features such as capitalization or white space usage.

At a later stage, or following the development of suitably intelligent tools, it may be possible to review the elements which have been marked as visually highlighted, and assign a more specific interpretive textual function to them. Examples of the range of textual functions of this kind include quotation, foreign words, linguistic emphasis, mention rather than use, titles, technical terms, glosses, etc.

The performance of such tools as morpho-syntactic taggers may occasionally be improved by pre-identification of these, and of other kinds of textual features which are not normally visually salient, such as names, addresses, dates, measures, etc. It remains debatable whether effort is better spent on improving the ability of such tools to handle arbitrary text, or on improving the performance of pre-tagging tools. Such tagging has other uses however: for example, once names have been recognized, it becomes possible to attach normalized values for their referents to them, thus facilitating development of systems which can link all references to the same individual by different names. This kind of *named entity recognition* is of particular interest in the development of message understanding and other NLP systems.

The process of encoding or tagging a corpus is best regarded as the process of making explicit a set of more or less interpretive judgments about the material of which it is composed. Where the corpus is made up of reasonably well understood material (such as contemporary linguistic usage), it is reasonably easy to distinguish such interpretive judgments from apparently objective assertions about its structural properties, and hence convenient to represent them in a formally distinct way. Where corpora are made up of less well understood materials (for example, in ancient scripts or languages), the distinction between structural and analytic properties becomes less easy to maintain. Just as, in some models of cognition at least, a text triggers meaning but does not embody it, so a text triggers multiple encodings, each of equal formal validity, if not utility.

Linguistic annotation of almost any kind may be attached to components at any level from the whole text to individual words or morphemes. At its simplest, such annotation allows the analyst to distinguish between orthographically similar sequences (for example, whether the word 'Pat' at the beginning of a sentence is a proper name, a verb, or an adjective), and to group orthographically dissimilar ones (such as the negatives 'not' and 'n't'). In the same way, it may be convenient to specify the base or lemmatized version of a word as an alternative for its inflected forms explicitly, (for example to show that 'is', 'was', 'being' etc. are all forms of the same verb), or to regularize variant orthographic forms, (for example, to indicate in a historical text that 'morrow', 'morwe' and 'morrowe' are all forms of the same token). More complex annotation will use similar methods to capture one or more syntactic or morphological analyses, or to represent such matters as the thematic or discourse structure of a text.

Corpus work in general requires a modular approach in which basic text structures are overlaid with a variety of such annotations. These may be conceptualized as operating as a series of layers or levels, or as a complex network of descriptive pointers, and a variety of encoding techniques may be used to express them (for example, XML or RDF schemas, annotation graphs, standoff markup...).

4.1. Categorization

In the TEI and other markup schemes, a corpus component may be categorized in a number of different ways. Its category may be implied by the presence of information in the header associated with the element in question (see further 5. Descriptive metadata). It may be inherited from a parent element occurrence, or explicitly assigned by an appropriate attribute. The latter case is the more widely used, but we begin by discussing some aspects of the former.

If we say that a text is a newspaper or a novel, it is self-evident that journalistic or novelistic properties respectively are inherited by all the components making up that text. In the same way, any structural division of an XML-encoded text can specify a value which is understood to apply to all elements within it. As an example, consider a corpus composed of small ads:

```
<adSection> <s>For sale</s> <ad> <s>Large French chest available... </s> </ad> <ad>
<s>Pair of skis, one careful owner...</s> </ad> </adSection>
```

In this example, the element <s> has been used to enclose all the textual parts of a corpus, irrespective of their function. However, an XML processor is able to distinguish <s> elements appearing in different contexts, and can thus distinguish occurrences of words which appear directly inside an <adSection> (such as "for sale") from those which appear nested within an

<ad> (such as "large French chest"). In this way, the XML markup provides both syntax and semantics for corpus analysis.

Attribute values may be used in the same way, to assert properties for the elements to which they are attached, and for their children. For example:

```
<div type="section" lang="FRA"> <head>Section en française</head> <s id="S1">Cette phrase est en français.</s> <s id="S2">Celle-ci également.</s> </div> <div type="section" lang="ENG"> <head>English Section</head> <s id="S3">This sentence is in English.</s> <s id="S4">As is this one.</s> <s id="S5" lang="FRA">Celle-ci est en français.</s> <s id="S6">This one is not.</s> </div>
```

An XML application can correctly identify which sentences are in which language here, by following an algorithm such as "the language of an <s> element is given by its lang attribute, or (if no lang is specified) by that of the nearest parent element on which it is specified".

As noted above, many linguistic features are inherent to the structure and organization of the text, indeed inseparable from it. A common requirement therefore is to associate an interpretive category with one or more elements at some level of the hierarchy. The most typical use of this style of markup is as a vehicle for representation of linguistic annotation, such as morphosyntactic code or root forms. For example:

```
<s ana="NP"> <w ana="VVD" lemma="analyse">analysed</w> <w ana="NN2" lemma="corpus">corpora</w> </s>
```

XML is, of course, a hierarchic markup language, in which analysis is most conveniently represented as a well-behaved singly-rooted tree. A number of XML techniques have been developed to facilitate the representation of multiple hierarchies, most notably *standoff* markup, in which the categorizing tags are not embedded within the text stream (as in the examples above) but in a distinct data stream, linked to locations within the actual text stream by means of hypertext style pointers. This technique enables multiple independent analyses to be represented, at the expense of some additional complexity in programming.

4.2. Validation of categories

A major advantage of using a formal language such as XML to represent analytic annotation within a text is its support for automatic validation, that is, checking that the categories used conform to a previously defined model of which categories are feasible in which contexts³. Where the categorization is performed by means of specific XML elements, the XML system itself can validate the legality of the tags, using a *schema* or *document type declaration*. Validation of attribute values or element content requires additional processing, for which analytic metadata is particularly important.

As an example, consider the phrase "analysed corpora", which might be tagged as follows:

```
<w ana="VVD">analysed</w> <w ana="NN2">corpora</w>
```

Morpho-syntactic analyses of this kind are relatively commonplace and well understood, so that (in this particular case) the encoder may feel that no further documentation or validation of the codes VVD or NN2 is needed. Suppose however that the encoder in this case wishes to do rather more than simply associate an opaque or undefined code with each <w> element.

As a first step, the encoder may decide to provide a list of all possible analytic codes, giving a gloss to each, as follows:

```
<interp id="VVD" value="past tense adjectival form of lexical verb"/> <interp id="NN2" value="plural form of common noun"/>
```

The availability of a control list of annotations, even a simple one like this, increases the sophistication of the processing that can be carried out with the corpus, supporting both documentation and validation of the codes used. If the analytic metadata is further enhanced to reflect the internal structure of the analytic codes, yet more can be done — for example, one could construct a typology of word class codes along the following lines:

```
<interpGrp id="NN" value="common noun"> <interp id="NN1" value="singular common noun"/> <interp id="NN2" value="plural common noun"/> </interpGrp>
```

The hierarchy could obviously be extended by nesting groups of the same kind. We might for example mark the grouping of common (NN) and proper (NP) nouns in the following way:

```
<interpGrp value="nominal"> <interpGrp id="NN"> <interp id="NN1" value="singular common noun"/> <interp id="NN2" value="plural common noun"/> </interpGrp> <interpGrp id="NP"> <interp id="NP1" value="singular proper noun"/> <interp id="NP2" value="plural proper noun"/> </interpGrp>
```

Alternatively, one could unbundle the linguistic interpretations entirely by regarding them as a set of typed *feature structures*, a popular linguistic formalism which is readily expressed in XML. This approach permits an XML processor automatically to identify linguistic analyses where features such as number or properness are marked, independently of the actual category code (the NN1 or NP2) used to mark the analysis.

5. Descriptive metadata

The social context within which each of the language samples making up a corpus was produced, or received, is arguably at least as significant as any of its intrinsic linguistic properties, if indeed the two can be entirely distinguished. In large mixed corpora such as the BNC, it is of considerably more importance to be able to identify with confidence such information as the mode of production or publication or reception, the type or genre of writing or speech, the socio-economic factors or qualities pertaining to its producers or recipients, and so on. Even in smaller or more narrowly focussed corpora, such variables and a clear

identification of the domain which they are intended to typify are of major importance for comparative work.

At the very least, a corpus text should indicate its provenance, (i.e. the original material from which it derives) with sufficient accuracy that the source can be located and checked against its corpus version. Existing bibliographic descriptions are easily found for conventionally published materials such as books or articles and the same or similar conventions should be applied to other materials. In either case, the goal is simple: to provide enough information for someone to be able to locate an independent copy of the source from which the corpus text derives. Because such works have an existence independent of their inclusion in the corpus, it is possible not only to verify but also to extend their descriptive metadata.

For fugitive or spoken material, where the source may not be so easily identified and is less likely to be preserved independently of the corpus, this is less feasible. It is correspondingly important that the metadata recorded for such materials should be as all inclusive as feasible. When transcribing spoken material, for example, such features as the place and time of recording, the demographic characteristics of speakers and hearers, the social context and setting etc. are of immense value to the analyst, and cannot easily be gathered retrospectively.

Where interpretative categories or descriptive taxonomies have been applied, for example in the definition of text types or genres, these must also be documented and defined if the user is to make full use of the material.

To record the classification of a particular text, one or more of the following methods may be used:

- a list of descriptive keywords, either arbitrary or derived from some specific source, such as a standard bibliography;
- a reference to one or more of internally-defined categories, declared in the same way as other analytic metadata, each defined as unstructured prose, or as a more structured set of *situational parameters*.

Despite its apparent complexity, a classificatory mechanism of this kind has several advantages over the kind of fixed classification schemes implied by simply assigning each text a fixed code, chiefly as regards flexibility and extensibility. As new ways of grouping texts are identified, new codes can be added. Cross classification is built into the system, rather than being an inconvenience. More accurate and better targetted enquiries can be posed, in terms of the markup. Above all, because the classification scheme is expressed in the same way as all the other encoding in the corpus, the same enquiry system can be used for both.

It will rarely be the case that a corpus uses more than one reference or segmentation scheme. However, it will often be the case that a corpus is constructed using more than one editorial policy or sampling procedure and it is almost invariably the case that each corpus text has a different source or particular combination of text-descriptive features or topics.

To cater for this variety, the TEI scheme allows for contextual information to be defined at a number of different levels. Information relating, either to all texts, or potentially to any number of texts within a corpus should be held in the overall corpus header. Information relating either to the whole of a single text, or to potentially any of its subdivisions, should be held in a single text header. Information is typically held in the form of elements whose names end with the letters *decl* (for 'declaration'), and have a specific type. Examples include `<editorialDecl>` for editorial policies, `<classDecl>` for text classification schemes, and so on.

The following rules define how such declarations apply:

- a single declaration appearing only in the corpus header applies to all texts;
- a single declaration appearing only in a text header applies to the whole of that text, and over-rides any declaration of the same type in a corpus header;
- where multiple declarations of the same type are given in a corpus header, individual texts or text components may specify those relevant to them by means of a linking attribute.

As a simple example, here is the outline of a corpus in which editorial policy E1 has been applied to texts T1 and T3, while policy E2 applies only to text T2:

```
<teiCorpus> <teiHeader> ... <editorialDecl id="E1"> ... </editorialDecl> <editorialDecl id="E3"> ... </editorialDecl> ... </teiHeader> <tei.2 id="T1"> <teiHeader> <!" no editorial declaration supplied "> </teiHeader> <text decls="E1"> ... </text> </tei.2> <tei.2 id="T2"> <teiHeader> <editorialDecl id="E2"> ... </editorialDecl> </teiHeader> <text> ... </text> </tei.2> <tei.2 id="T3"> <teiHeader> <!" no editorial declaration supplied "> </teiHeader> <text decls="E1"> ... </text> </tei.2>
```

The same method may be applied at lower levels, with the `decls` attribute being specified on lower level elements within the text, assuming that all the possible declarations are specified within a single header.

A similar method may be used to associate text descriptive information with a given text, (though not with part of a text). Corpus texts are generally selected in order to represent particular classifications, or text types, but the taxonomies from which those classifications come are widely divergent across different corpora.

Finally, we discuss briefly the methods available for the classification of units of a text more finely grained than the complete text. These are of particular importance for transcriptions of spoken language, in which it is often of particular importance to distinguish, for example,

speech of women and men, or speech produced by speakers of different socio-economic groups. Here the key concept is the provision of means by which information about individual speakers can be recorded once for all in the header of the texts they speak. For each speaker, a set of elements defining a range of such variables as age, social class, sex etc. can be defined in a <participant> element. The identifier of the participant is then used as the value for a who attribute supplied on each <u> element enclosing an utterance by the participant concerned. To select utterances by speakers according to specified participant criteria, the equivalent of a relational join between utterance and participant must be performed, using the value of this identifier.

The same method may be applied to select speech within given social contexts or settings, given the existence in the header of a <settingDesc> element defining the various contexts in which speech is recorded, which can be referenced by the decls attribute attached to an element enclosing all speech recorded in a particular setting.

6. Metadata categories for language corpora: a summary

As we have noted, the scope of metadata relevant to corpus work is extensive. In this final section, we present an overview of the kinds of 'data about data' which are regarded as most generally useful.

Multiple *levels* of metadata may be associated with a corpus. For example, some information may relate to the corpus as a whole (for example, its title, the purpose for which it was created, its distributor, etc); other information may relate only to individual components of it (for example, the bibliographic description of an individual source text), or to groups of such components (for example, a taxonomic classification).

In the following lists, we have supplied the TEI/XCES element corresponding with the topic in question. This is not meant to imply that all corpora should conform to TEI/XCES standards, but rather to add precision to the topics addressed.

6.1. Corpus identification

Under this heading we group information that identifies the corpus, and specifies the agencies responsible for its creation and distribution.

- name of corpus (<titleStmnt/title>)
- producer (<titleStmnt/respStmnt>). The agency (individuals, research group, "principle investigator", company, institution etc.) responsible for the intellectual content of the corpus should be specified. This may also include information about any funding body or sponsor involved in producing the corpus.

- distributor (<publicationStmnt>). The agency (individual, research group, company, institution etc) responsible for making copies of the corpus available. The following information should typically be provided:
 - name of agency <publisher, distributor,>
 - contact details (postal address, email, telephone, fax) (<pubPlace>)
 - date first made available by this agency (<date>)
 - any specific identifier (e.g. a URN) used for the published version (<idno>)
 - availability: a note summarizing any restrictions on availability, e.g. where the corpus may not be distributed in some geographic zones, or for some specific purposes, or only under some specific licensing conditions.

If a corpus is made available by more than one agency, this should be indicated, and the information above supplied for at least one of them. If specific licensing conditions apply to the corpus, a copy of the licence or other agreement should also be included.

6.2. Corpus derivation

Under this heading we group information that describes the sources sampled in creating the corpus.

Written language resources may be derived from any of the following:

- books, newspapers, pamphlets etc. originally in printed form;
- unpublished handwritten or 'born-digital' materials;
- web pages or other digitally distributed materials;
- recorded or broadcast speech or video.

A description of each different source used in building a corpus should be supplied. This may take the form of a full TEI <sourceDescription> attached to the relevant corpus component, or it may be supplied in ancillary printed documentation, but its presence is essential. In a language corpus, samples are taken out of their context; the description of their source both restores that context and enables a degree of independent verification that the sample correctly represents the original.

6.2.1. Bibliographic description

For conventionally printed and published material, a standard bibliographic description should be supplied or referenced, using the usual conventions (author, title, publisher, date, ISBN, etc.), and using a standard citation format such as TEI, BibTeX, MLA etc. For other kinds of material, different data is appropriate: for example, in transcripts of spoken data it is customary to supply demographic information about each speaker, and the context in which the speech interaction occurs. Standards defining the range of such information useful in particular research communities should be followed where appropriate.

Language corpora are generally created in order to represent language in use. As such, they often require more detailed description of the persons responsible for the language production they represent than a standard bibliographic description would provide. Demographic descriptions of the participants in a spoken interaction are clearly essential, but even in a work of fiction, it may also be useful to specify such characteristics for the characters represented. In both cases, the 'speech situation' may be described, including such features as the target and actual audience, the domain, mode, etc.

6.2.2. Extent

Information about the size of each sample and of the whole corpus should be provided, typically as a part of the metadata discussed in [6.3.2. Sampling and extent](#).

6.2.3. Languages

The natural language or languages represented in a corpus should be explicitly stated, preferably with reference to existing ISO standard language codes (ISO 639). Where more than one language is represented, their relative proportions should also be stated. For multilingual aligned or parallel corpora, source and target versions of the same language should be distinguished. (<langUsage>)

6.2.4. Classification

As noted earlier, corpora are not haphazard collections of text, but have usually been constructed according to some particular design, often related to some kind of categorization of textual materials. Particularly in the case where corpus components have been chosen with respect to some predefined taxonomy of text types, the classification assigned to each selected text should be formally specified. (The taxonomy itself may also need to be defined, in the same way as any other formal model; see further [6.3.6. Classification \(etc.\) Scheme](#) below).

A classification may take the form of a simple list of descriptive keywords, possibly chosen from some standard controlled vocabulary or ontology. Alternatively, or in addition, it may take the form of a coded value taken from some list of such values, standard or non-standard. For example, the Universal Decimal Classification might be used to characterize topics of a text, or the researcher might make up their own ad hoc classification scheme. In the latter case an associated set of definitions for the classification codes used must be supplied.

6.3. Corpus encoding

Under this heading we group the following descriptive information relating to the way in which the source documents from which the corpus was derived have been processed and managed:

- Project goals and research agenda (<projectDesc>);
- Sampling principles and methods employed (<samplingDecl>);
- Editorial principles and practices (<editorialDecl>);
- XML or SGML tagging used (<tagsDecl>);
- Reference scheme applied (<refsDecl>);
- Classification scheme used (<classDecl>).

6.3.1. Project Goals

Corpora are usually designed according to some specific design criteria, rather than being randomly assembled. The project goals and research agenda associated with the creation of a corpus should therefore be explicitly stated. The persons or agencies directly responsible will already have been mentioned in the corpus identification; the purpose of this section is to provide further background on such matters as the purposes for which the corpus was created, its design goals, its theoretical framework or context, its intended usage, target audience etc. Although such information is of necessity impressionistic and anecdotal, it can be very helpful to the user seeking to determine the potential relevance of the resource to their own needs.

6.3.2. Sampling and extent

Where a corpus has been made (as is usually the case) by selecting parts of pre-existing materials, the sampling practice should be explicitly stated. For example, how large are the samples? what is the relationship between size of sample and size of original? Were all samples taken from the beginning, middle, or end of texts? On what basis were texts selected for sampling? etc.

The corpus metadata should also include unambiguous and verifiable information about the overall size of the corpus, the size of the sources from which it was derived, and the frequency distribution of sample sizes. Size should be expressed in meaningful units, such as orthographically defined words, or characters.

6.3.3. Editorial practice

By editorial principles and practices we mean the practices followed when transforming the original source into digital form. For textual resources, this will typically include such topics as the following, each of which may conveniently be given as a separate paragraph.

correction

how and under what circumstances corrections have been made in the text.

normalization

the extent to which the original source has been regularized or normalized.

segmentation

how has the text has been segmented, for example into sentences, tone-units, graphemic strata, etc.

quotation

what has been done with quotation marks in the original — have they been retained or replaced by entity references, are opening and closing quotes distinguished, etc.?

hyphenation

what has been done with hyphens (especially end-of-line hyphens) in the original — have they been retained, replaced by entity references, etc.?

interpretation

what analytic or interpretive information has been added to the text — only a brief characterization of the scope of such annotation is needed here; a more formal specification for such annotation may be usefully provided elsewhere, however.

There is no requirement that *all* (or any) of the above be formally documented and defined. It is however, very helpful to identify whether or not information is available under each such heading, so that the end user for whom a particular category may or may not be significant can make an informed judgment of the usefulness to them of the corpus.

6.3.4. Markup scheme

Where a resource has been marked up in XML or SGML, or some other formal language, the markup scheme used should be documented in full, unless it is an application of some publicly defined markup vocabulary such as TEI, CES, Docbook, etc. Non XML or SGML markup is not generally recommended.

For XML or SGML corpora not conforming to a publicly available schema, the following should be made available to the user of the corpus:

- a copy in electronic form of a DTD or XML Schema which can be used to validate each resource supplied;
- a document providing definitions for each element used in the DTD or schema (The TEI element definitions may be used as a model, but any equivalent description may be used);
- any additional information needed to correctly process and interpret the markup scheme.

For XML or SGML which does conform to a publicly available scheme, the following information should be supplied:

- name of the scheme and reference to its definition;
- whether the scheme has been customized or modified in any way;
- where modification has been made, a description of the modification or customization made, including any ancillary documentation, DTD fragments, etc.

For schemes permitting user modification or extension (such as the TEI), documentation of the additional or modified elements provided must also be provided.

Finally, for resources in XML or SGML, it is useful to provide a list of the elements actually marked up in the resource, indicating how often each one is used. This can be used to validate the coverage of the category of information marked up within the corpus. Such a list can then be compared with one generated automatically during validation of the corpus in order to confirm integrity of the resource. The TEI `<tagsDecl>` element is useful for this purpose.

6.3.5. Reference Scheme

By *reference scheme* we mean the recommended method used to identify locations within the corpus, for example text identifier plus sentence-number within text, physical line number within file, etc. Reference systems may be explicit, in that the reference to be used for (say) a given sentence is encoded within the text, or implicit, in that, if sentences are numbered sequentially, it is sufficient only to mark where the next sentence begins. Reference systems may depend upon logical characteristics of the text (such as those expressed in the mark up) or physical characteristics of the file in which the text is stored (such as line sequence); clearly the former are to be preferred as they are less fragile.

A corpus may use more than one reference system concurrently, for example it is often convenient to include a referencing system defined in terms of the original source material (such as page number within source text) as well as one defined in terms of the encoded corpus.

6.3.6. Classification (etc.) Scheme

As noted above, a classification scheme may be defined externally (with reference to some preexisting scheme such as bibliographic subject headings) or internally. Where it is defined internally, a structure like the TEI <taxonomy> element may be used to document the meaning and structure of the classifications used.

Exactly the same considerations apply to any other system of analytic annotation. For example in a linguistically annotated corpus, the classification scheme used for morphosyntactic codes or linguistic functions may be defined externally, by reference to some standard scheme such as EAGLES or the ISO Data Category Registry, or internally by means of an explicit set of definitions for the categories employed.

7. Conclusions

Metadata plays a key role in organizing the ways in which a language corpus can be meaningfully processed. It records the interpretive framework within which the components of a corpus were selected and are to be understood. Its scope extends from straightforward labelling and identification of individual items to the detailed representation of complex interpretive data associated with their linguistic components. As such, it is essential to proper use of a language corpus.

Notes

1. In International Standard Bibliographic Description, the term *computer file* is used to refer to any computer-held object, such as a language corpus, or a component of one.
2. [Dunlop 1995](#) and [Burnard 1999](#) describe the use of the TEI Header in the construction of the BNC.
3. Checking that the categories have been correctly applied, i.e. that for example the thing tagged as a 'foo' actually *is* a 'foo', is not in general an automatable process, since it depends on human judgment as noted above.

Chapter 4: Character encoding in corpus construction (Anthony McEnery and Richard Xiao, Lancaster University © Anthony McEnery and Richard Xiao 2004)

1. Introduction

Corpus linguistics has developed, over the past three decades, into a rich paradigm that addresses a great variety of linguistic issues ranging from monolingual research of one language to contrastive and translation studies involving many different languages. Today, while the construction and exploitation of English language corpora still dominate the field of corpus linguistics, corpora of other languages, either monolingual or multilingual, have also become available. These corpora have added notably to the diversity of corpus-based language studies.

Character encoding is rarely an issue for alphabetical languages, like English, which typically still use ASCII characters. For many other languages that use different writing systems (e.g. Chinese), encoding is an important issue if one wants to display the corpus properly or facilitate data interchange, especially when working with multilingual corpora that contain a wide range of writing systems. Language specific encoding systems make data interchange problematic, since it is virtually impossible to display a multilingual document containing texts from different languages using such encoding systems. Such documents constitute a new Tower of Babel which disrupts communication.

In addition to the problem with displaying corpus text or search results in general, an issue which is particularly relevant to corpus building is that the character encoding in a corpus must be consistent if the corpus is to be searched reliably. This is because if the data in a corpus is encoded using different character sets, even though the internal difference is indiscernible to human eyes, a computer will make a distinction, thus leading to unreliable results. In many cases, however, multiple and often competing encoding systems complicate corpus building, providing a real problem. For example, the main difficulty in building a multilingual corpus such as EMILLE is the need to standardize the language data into a single character set (see [Baker, Hardie & McEnery et al 2004](#))¹. The encoding, together with other ancillary data such as markup and annotation schemes, should also be documented clearly. Such documentation must be made available to the users.

A legacy encoding is typically designed to support one writing system, or a group of writing systems that use the same script (see discussion below). In contrast, Unicode is truly multilingual in that it can display characters from a very large number of writing systems.

Unicode enables one to surmount this Tower of Babel by overcoming the inherent deficiencies of various legacy encodings². It has also facilitated the task of corpus building (most notably for multilingual corpora and corpora involving non-Western languages). Hence, a general trend in corpus building is to encode corpora (especially multilingual corpora) using Unicode (e.g. EMILLE).

Corpora encoded in Unicode can also take advantage of the latest Unicode-compliant corpus tools like *Xaira* ([Burnard and Dodd 2003](#)) and *WordSmith Tools* version 4.0 ([Scott 2003](#)). In this chapter, we will consider character encoding from the viewpoint of corpus linguistics rather than programming, which means that the account presented here is less technical and that some of the proposals we make may differ slightly from those that would be ideal for programmers.

This chapter first briefly reviews the history of character encoding. Following from this is a discussion of standard and non-standard native encoding systems, and an evaluation of the efforts to unify these character codes. Then we move on to discuss Unicode as well as various Unicode Transformation Formats (UTFs). As a conclusion, we recommend that Unicode (UTF8, to be precise) be used in corpus construction.

2. Shift in: what is character encoding about?

The need for electronic character encoding first arose when people tried to send messages via telegraph lines using, for example, the Morse code³. The Morse code encodes alphabets and other characters, like major punctuation marks, as dots and dashes, which respectively represent short and long electrical signals. While telegraphs already existed when the Morse code was invented, the earlier telegraph relied on varying voltages sent via a telegraph line to represent various characters. The earlier approach was basically different from the Morse code in that with this former approach the line is always "on" whereas with the latter, the line is sometimes "on" and sometimes "off". The binary "on" and "off" signals are what, at the lowest level, modern computers use (i.e. 0 and 1) to encode characters. As such, the Morse code is considered here as the beginning of character encoding. Note, however, that character encoding in the Morse code is also different from how modern computers encode data. Whilst modern computers use a succession of "on" and "off" signals to present a character, the Morse code uses a succession of "on" impulses (e.g. the sequences of .- / -... / -.-. stand respectively for capital letters A, B and C), which are separated from other sequences by "off" impulses.

A later advance in character encoding is the Baudot code, invented by Frenchman Jean-Maurice-Émile Baudot (1845-1903) for teleprinters in 1874. The Baudot code is a 5-bit character code that uses a succession of "on" and "off" codes as modern computers do (e.g.

00011 without shifting represents capital letter A). As the code can only encode 32 (i.e. 2⁵) characters at one level (or "plane"), Baudot employs a "lock shift scheme" (similar to the SHIFT and CAPS LOCK keys on your computer keyboard) to double the encoding capacity by shifting between two 32-character planes. This lock shift scheme not only enables the Baudot code to handle the upper and lower cases of letters in the Latin alphabet, Arabic numerals and punctuation marks, it also makes it possible to handle control characters, which are important because they provide special characters required in data transmission (e.g. signals for "start of text", "end of text" and "acknowledge") and make it possible for the text to be displayed or printed properly (e.g. special characters for "carriage return" and "line feed"). Baudot made such a great contribution to modern communication technology that the term *Baud rate* (i.e. the number of data signalling events occurring in a second) is quite familiar to many of us.

One drawback of 5-bit Teletype codes such as the Baudot code is that they do not allow random access to a character in a character string because random access requires each unit of data to be complete in itself, which prevents the use of code extension by means of locking shifts. However, random is essential for modern computing technology. In order to achieve this aim, an extra bit is needed. This led to 6-bit character encoding, which was used for a long time. One example of such codes is the Hollerith code, which was invented by American Herman Hollerith (1860-1929) for use with a punch card on a tabulating machine in the U.S. Census Bureau. The Hollerith code could only handle 69 characters, including upper and lower cases of Latin letters, Arabic numerals, punctuation marks and symbols. This is slightly more than what the Baudot code could handle. The Hollerith code was widely used up to the 1960s.

However, the limited encoding capacity of 6-bit character codes was already felt in the 1950s. This led to an effort on the part of telecommunication and computing industries to create a new 7-bit character code. The result of this effort is what we know today as the ASCII (the American Standard Code for Information Interchange) code. The first version of ASCII (known as ASCII-1963), when it was announced in 1963, did not include lower case letters, though there were many unallocated positions. This problem, among others, was resolved in the second version, which was announced in 1967. ASCII-1967, the version many people still know and use today, defines 96 printing characters and 32 control characters. Although ASCII was designed to avoid shifting as used in Baudot code, it does include control characters such as shift in (SI) and shift out (SO). These control characters were used later to extend the 7-bit ASCII code into the 8-bit code that includes 190 printing characters (cf. [Searle 1999](#)).

The ASCII code was adopted by nearly all computer manufacturers and later turned into an international standard (ISO 646) by the International Standard Organization (ISO) in 1972. One exception was IBM, the dominant force in the computing market in the 1960s and 1970s⁴. Either for the sake of backward compatibility or as a marketing strategy, we do not know which for sure, IBM created a 6-bit character code called BCDIC (Binary Coded Decimal Interchange Code) and later extended this code to the 8-bit EBCDIC (Extended Binary Coded Decimal Interchange Code). As EBCDIC is presently only used for data exchange between IBM machines, we will not discuss this scheme further.

The 7-bit ASCII, which can handle 128 (i.e. 2⁷) characters, is sufficient for the encoding of English characters. With the increasing need to exchange data internationally, which usually involves different languages, as well as using accented Latin characters and non-Latin characters, this encoding capacity quickly turned out to be inadequate. As noted above, the extension of the 7-bit ASCII code into the 8-bit code significantly increased its encoding capacity. This increase was important, as it allowed accented characters in European languages to be included in the ASCII code. Following the standardization of the ASCII code and ISO 646, ISO formulated a new standard (ISO 2022) to outline how 7- and 8-bit character codes should be structured and extended so that native characters could be included. This standard was later applied to derive the whole ISO 8859 family of extensions of the 8-bit ASCII/ISO 646 for European languages. ISO 2022 is also the basis for deriving 16-bit (double-byte) character codes used in East Asian countries such as China, Japan and Korea (the so called CJK language community).

3. Legacy encoding: complementary and competing character codes

The first member of the ISO 8859 family, ISO 8859-1 (unofficially known as Latin-1), was formulated in 1987 (and later revised in 1998) for Western European languages such as French, German, Spanish, Italian and the Scandinavian languages, among others. Since then, the 8859 family has extended to 15 members. However, as can be seen in Table 2 (cf. [Gillam 2003: 39-40](#)), these character codes mainly aim at writing systems of European languages.

It is also clear from the table that there is considerable overlap between these standards, especially the many versions of the Latin characters. Each standard simply includes a slightly different collection of characters to optimise the performance of a particular language or group of languages. Apart from the 8859 standards, there also exist ISO 2022-compliant character codes (national variants of ISO 646) for non-European languages, including, for example, Thai (TIS 620), Indian languages (ISCII), Vietnamese (VISCII) and Japanese (JIS

X 0201). In addition, as noted in the previous section, computer manufacturers such as IBM, Microsoft and Apple have also published their own character codes for languages already covered by the 8859 standards. Whilst the members of the 8859 family can be considered as complementary, these manufacturer tailored "code pages" are definitely competing character codes.

ISO-8859-x	Name	Year	Languages covered
1	Latin-1	1987	Western European languages
2	Latin-2	1987	East European languages
3	Latin-3	1988	Southern European languages
4	Latin-4	1988	Northern European languages
5	Latin/Cyrillic	1988	Russian, Bulgarian, Ukrainian, etc.
6	Latin/Arabic	1987	Arabic
7	Latin/Greek	1987	Greek
8	Latin/Hebrew	1988	Hebrew
9	Latin-5	1989	Turkish (Replaces Latin-3)
10	Latin-6		Northern European languages (Unifies Latin-1 and Latin-4)
11	Latin/Thai		Thai
12	Currently unassigned		May be used in future for Indian or Vietnamese
13	Latin-7		Baltic languages (Replaces Latin-4 and supplements Latin-6)
14	Latin-8		Celtic characters
15	Latin-9	1998	Western European languages (Replaces Latin-1 and adds the euro symbol plus a few missing French and Finnish characters)
16	Latin-10		Eastern European languages (Replaces Latin-2 and adds

			the euro symbol plus a few missing Romanian characters
--	--	--	--

Table 2. ISO 8859 standards

The counterparts of the 8859 standards for CJK languages are also wrapped around ISO 2022, including, for example, ISO 2022-JP, ISO-2022-CN and ISO-2022-KR. These standards are basically 7-bit encoding schemes used for email message encoding. Whilst the 7 or 8-bit character codes are generally adequate for English and other European languages, CJK languages typically need 16-bit character codes, as all of these languages use Chinese characters, which may well exceed tens of thousands. The number of Chinese characters in 1994 was 85,000. Most of these characters, however, are only used infrequently. Studies show that 1,000 characters cover 90%, 2,400 characters cover 99%, 3,800 characters cover 99.9%, 5,200 characters cover 99.99%, and 6,600 characters cover 99.999% of written Chinese (cf. [Gillam 2003: 359](#)). Nevertheless, even the lower limit for literacy, 2,400 Chinese characters, considerably exceeds the number of characters in European languages. Unsurprisingly, double-byte (16-bit) encoding is mandatory for East Asian languages. The double byte scheme is also combined with 7 or 8 bit encoding so that Western alphabets are covered as well. Encoding schemes of this kind are called multi-byting schemes.

Character encoding of East Asian languages started in Japan when the Japanese Industrial Standard Committee (JISC) published JIS C 6220 in 1976 (which was later renamed in 1987 as JIS X 0201-1976). JIS C 6220 is an 8-bit character code which does not include any Chinese characters (or *kanji* as the Japanese call them). Shortly after that, in 1978, JISC published the first character code that includes *kanji* (divided into different levels), JIS C 6226-1978, which shifts between the national variant of ISO 646 and the 8-bit character set of level 1 *kanji*. JIS C 6226 was redefined in 1981 (then JIS C 6226-1983) and renamed in 1987 as JIS X 0208-1983. When level 2 *kanji* was added to level 1 in 1990, the standard became JIS X 0208-1990, including 6,355 *kanji* of two levels. Another 5,801 *kanji* were added when a supplementary standard, JIS X 0212-1990, was published in the same year. The publication of JIS X 0213 (7-bit and 8-bit double byte coded extended Kanji sets for information interchange) in 2000 added 5,000 more Chinese characters.

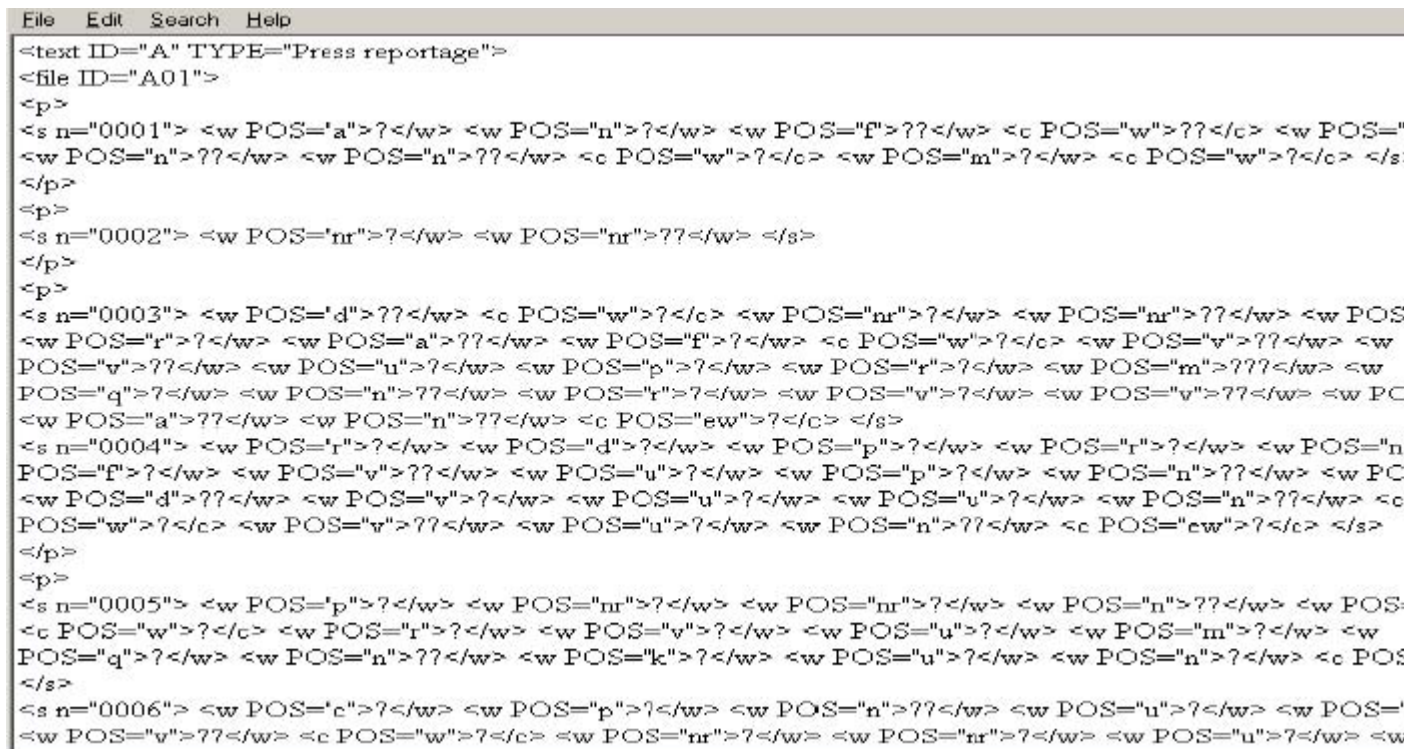
Whilst JIS X 0208/0213 shift between the 7-bit Japanese variant of ISO 646 and the 16-bit character set, the Shift-JIS encoding invented by Microsoft includes both JIS X 0201 (single byte) and JIS X 0208 (double byte), with the single byte character set considered as "halfwidth" while the double byte character set as "full-width".

The character codes in other East Asian countries and regions that use Chinese characters are all based on the JIS model. China published its standard GB 2312 (GB means *guojia*

biaozhun "national standard") in 1981; (South) Korea published KS C 5601 in 1987; Taiwan published CNS 11643 in 1992.

It is also important to mention the EUC (Extended Unix Code) character encoding scheme, which was standardized in 1991 for use on Unix systems. EUC is also based on ISO 2022 and includes the following local variants: EUC-JP for Japan, EUC-CN for China, EUC-TW for Taiwan, and EUC-KR for Korea. In addition, two other character codes have been created to encode Chinese characters. One is the Big5 standard (formulated by five big computer manufacturers), which actually predated and was eventually included in CNS 11643. Big5 is used to encode traditional Chinese (mainly used in Taiwan and Hong Kong). The other is HZ (i.e. *Hanzi* "Chinese character") used for simplified Chinese. Both Big5 and HZ are 7-bit encoding systems.

It is clear from the discussion above that these European or East Asian character encodings are designed to support one writing system, or a group of writing systems that use the same script. These language specific character codes are efficient in handling the writing system(s) for which they are designed. However, with accelerating globalisation and the increasing need for electronic data interchange internationally, these legacy character codes have increasingly become the source of confusion and data corruption, as widely observed (e.g. [Gillam 2003: 52](#)) and experienced by many of us. Have you ever opened a text file that you cannot read, as shown in Figures 1-2? How about the partially unreadable texts as in Figures 3 and 4?



```
File Edit Search Help
<text ID="A" TYPE="Press reportage">
<file ID="A01">
<p>
<s n="0001"> <w POS='a'>?</w> <w POS='n'>?</w> <w POS='f'>??</w> <c POS='w'>??</c> <w POS='
<w POS='n'>??</w> <w POS='n'>??</w> <c POS='w'>?</c> <w POS='m'>?</w> <c POS='w'>?</c> </s>
</p>
<p>
<s n="0002"> <w POS='nr'>?</w> <w POS='nr'>??</w> </s>
</p>
<p>
<s n="0003"> <w POS='d'>??</w> <c POS='w'>?</c> <w POS='nr'>?</w> <w POS='nr'>??</w> <w POS
<w POS='r'>?</w> <w POS='a'>??</w> <w POS='f'>?</w> <c POS='w'>?</c> <w POS='v'>??</w> <w
POS='v'>??</w> <w POS='u'>?</w> <w POS='p'>?</w> <w POS='r'>?</w> <w POS='m'>???</w> <w
POS='q'>?</w> <w POS='n'>??</w> <w POS='r'>?</w> <w POS='v'>?</w> <w POS='v'>??</w> <w PC
<w POS='a'>??</w> <w POS='n'>??</w> <c POS='ew'>?</c> </s>
<s n="0004"> <w POS='r'>?</w> <w POS='d'>?</w> <w POS='p'>?</w> <w POS='r'>?</w> <w POS='n
POS='f'>?</w> <w POS='v'>??</w> <w POS='u'>?</w> <w POS='p'>?</w> <w POS='n'>??</w> <w PC
<w POS='d'>??</w> <w POS='v'>?</w> <w POS='u'>?</w> <w POS='u'>?</w> <w POS='n'>??</w> <c
POS='w'>?</c> <w POS='v'>??</w> <w POS='u'>?</w> <w POS='n'>??</w> <c POS='ew'>?</c> </s>
</p>
<p>
<s n="0005"> <w POS='p'>?</w> <w POS='nr'>?</w> <w POS='nr'>?</w> <w POS='n'>??</w> <w POS
<c POS='w'>?</c> <w POS='r'>?</w> <w POS='v'>?</w> <w POS='u'>?</w> <w POS='m'>?</w> <w
POS='q'>?</w> <w POS='n'>??</w> <w POS='k'>?</w> <w POS='u'>?</w> <w POS='n'>?</w> <c POS
</s>
<s n="0006"> <w POS='c'>?</w> <w POS='p'>?</w> <w POS='n'>??</w> <w POS='u'>?</w> <w POS='
<w POS='v'>??</w> <c POS='w'>?</c> <w POS='nr'>?</w> <w POS='nr'>?</w> <w POS='u'>?</w> <w
```


Figure 1. Chinese characters displayed as question marks

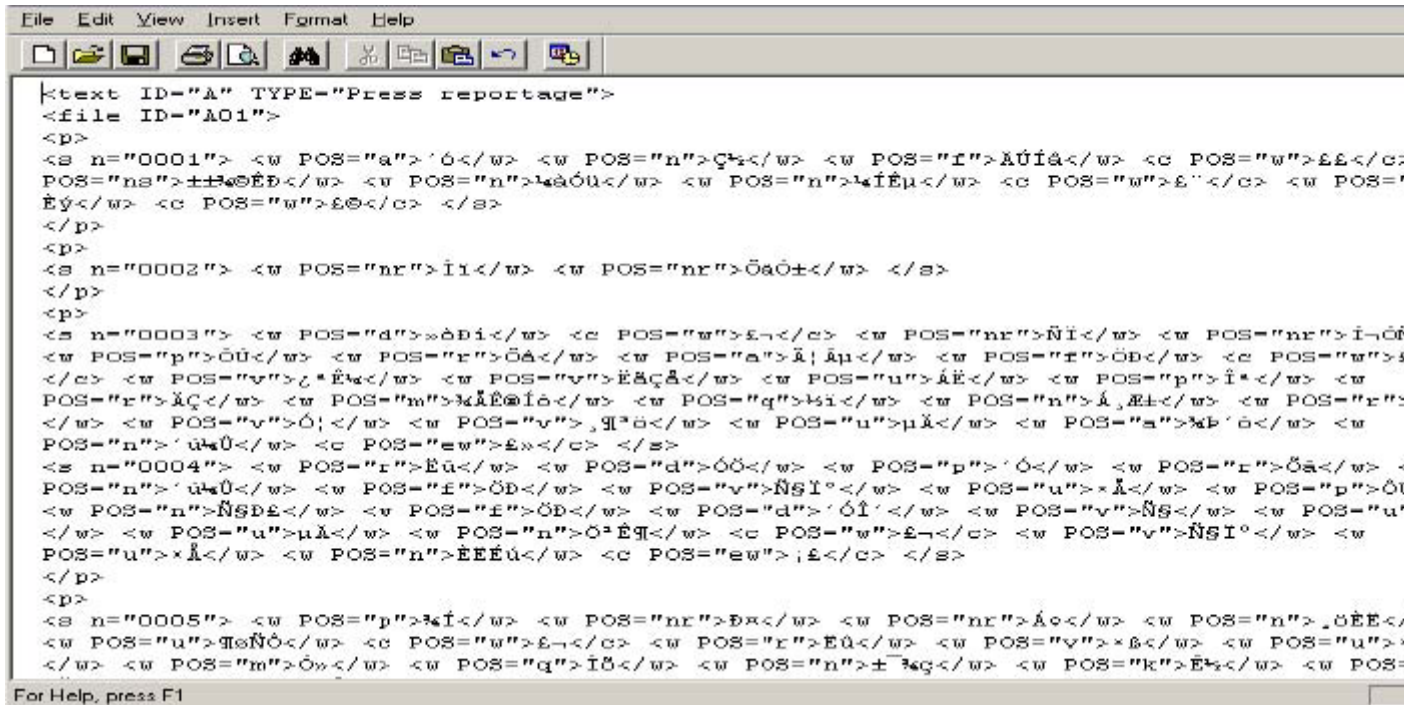


Figure 2. Chinese characters displayed incorrectly

With legacy encodings, each language has its own character set, sometimes even in more than one variant (e.g. GB2312 and HZ). Unsurprisingly, characters in a document encoded using one native character code cannot be displayed correctly with another encoding system, thus causing problems for data exchange between languages. Different operating systems may also encode the same characters in their own ways (e.g. Microsoft Windows vs. Apple Macintosh).

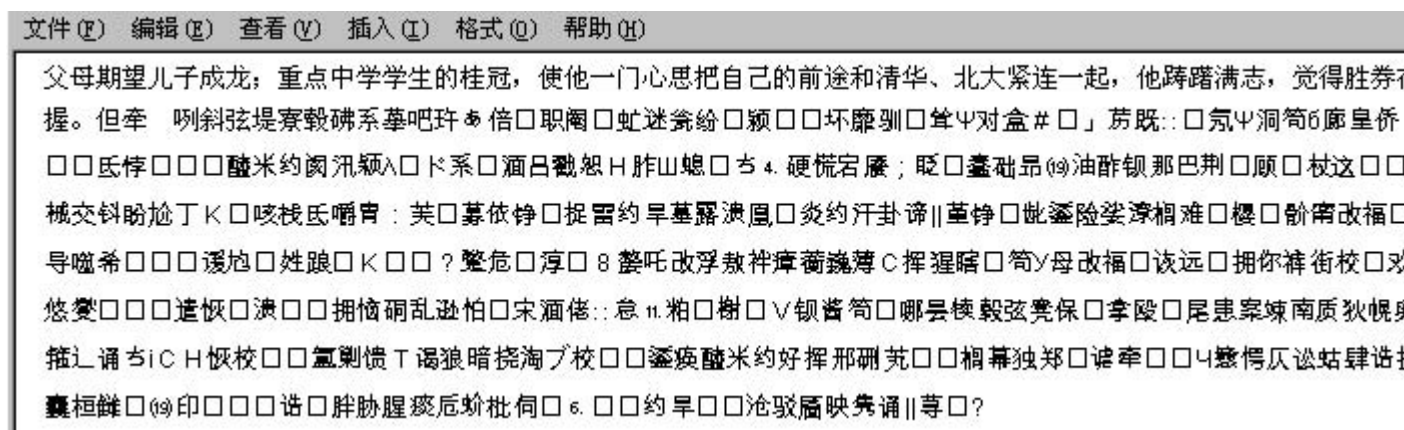


Figure 3. A partially corrupted Chinese paragraph

मुंबई दंगों के अभयिक्तों के पाक में होने के सबूत

fUuk={eg stka çgqhtu (meceytRo) fUu vtm Rm ct; fUu vgtôÉ; mcq; ni rfU 1993 buk bwkcRo buk nw
=kdtuk fUu bwlg yrCgwç; =tW= Rc{trnb, xtRdh bibl ytlh Atuxt NfUej ;:t Rkrzgl YghjtRkm rJbtl yvnhK
fUu vtkatuk yrCgwç; yts Ce vtrfUô;tl buk nik>

सीबीआई प्रवक्ता के अनुसार ठोस सबूतों और लगातार मलि रही सूचनाओं के आधार पर वह कह सकते हैं कदिउद इब्राह
टाईगर मैमन और छोटा शकील अब भी पाकस्तान में है, लेकिन पाकस्तान ने इन्टरपोल के निर्देशों का पालन कर उनकी
गरिफ्तारी के लिए कोई कदम नहीं उठाए है।

ब्यूरो के अनुसार आईसी-814 अपहरण कांड के अभयिक्त भी पाकस्तान में ही कही छपि है, जिन्हें पाकस्तान को गरिफ्त
करना चाहिए। भारत द्वारा पाकस्तान को हाल ही में सौपी गई बीस आतंकवादियों की सूची के बारे में सीबीआई का कहना
इन आतंकवादियों के खिलाफ प्रथम दृष्ट्या साक्ष्य मौजूद है, जो पाकस्तान के लिए उन्हें गरिफ्तार करने के लिए काफी
(वार्ता)

Figure 4. A partly corrupted Hindi text

Even machines using the same operating system may have different regional settings, thus using different character codes. A further problem with legacy encodings is their idiosyncratic fonts⁵. Sometimes even when the regional settings are correct, a text still cannot be displayed correctly without an appropriate font. In a word, legacy encodings, while they handle particular language(s) efficiently, constitute a Tower of Babel that disrupts international communication. As such, Herculean efforts have been made to unify these mutually incompatible character codes with the aim of creating a unified, global standard of character code.

4. Globalisation: efforts to unify character codes

Efforts to unify character codes started in the first half of the 1980s, which unsurprisingly coincides with the beginning of the Internet. Due to a number of technical, commercial and political factors, however, these efforts were pursued by three independent groups from the US, Europe and Japan. In 1984, a working group (known as WG2 today) was set up under the auspices of ISO and International Electrotechnical Commission (IEC) to work on an international standard which has come to be known as ISO/IEC 10646. In the same year, a research project named TRON was launched in Japan, which proposed a multilingual character set and processing scheme. A similar group was established by American computer manufacturers in 1988, which is known today as the Unicode Consortium.

The TRON multilingual character set, which uses escape sequences to switch between 8 and 16 bit character sets, is designed to be "limitlessly extensible" with the aim of including all scripts used in the world ([Searle 1999](#))⁶. However, as this multilingual character set

appears to favour CJK languages more than Western languages, and because US software producers, who are expected to dominate the operating system market in the unforeseeable future, do not support it, it is hard to imagine that the TRON multilingual character set will win widespread popularity except in East Asian countries.

ISO aimed at creating a 32-bit universal character set (UCS) that could hold space for as many as 4,294,967,296 characters, which is large enough to include all characters in modern writing systems in the world. The new standard, ISO/IEC 10646, is clearly related to the earlier ISO 646 standard discussed above. The original version of the standard (ISO/IEC DIS 10646 Version 1), nevertheless, has some drawbacks (see [Gillam 2003: 53](#) for details). It was thus revised and renamed as ISO/IEC 10646 Version 2, which is now known as ISO/IEC 10646-1: 1993. The new version supports both 32-bit (4 octets, thus called UCS-4) and 16-bit forms (2 octets, thus called UCS-2).

The term Unicode (Unification Code) was first used in a paper by Joe Becker from Xerox. The Unicode Standard has also built on Xerox's XCCS universal character set. Unicode was originally designed as a fixed length code, using 16 bits (2 bytes) for each character. It allows space for up to 65,536 characters. In Unicode, characters with the same "absolute shape" — where differences are attributable to typeface design — are "unified" so that more characters can be covered in this space (see [Gillam 2003: 365](#)). In addition to this native 16-bit transformation format (UTF-16), two other transformation formats have been devised to permit transmission of Unicode over byte-oriented 8-bit (UTF-8) and 7-bit (UTF-7) channels (see the next section for a discussion of various UTFs)[7](#). In addition, Unicode has also devised a counterpart to UCS-4, namely UTF-32.

From 1991 onwards, the efforts of ISO 10646 and Unicode were merged, enabling the two to synchronize their character repertoires and the code points these characters are assigned to[8](#). Whilst the two standards are still kept separate, great efforts have also been made to keep the two in synchronization. As such, despite some superficial differences (see [Gillam 2003: 56](#) for details), there is a direct mapping, starting from The Unicode Standard version 1.1 onwards, between Unicode and ISO 10646-1. Although UTF-32 and UCS-4 did not refer to the same thing in the past, they are practically identical today. While Unicode UTF-16 is slightly different from UCS-2, UTF-16 is actually UCS-2 plus the surrogate mechanism (see the next section for a discussion of the surrogate mechanism).

Unicode aims to be usable on all platforms, regardless of manufacturer, vendor, software or locale. In addition to facilitating electronic data interchange between different computer systems in different countries, Unicode has also enabled a single document to contain texts

from different writing systems, which was nearly impossible with native character codes⁹. Unicode make a truly multilingual document possible¹⁰.

Today, Unicode has published the 4th version of its standard. Backed up by the monopolistic position of Microsoft in the operating system market, Unicode appears to be "the strongest link". The current state of affairs suggests that Unicode has effectively "swallowed" ISO 10646. As long as Microsoft dominates the operating system market, it can be predicted that where there is Windows (Windows NT/2000 or later version), there will be Unicode. Consequently, we would recommend that all researchers engaged in electronic text collection development use Unicode.

5. Unicode Transformation Formats (UTFs)

Having decided that one should use Unicode in corpus construction, we need to address yet another important question — what transformation format should be used? Unicode not only defines the identity of each character and its numeric value (code point), it also formulates how this value is represented in bits when the character is stored in a computer file or transmitted over a network connection. Formulations of this kind are referred to as Unicode Transformation Formats, abbreviated as UTFs. For example, with UTF-16, every Unicode character is represented by the 16-bit value of its Unicode number while with UTF-8, Unicode characters are represented by a stream of bytes. The Unicode Standard provides, in chronological order, three UTFs — UTF-16, UTF-8 and UTF-32¹¹. They encode the same common character repertoire and can be efficiently transformed into one another without loss of data. The Unicode Standard suggests that these different encoding forms are useful in different environments and recommends a "common strategy" to use UTF-16 or UTF-8 for internal string storage, but to use UTF-32 for individual character data types. As far as corpus construction is concerned, however, UTF-8 is superior to the other two, as we will see shortly.

As noted previously, Unicode was originally designed as a 16-bit fixed length standard. UTF-16 is the native transformation format of Unicode. As such, in Microsoft applications, UTF-16 is known simply as "Unicode", while UTF-8 is known as "Unicode (UTF-8)". The 16-bit encoding form uses 2 bytes for each code point on the BMP (Basic Multilingual Plane)¹², regardless of position. Shortly after the Unicode Standard came into being, it became apparent that the encoding space allowed by the 16-bit form (65,536 positions) was inadequate. In the Unicode Standard Version 2, therefore, the "surrogate mechanism" was invented, which reserved 2,048 positions in the encoding space and divided these positions into two levels: high and low surrogates, with each allocated 1,024 positions. A high surrogate is always paired with a low surrogate. Whilst unpaired surrogates are meaningless,

different combinations (pairings) of high and low surrogates enable considerably more characters to be represented (usually infrequently used characters are encoded using pairs of 16-bit code points whereas frequently used characters are encoded with a single unit point). As a high surrogate is unmistakably the first byte, and similarly, a low surrogate can only be the second byte of a double-byte character, UTF-16 is able to overcome the deficiencies of variable length encoding schemes. A missing high or low surrogate can only corrupt a single character unlike, for example, the legacy encoding systems for Chinese characters, where such errors typically turn large segments of text into rubbish (see [Figure 3](#)).

UTF-32 is something of a novelty designed as a counterpart to UCS-4 to keep the two standards in synchronization. Unlike UTF-16, which encodes infrequently used characters via pairs of unit points, UTF-32 uses a single code point for each character, thus making data more compact. Nevertheless, this advantage is immediately traded off, as UTF-32 devours memory and disk space.

An important concept specifically related to Unicode-16/32 is byte order. Computers handle data on the basis of 8-bit units, known as octets. Each memory location occupies an octet, or 8 bits. A 16-bit Unicode character takes up 2 memory locations while a 32-bit character occupies 4 memory locations. The distribution of a 16/32-bit character across the 2 or 4 memory locations may vary from one computer to another. Some machines may write the most significant byte into the lowest numbered memory location (called *big-endian*, or UTF-16/ 32BE) whereas others may write the most significant byte into the highest numbered memory location (*little-endian*, or UTF-16/32LE). This is hardly an issue for data stored in computer memory, as the same processor always handles the distribution of a character consistently. When the data is shared between computers with different machine architectures via storage devices or a network, however, this may cause confusion. Unicode does provide mechanisms to indicate the endian-ness of a data file, either by explicating it as UTF-16/32BE/UTF-16/ 32LE, or using a byte order marker (BOM). The default value is big-endian. Even with a BOM, however, confusion may sometimes arise as earlier versions of the Unicode Standard define a BOM differently from version 3.2 and later. As noted earlier in this section, UTF-16 also involves surrogates. As such UTF-16 and UTF-32 are more complex architecturally than UTF-8.

While UTF-32 is wasteful of memory and disk space for all languages, UTF-16 also doubles the size of a file containing single-byte characters (such as English), though for CJK languages that have already used 2-byte encodings traditionally, the file size remains more or less the same.

In addition to the architectural complexity and the waste of storage capacity, a more important point to note regarding UTF-16/32 is that they are not backward compatible, i.e. data encoded with UTF-16/32 cannot be easily used with existing software without extensive rewriting (just imagine the extra workload involved in rewriting *Sara* into *Xaira* and updating WordSmith version 3 to version 4, such rewrites are not trivial). As noted previously, backward compatibility was powerful enough to force IBM to create EBCDIC in parallel to ASCII. Even in its early life, Unicode realised that it was important to have an encoding system which is backward compatible with ASCII. That is why UTF-8 came into being.

UTF-8 is 100% backward compatible with ASCII. It transforms all Unicode characters into a variable length encoding of bytes. UTF-8 encodes the single-byte ASCII characters using the same byte values as ASCII. Other characters on the Basic Multilingual Plane (BMP) are encoded with 1-3 bytes while all non-BMP characters take up 4 bytes. Like UTF-16, UTF-8 is also able to overcome the defects of legacy multi-byting encoding systems by stipulating the specific positions a range of values can take in a character (cf. [Gillam 2003: 198](#)). As such, a Unicode text using UTF-8 can be handled efficiently as any other 8-bit text. UTF-8 is the universal format for data exchange in Unicode, removing all of the inconveniences of Unicode in the sense that it is backward compatible with existing software while at the same time it enables existing programs to take advantage of a universal character set. UTF-8 is also a recommended way of representing ISO/IEC 10646 characters for UCS-2/4 because it is easy to convert from and into UCS. As such, UTF-8 will always be with us and is likely to remain the most popular way of exchanging Unicode data between entities in a heterogeneous environment (cf. [Gillam 2003: 204](#)).

Returning to the issue of efficiency and storage space, it is clear from the above that UTF-8 handles ASCII text as efficiently as ASCII, and because of its feature of backward compatibility, the extra workload required to rewrite software can be saved. Note, however, that UTF-8 is not necessarily a way to save storage space for some writing systems. For example, accented characters take only 1 byte in the ISO 8859 standards whereas they occupy 2 bytes in UTF-8. Legacy encoding systems encode a Chinese character with 2 bytes while UTF-8 uses 3 bytes. However, it can be sensibly argued that a compromise has to be made if one is to have a truly multilingual character code like Unicode. UTF-8 is, we believe, just such a sensible compromise.

6. *Shift out: conclusions and recommendations*

This chapter is concerned with character encoding in corpus construction. It was noted that appropriate and consistent character encoding is important not only for displaying corpus text and search results, it is also for corpus exploration. We first reviewed character encoding in a

historical context, from the Morse code to ASCII. Following from this we introduced various legacy encodings, focusing on the ISO 2022-compliant ISO 8859 standards for European languages and the native character codes for CJK languages. These encoding systems are either complementary to or competing with each other. It was found that while native character codes are efficient in handling the language(s) they are designed for, they are actually inadequate for the purpose of electronic data interchange in a steadily globalising environment. This led to an evaluation of the efforts to create a unified multilingual character code, which concluded that Unicode is the best solution. Following from this we reviewed three UTFs, on the basis of which we recommended UTF-8 as a universal format for data exchange in Unicode, and for corpus construction so as to avoid the textual Tower of Babel.

Notes

1. See the corpus website <http://www.ling.lancs.ac.uk/corplang/emille> for more details of the EMILLE corpus.
2. *Legacy encoding* is used here interchangeably with language specific, or native character code.
3. The Morse code was invented by American Samuel Finley Breese Morse (1791-1872).
4. IBM is an acronym for International Business Machines, which was established on the basis of a company formed, in 1896, by Herman Hollerith after his success.
5. A font is an ordered collection of character glyphs that provides a graphical representation of characters in a character set.
6. In character encoding, an escape sequence is a sequence of more than one code point representing a control function. Escape sequences are used to switch different areas in the encoding space between the various sets of printing characters. They are so called because the ASCII ESC character was traditionally used as the first character of an escape sequence.
7. A communication is said to be *byte-oriented* when the transmitted information is grouped into full bytes rather than single bits (i.e. *bit-oriented*), as in data exchange between disks or over the Internet.
8. *Code point*, or *encoded value*, is the numeric representation of a character in a character set. For example, the code point of capital letter A is 0x41.
9. Whilst it is true that English and Chinese texts, for example, can be merged in a single document with a Chinese encoding system, some English characters may not be displayed correctly. For example, the pound symbol, together with the first numeral following it, is displayed as a question mark.

[10](#). See Norman Goundry's article posted at the Hastings Research website and Ken Whistler's comments posted to Slashdot for arguments for and against Unicode (see the [Bibliography](#)).

[11](#). You might have come across the term UTF-7. This encoding form is specifically designed for use in 7-bit ACSII environments (notably for encoding email messages) that cannot handle 8-bit characters. UTF-7 has never become part of the Unicode Standard.

[12](#). The Unicode encoding space is composed of different layers technically referred to as *planes*. The Basic Multilingual Plane (BMP) is the official name of Plane 0, "the heart and soul of Unicode" ([Gillam 2003](#)), which contains the majority of the encoded characters from most of the modern writing systems (with the exception of the Han ideographs used in Chinese, Japanese and Korea).

Chapter 5: Spoken language corpora (Paul Thompson, University of Reading © Paul Thompson 2004)

1. Introduction

In this chapter I will look at some of the issues involved in developing a corpus of spoken language data. 'Spoken language' is here taken to mean any language whose original presentation was in oral form, and although spoken language data can include recordings of scripted speech, I will focus mainly on the use of recordings of naturally occurring spoken language. The chapter is merely a brief introduction to a highly complex subject; for more detailed treatments, see the collection of papers in [Leech, Myers and Thomas \(1995\)](#).

Spoken language data are notoriously difficult to work with. Written language data are typically composed of orthographic words, which can easily be stored in electronic text files. As other papers in this collection show, there may be problems in representing within a corpus the original presentational features of the text, such as layout, font size, indentations, accompanying diagrams, and so on, but the problem is primarily one of describing what can be seen. In the case of spoken language data, however, the primary problem is one of representing in orthographic or other symbolic means, for reading on paper or screen, what can be heard, typically, in a recording of a speech event in the past. Whereas the words in a written language corpus have an orthographic existence prior to the corpus, the words that appear in an orthographic transcription of a speech event constitute only a partial representation of the original speech event. To supplement this record of the event, the analyst can capture other features, by making either a prosodic or phonetic transcription, and can also record contextual features. However, as Cook ([1995](#)) convincingly argues in his discussion of theoretical issues involved in transcription, the record remains inevitably partial.

Not only the transcription but also the process of data capture itself is problematic: an audio recording of a speech event is only an incomplete view of what occurred, not only because of possible technical deficiencies, but also because visual and tactile features are lost. To compensate for this, video can also be used, but a video recording also presents a view of the event that in most cases cannot capture the views of the participants in the event themselves.

Bearing in mind, then, the complexities of working with spoken language data, the corpus developer needs to approach the task of compiling a spoken language corpus with some circumspection, and design the project carefully. Clearly much will depend upon the purposes for which the corpus is being developed. For a linguist whose interest is in the

patterning of language and in lexical frequency over large quantities of data, there will be little need for sophisticated transcription, and the main consideration will be the quantity and speed of transcription work. Sinclair (1995) advocates simple orthographic transcriptions without indication even of speaker identity, in order to produce large amounts of transcripts. The phonetician, on the other hand, requires less data, but a high degree of accuracy and detail in the phonetic transcription of recordings, with links, where possible, to the sound files. For a discourse analyst, richly detailed information on the contextual features of the original events will be needed. Underlying the development of the corpus, therefore, will be tensions between the need for breadth and the levels of detail that are possible given the resources available. An excess of detail can make the transcripts less readable, but a parsimonious determination in advance of what level of detail is required also runs the danger of removing from the data information that has potential value to the analyst at a later stage.

The amount of documentation compiled (such as explanation of the coding schemes used, records of data collection procedures, and so on) will also depend on whether or not the resources are to be made available to the public, for example, or whether the corpus has to be deposited with the funding body that is sponsoring the research project.

Leech, Myers and Thomas (1995) describe five stages in the development and exploitation of what they term 'computer corpora of spoken discourse':

1	Recording
2	Transcription
3	Representation (mark-up)
4	Coding (or annotation)
5	Application

Table 3: Stages in the development of spoken corpora (after Leech, Myers and Thompson)

This provides a useful framework for our discussion of the issues involved in developing a spoken language corpus, although we will change the headings for two stages and we will also collapse two stages (3 and 4) into one.

For the first stage, it is necessary to discuss both the technicalities of audio/video recording, and also the collection of contextual information, and of the consent of participants; consequently, we will call this section 'Data collection'. Following the collection of spoken language data, the transcription process begins. The third stage, 'Representation', involves the computerization of the transcription, which makes it machine-readable. Consequent to

this, the analyst may wish to add further information to the original transcription, such as classification of the speech acts in the data, or ascription of each word to a grammatical class, and this stage is referred to as 'Annotation'. For brevity's sake, the two stages are treated together here, under the heading of 'Markup and annotation'. In the final stage, which will be headed 'Access', the emphasis will be on access to the corpus, rather than on application, as it is not possible to discuss the range of possible approaches to application, but it is important to think of whether or not the corpus will be made available to other researchers, and in what form it can be made available if so desired.

The headings for the following discussion, therefore, will be:

1	Data collection
2	Transcription
3	Markup and annotation
4	Access

2. Data collection

Before gathering the data, it is important to ensure that you receive informed consent from those who feature clearly either in the transcripts or the video recordings. This can sometimes compromise the purpose of the data collection in research projects that investigate spontaneously occurring speech events, since participants may behave differently if aware that they are being recorded; the BAAL *Recommendations on Good Practice in Applied Linguistics* (<http://www.baal.org.uk/goodprac.htm#6>) gives useful guidance on the researcher's responsibilities to informants. Typically, proof of consent is kept on paper, but in some cases it can be kept on the original recording. If a university seminar is being recorded, for example, it may be easier to film a member of the research team asking all participants for consent rather than to ask all the participants to sign forms.

The development of audio recording technology has had a profound effect on linguistics, as it has made possible the capture of previously ephemeral language events. A criticism of early collections of data, however, was that the recording quality was often not good and it was therefore difficult for transcribers to hear the words clearly. Where high quality data were required, studio recordings were used. But it should be noted that it is often difficult to capture the more spontaneous types of speech event in the studio.

Technological advances mean that there are more options available now. The EAGLES recommendations for spoken texts suggest the use of headphone microphones for best

quality recording, and this would suit data capture under controlled conditions. For recording of naturalistic data, an alternative is to use flat microphones, which are far less obtrusive. As recording devices become smaller, it is possible also to wire up each participant in an event; Perez-Parent ([2002](#)) placed a minidisk (MD) recorder and lapel microphone on each child in the recording of primary school pupils in the Literacy Hour and then mixed the six channels to produce a high quality reproduction of the audio signals. She also recorded the event on video (with a single video camera) and later aligned the orthographic transcript with the six audio channels and the video.

The EAGLES recommendations ([Gibbon, Moore and Winski 1998](#)) also propose that digital recording devices be used, as analogue speech recordings tend to degrade, and are not as easy to access when they need to be studied. Digital recordings can be copied easily on to computers, and backed up on to CDs or DVDs, with minimal loss of quality. Interestingly, they recommend the use of DAT tapes for data capture. The document was published in 1996, and DAT may have been the best choice at the time, but there are several other options available now, including the use of MD technology. This illustrates the difficulty of making recommendations about the best technology to employ — advances in technology make it impossible to give advice that will remain up-to-date. Rather than make recommendations about which data capture technology to use, therefore, I suggest that you seek advice from technical staff, especially sound engineers, and search the Internet for guidance.

With the development of cheaper video cameras and of technology for the digitization of video, the use of video in data capture is becoming more common, and the possibilities for including the video data in a corpus are increasing. Before using a video, however, a number of questions need to be posed:

- Firstly, what relation will the video data have to the transcripts and to the corpus? Will the video data act simply as an aid to the transcriber, providing an extra source of information, or will the video contain information that could not otherwise be captured?
- Secondly, whose perspective is represented by the video camera angle: that of the observer or that of the participants? Cook ([1995](#)) points out that video angles tend to give the view primarily of the observer. If the aim is to gain the perspectives of the participants, how should the camera(s) be positioned?
- Thirdly, will the transcript be aligned to the video? Will the transcript include coding of the gestural and other non-linguistic features? Such coding would make the video machine-readable.

As indicated above, design issues are subject to tensions between the desire for fuller representation of the event and the threat of excessive quantities of data, and of excessive amounts of time and work.

In addition to the recording of the event, a certain amount of background and circumstantial information will be needed as well. In advance of the recording work, it is recommended that procedures be set up for the collection of this information, especially in cases where recordings are to be made by people other than the main team of researchers. The BNC, for example, contains recordings made by speaker participants themselves, of conversations with friends and colleagues, and these people had to record speaker information and consent on forms supplied to them by the project. If the event is to be recorded in audio only, the observer(s) will need to make notes on the event that could assist the transcriber, and which could help to explicate the interactions recorded. Finally, detailed notes should also be kept at the recording stage about the equipment used, the conditions, and about any technical problems encountered, as this information could be of relevance at a later stage, for example, in classifying recordings by quality. It is important, too, to determine in advance what information is required and to make notes under various headings, following a template, to ensure that there is a consistency in type and degree of detail of information for each recording.

3. Transcription

By placing transcription after the section on data collection, I do not want to suggest that each section is neatly compartmentalized and that you do not need to consider transcription issues until after the data have been collected. In the planning stages, it will usually be necessary to decide what features of speech are to be focused on, and therefore what aspects of the speech event need to be captured, and to what level of detail. However, it is convenient to deal with transcription separately.

Firstly, let us consider the *design* of a transcription system. Edwards ([1993](#)) states three principles:

1. Categories should be:
 - systematically discriminable
 - exhaustive
 - systematically contrastive
2. Transcripts should be readable (to the researcher)
3. For computational tractability, mark-up should be
 - systematic
 - predictable.

These three principles refer to the creation of categories, and to questions of readability, both for the human researcher and for the computer. This last point, that of machine readability, will be taken up in the next section.

At the first level, a decision must be made as to whether the transcription is to be *orthographic*, *prosodic*, or *phonetic*, or more than one of these. If a combination is to be used, this means that two, possibly three levels of transcriptions must be aligned somehow. This can be done, for example, by placing the levels of transcription on different lines, or in different columns. Either of these options will have implications for mark-up of the data (see the following section below).

For an orthographic transcription, decisions will have to be taken over spelling conventions. The easiest solution to this problem is to choose a major published dictionary and follow the spelling conventions specified there. This will at least provide guidance on standard orthographic words, but there will be several features of spoken language that are not clearly dealt with, and decisions must be taken over how best to represent them in orthographic form. How, for example, to represent a part of an utterance that sounds like 'gonna'? Should this be standardized to 'going to'? Would standardization present an accurate representation of the language of the speaker? If, on the other hand, a decision is taken to use 'gonna' in some cases, and 'going to' in others, what criteria are to be employed by the transcriber for distinguishing one case from the other (this is what Edwards points to in the expression 'systematically discriminable' above)? The more people there are transcribing the data, the more important it is to provide explicit statements on the procedures to be followed.

Reference books may also not provide all the information that is needed. Where words from languages other than the main language(s) of the speakers appear, what spelling is to be used? This is particularly a problem for languages that employ different orthographic systems, such as Japanese. The Japanese city of 広島 is usually written in English as Hiroshima (using the Hepburn romanization system), but can also be written as Hirosima, following *kunreisiki* (also known as Kunrei-shiki) romanization. According to a report on different systems of romanization conducted by the United Nations Group of Experts on Geographical Names (UNGEGN, <http://www.eki.ee/wgrs/>), the latter is the official system, but the former is most often used in cartography. There is no right or wrong choice, but a decision must be made which can then be set down in writing, so that all transcribers adopt the same conventions, which in turn will lead to consistency in the transcription process.

Decisions will also need to be taken over how to represent non-verbal data, such as contextual information, paralinguistic features, gaps in the transcript, pauses, and overlaps. Let us take the example of pauses. One reason why pauses are important in spoken

language data is that they indicate something of the temporal nature of spoken language and it is this temporality that distinguishes spoken language from written language. Pauses can be classified as 'short' or 'long', but this begs the questions of where the dividing line between the two lies, as well as where the distinction between a short pause and 'not a pause' can be drawn, and to whom the pause appears to be either 'long' or 'short'. Typically, in transcripts for linguistic analysis, a short pause can range in length from less than 0.2 seconds to less than 0.5 seconds, depending on whether the researcher is interested in turn-taking or in information packaging (Edwards 1993: 24). To avoid the problem of terming a pause either 'short' or 'long', the exact length of the pause can be indicated, but strict measurement of pauses could be highly time-consuming and does not necessarily help the analyst to assess the quality of the pause relative to the speech rate of a speaker, or the perceptions of listeners.

Du Bois et al 1990	Short pause	Longer pause	Timed pause
 (1.5)
MacWhinney (1991)	Short pause	Longer pauses	Timed pause
	#	##, ###, #long	#1_5
Rosta (1990)	Short pause	Long pause	
	<,>	<,,>	
Svartvik and Quirk (1980)	Brief pause	Unit pause	Longer pauses
	.	—	—· —·

Figure 5. Four approaches to the symbolic representation of pauses and pause length, adapted from Johansson (1995)

Apart from the question of how a pause is to be classified, there is also the issue of determining a set of symbolic representations of pauses for use in the transcript. Johansson (1995) gives the following examples of sets of conventions used by four different researchers:¹

Edwards (1993:11) raises the issue of how speaker turns can be represented spatially in transcription in the Figure below. These can be portrayed in contrasting systems of spatial arrangement of speakers' turns, as shown below. In the first of these, the vertical arrangement, each speaker's turn appears in sequence below the previous turn, and this, Edwards suggests, implies parity of engagement and influence. The columnar representation, on the other hand, helps to highlight asymmetries in relationships between

the participants, although it is difficult to represent conversations with more than two speakers.

Vertical

A: Did you just get [back]? B: [Yes] or rather 2 hours ago. It was a great film. A: Really?

Column

Speaker A

Did you just get [back]?

Speaker B

[Yes] or rather 2 hours ago. It was a great film.

Really?

Figure 6: spatial arrangement of turns, in vertical and column formats, after Edwards 1993.

In summary, then, a number of transcription conventions need to be established, guided by the principles that Edwards has described (above). Throughout the transcription process, as is also case for the data collection stage, it is important that records are kept carefully so that discrepancies can be dealt with. This is especially true in cases where there are a number of transcribers working on the same project. Where the categories established and the codes adopted consist of non-finite sets, it is advisable to set up some form of easily accessible database or web-based list that all members of the team can add to, where they can add new entries, with commentary, as new set members appear.

The key word is *consistency*. While the human reader may easily notice that a particular contraction (for example, *can't*) has sometimes been mistyped as *ca'nt*, a computer cannot detect the error unless it is programmed to do so. It is essential, therefore, that clear guidelines are established, to reduce the risk of inconsistency. Furthermore, it is important to implement a thorough procedure for checking each transcription. This might involve each transcriber checking the work of a fellow transcriber, or it may be a case of the researchers methodically monitoring the work of the transcribers. To ensure that the correct procedure is followed, it is useful to have checkers record the completion of the review in the documentation of the corpus. In the header for each transcript file in the Michigan Corpus of Academic Spoken English (MICASE), for example, the names of the transcribers, and the checkers, and dates for completion of each of these stages, are given. Pickering et al ([1996](#)) provide a thorough account of procedures they followed in assessing the reliability of prosodic transcription in the Spoken English Corpus.

4. Representation and annotation

So far we have considered only the problems of transcription, and we have concentrated on issues relating to how the data can be represented in different ways. The discussion has centred on representations that can be read or heard by the human eye or ear, and has largely ignored the question of machine readability. While it is easy enough for example, to display information in single or multiple columns, as in the example given above, using a particular word-processing package, it cannot be assumed that all users of a corpus will have the same package nor that their analytical software will be able to cope with the coding of the data used by the word processor. It is necessary therefore to choose a method for marking up the data and adding annotations that is relatively independent of particular operating systems and commercial packages. Mark-up languages such as HTML and XML are widely accepted means to achieve this. Both HTML and XML are forms of SGML (Standardized General Markup Language) and one of the advantages that XML has over HTML is that it is, as its name (eXtensible Markup Language) suggests, extensible. Users can extend the range of elements, attributes and entities that are permitted in a document as long as they state the rules clearly. XML is now the chosen form of mark-up for future releases of the BNC and many other corpora that follow the Guidelines of the Text Encoding Initiative. There is no space to describe these guidelines in detail, and they have been discussed in other chapters (Burnard, [chapter 3](#)) in this book. What is worthy of note here, however, is that the TEI Guidelines provide a set of categories for the description of spoken language data, and that they form a powerful and flexible basis for encoding and interchange.

It is usually hoped that corpora will be made available for other researchers to use, and in this case it is necessary to create a corpus that is in a suitable format for interchange of the resource. There are also closely related issues to do with preservation of the resource (see [chapter 6](#)). I recently requested a copy of an Italian corpus called PIXI [corpus may be ordered from [OTA](#)] from the Oxford Text Archive. The files for the corpus are in WordPerfect format, and I opened them using both a text editor, and WordPerfect 9. As the 'Readme' file informs me, there are some characters in the files that are specific to WordPerfect and which do not convert to ANSI. Some of these characters were used in this corpus to mark overlap junctures.

The version I see shows:

```
<S C><p> $$Li avevo gia?presi, esatto.%% Poi pero? $ne avevo ordinati -% <S A><p> $Allora aspetti che guardiamo% se e?rimasto:
```

This shows how the particularities of word-processors and the character sets that they use create problems in interchange between different programmes, even between different versions of the same word-processing package.

Where interchange is an issue, then, consideration must be given to finding ways to encode the transcription in such way that the files can be exchanged between computers running on different operating systems, and using different programmes for text browsing and analysis.

A second possible desideratum is that data can be represented in a variety of ways. As noted above in [Figure 5](#), there are several different conventions for indicating pauses in a transcript. If one has access to transcripts transcribed following MacWhinney's conventions, and wanted to convert one's own transcripts that had been transcribed using a different set of conventions, it would be useful to be able automatically generate a representation of one's data that conformed to MacWhinney's system. This would require that all short pauses be transformed from their original representation to a single hash sign, and timed pauses be shown as a single hash sign followed by the measurement of the pause.

For both these purposes, use of the TEI Guidelines in order to encode the transcripts following standardized procedures offers a good solution. The TEI recommendations provide a comprehensive set of conventions for the underlying representation of data. The analyst then has the potential to represent the data in any number of ways, through the use of stylesheets. If you want to present your transcript in the MacWhinney style, you can create a stylesheet which specifies, among other things, that all pauses which have been marked up following TEI guidelines as `<pause dur="short"/>` be converted to single hash marks. A second stylesheet, for the Du Bois et al system, would transform `<pause dur="short"/>` to ..., and obviously the same set of transformations would be specified for any other feature of the transcript, for which an equivalent exists in the Du Bois et al system.

Johansson ([1995](#)) presents a clear exposition on the TEI guidelines for the encoding of spoken language data. While some of the details are slightly outdated (for the latest statement, see the online TEI Guidelines, particularly sections 11, 14, 15 and 16, at <http://www.tei-c.org/P4X/>; the chapter 'A Gentle Introduction to XML' , is also recommended). The tagset specified in the TEI guidelines for transcriptions of speech covers the following components:

- utterances
- pauses
- vocalized but non-lexical phenomena such as coughs
- kinesic (non-verbal, non-lexical) phenomena such as gestures
- entirely non-linguistic events occurring during and possibly influencing the course of speech (e.g. sound of truck reversing in road next to lecture hall)
- writing
- shifts or changes in vocal quality (TEI Guidelines 11.1).

This tagset is a list of options, and it is also extensible. Johansson demonstrates how other tags can be used to mark tonic units, for example, and how entities can be created to indicate intonational features of speech. Furthermore, a TEI document has to contain both a header and a body, and the header will contain the background information about the recording/event (making such information easily accessible) and a clear statement of what mark-up conventions are followed in the document.

The TEI guidelines have been criticized by some as over-prescriptive and excessively complicated, and Johansson ([1995](#)) addresses some of these criticisms. In defence of the TEI, it can be said that, although the verbosity of the coding leads to bloated files, the guidelines allow for reasonable degrees of flexibility, and promise to increase levels of interchangeability. In addition, the shift from SGML to XML for mark-up of the texts has given both compilers and users many more options. There are several commercial XML editing packages, and the latest versions of mainstream word-processors and internet browsers are XML-aware, which means that it is now easy to edit and to view XML files (the same was not true of SGML documents). Furthermore, with its growing uptake in industry, XML looks likely to become a standard for document mark-up. To create stylesheets for the specification of desired output formats, XSL (eXtensible Stylesheet Language; see <http://www.w3.org/Style/XSL/> for details) can be used. For updated information on TEI and its relation to XML, see Cover pages, <http://xml.coverpages.org/tei.html>.

Where interchangeability is an important feature of the corpus, then, it is advisable to follow the TEI Guidelines. This will mean that some members of the research team will need to become familiar with the technicalities of XML editing, of DTD creation (document type definition) and, for the development of stylesheets, XSL (there are a number of free TEI stylesheets at <http://www.tei-c.org/Stylesheets/>). Other corpus developers may feel that this is beyond their needs or capabilities, however, and seek alternative methods. If a corpus of phonetic transcriptions is to be used only by a limited number of phoneticians, each of whom use the same software for reading and analyzing the corpus, it is not necessary to invest time in training the transcribers in XML. Questions that may need to be addressed, however, are:

- Can information be extracted from the files easily?
- Are tags unique and distinct?
- Do certain tags require the addition of an identifier for each instance?
- Is it useful to be able to remove all tags quickly?
- Is all the background information (details about participants, the context, etc) easily accessible?

For information to be easily extracted, it is important that tags are unique and can be searched for without risk of ambiguity. It is important that the beginning and ending of a string of data that is coded in a particular way are clearly marked. If sequences of a feature are of interest, it may be helpful to give each occurrence a unique identifier (e.g., in a study of teacher talk, each instance of a teacher *initiation* move could be given an ID number as in <tchr_init id="23">). The use of angle brackets for enclosing tags makes it easier to remove all tags in a single action (the text editor *NoteTab* — <http://www.notetab.com/> — has a 'Strip all tags' command that works on angle bracket delimited tags; similarly, *WordSmith Tools* has the option to ignore all strings with angle brackets). Background information, where relevant, should be easily accessible, either because it is contained within the files, or because it is stored in a database that is linked to the corpus. Failing this, the file naming system should provide clear indications of the contents of each file.

Transcripts can now be linked to the original digitized recordings, either on audio or video. Roach and Arnfield (1995) describe their procedure for automatic alignment of audio with prosodic transcriptions, although this is highly sophisticated operation. A cruder alternative is to place markers in the transcript that point to precise timings within the sound files. To align video, audio and transcript, programmes such as the freeware programmes *Anvil* (<http://www.dfki.uni-sb.de/~kipp/anvil/>) and *Transana* (<http://www2.wcer.wisc.edu/Transana/>) can be used.

Lastly, a word on linguistic annotation. With written text, it is a relatively easy task to prepare a text for automated POS tagging using a tagger such as the CLAWS tagger that was used for the British National Corpus (<http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/>) but spoken language data require a considerable amount of preprocessing, before an automatic tagger can deal with the input. Questions need to be posed about how to deal with false starts, repetitions (e.g. 'the the the the'), incomplete clauses, and so on: should they be removed? Should they be 'corrected'? For purposes of parsing, as Meyer (2002: 94-96) explains, the questions are redundant - it is simply essential that the transcript be rendered 'grammatical', because a parser cannot deal with the input otherwise. A further question to consider in relation to annotation and the removal of 'ungrammaticality' from the transcripts is: if features are to be removed, are they permanently deleted or are they removed temporarily and restored to the transcript after the POS tagging has been completed?

5. Access

The final stage in the process of establishing a corpus of spoken language data, depending on the purposes for which the corpus is to be used, is that of making the corpus available to

others. Funded research work often requires the researchers to deposit copies of the data with the funding body, but a more interesting possibility is to make the data accessible to the wider academic community. The printed version of the transcripts could be published, for example, but this restricts the kinds of analysis that can be made, and it is preferable to publish electronic versions where possible. The Oxford Text Archive (<http://www.ota.ox.ac.uk/>) acts as the repository for electronically-stored corpora and other forms of data collections in the UK, and there are similar centers in other countries. These are simply repositories, and do not provide an analytical interface to the data. An exciting new development is the MICASE corpus (<http://www.lsa.umich.edu/eli/micase/index.htm>), a collection of transcripts of lectures, seminars and other academic speech events, which is searchable on-line, through a web interface, that allows searches to be refined through specification of a range of parameters. Such open access promises to make the analysis of spoken language data easier for a wider audience, at no extra cost other than the Internet connection. The transcripts can also be downloaded in either HTML or SGML format.

A criticism levelled against the Survey of English Usage and the BNC was that the listener could not access the audio recordings to hear the original and to make comparisons with the transcript. It is clearly of benefit to other researchers for the original recordings to be made available, as this is an extra source of information, and some corpus projects have provided such opportunities for access. The COLT corpus will be accessible through the Internet and a demo version is online at: <http://torvald.aksis.uib.no/colt/>. With each concordance line that appears, a link to a short audio file (.wav format) is included. MICASE, according to its website, plans to make the sound files available on CD-ROM to academic researchers, and also to make most of the sound files available on the web in RealAudio format: several of these files have been placed on the site, but they are not linked in any way to the transcripts and it is not possible to search through the sound files in any way. Due to certain speaker consent restrictions, furthermore, not all recordings can be published. [Thompson, Anderson & Bader \(1995\)](#) is an interesting account of the process of making audio recordings available on CD-ROM.

The linking of the transcript to audio or audio files is an area for major development in the coming years, as retrieval and delivery technologies become more powerful and sophisticated. Spoken language is more than simply the written word, and the audio and video recordings of the original spoken language events offer invaluable resources for a richer record. As mentioned above, there are programmes such as *Transana* and *Anvil* which allow the analyst to link audio, video and transcript, but they also tie the user into the software. In order to view the project files, it is necessary to have a copy of the software. What is ideally needed is an independent means of linking transcript, audio and video that

uses XML, Java or other web-friendly technologies efficiently so that the tools and resources can be accessed by anyone with a browser, and necessary plug-ins.

As is the case with recording technologies, as discussed above, it is difficult to make any recommendations about the best formats for storage and delivery of a corpus and related data. For transfer of large quantities of data, such as audio or video recordings, CD-ROM and DVD offer the best options at present, but it is likely that new technologies for storage and transfer of data will develop.

A major consideration in storing audio or video data is the size of each file (or clip). Obviously, the larger the clip, the longer it will take to load, and the more the demands that will be placed on processing capability. Compression technologies, such as mp3 for audio, make it possible to create smaller files. Web delivery of audio and video material to supplement a corpus is possible but at the moment the transfer rates are prohibitively slow. Streaming video has the potential to deliver data reasonably quickly for viewing purposes, but would be cumbersome to work with if alignment with the transcript were required. For the moment, at least, it seems that the best way to make a multimodal corpus available is through CD or DVD media.

In summary then:

- The development of a corpus of spoken language data is a complex task and it requires careful planning.
- Planning is important to ensure that all relevant information is collected, and to assist in maintaining high levels of consistency (consistency of data quality, of transcription procedures and conventions, of markup and of information about the events)
- A choice needs to be made between breadth and depth — between capturing large amounts of spoken language data, and annotating the data in great detail
- Where resources are available, try to keep as much data as possible and make it possible to represent the data in a variety of ways
- The possibilities for linking the transcript with audio/video files are limited at present, but should develop rapidly in coming years.

Chapter 6: Archiving, distribution and preservation (Martin Wynne, University of Oxford © Martin Wynne 2004)

1. Introduction

Once you have created your corpus, what happens next? This chapter attempts to explain how good planning can ensure that, for as long as possible into the future, a corpus is useful and usable for a wide range of potential users.

Usually the creation of the corpus was not an end in itself, but was conceived as part of a research project, and it is only when the corpus building has finished that the real work begins. But the corpus is likely to be of potential value to many more researchers outside of the corpus creator's research group, so it is also advisable to plan to make sure that other users can make use of it too. Ensuring the initial and ongoing availability and usefulness of the corpus is the subject of this chapter.

It is not recommended that you start to address these issues only at the end of the corpus building project. A successful project to create a digital resource will usually have planned for the entire life-cycle of the resource, including what happens after the resource is created.

At the planning stage, it is important to ask whether, under the project plan, the corpus is likely still to be available and usable in one, or ten, or twenty years' time. Potential risks to its future viability include termination of funding, changes in staff or management, changes in technical infrastructure, obsolescence of the technologies associated with the resource and changes in standards. It is possible to be specific and say that it is certain that, at some point, the project funding will end, some of the staff will leave, the computers will be replaced, the servers will be upgraded, the software used to access the corpus will change, the interests and priorities of the staff involved will change and they will eventually get different jobs or retire.

To ensure ongoing availability and usability of the resource, it is desirable to remove reliance on particular individuals, institutional arrangements or technologies. This can only really be effectively managed in the context of an archive which is a trusted repository and which has a long-term access and preservation strategy for its collections.

The following sections attempt to cover some of the important issues which it is useful to consider at the planning stage of the corpus building project.

2. Planning for the future

Stop developing the corpus!

The first thing to say here may appear obvious, but it is sometimes necessary to remind corpus builders to stop developing the corpus. While it is important to achieve as low a rate of errors as possible, there is a danger of excessive perfectionism, which can lead to a situation in which the corpus is never finished, preventing its use and reuse. There may be similar problems if a corpus is made available, but then repeatedly revised, preventing the comparison or replication of results based on its analysis.

It is of course possible to conceive of a corpus which changes in a principled and useful way. For example, a monitor corpus is repeatedly updated with new texts and is constructed in such a way that language change over time can be analysed. For a dynamic resource of this type to be useful, it needs to develop in a managed, predictable and well-documented fashion, and in a way which is transparent to the users.

The corpus creator may plan to add annotations to the text. It is also likely that a well-constructed resource which is made available will have annotation added to it by other researchers. It is good practice to release a version of the corpus without annotation, however, for several reasons. Firstly, there are likely to be many users who do not wish to use the annotation, or indeed who use tools which find it difficult to process a corpus with certain types of annotations. Secondly, the annotation process may involve changing the text in some ways, such as changing the word tokenisation, or removing certain elements. The latter can happen deliberately, or accidentally, and may not be easy to detect. It is therefore important that an original version of the corpus be available for reference purposes.

Delays in finishing a corpus can be caused by checking and correcting errors in the text and markup. If it possible to have a clear idea from the start of a realistic level of quality which is required and an accurate means of measuring this, then it is much easier to know when the acceptable level is reached. Do bear in mind that this may have to be 'good enough' rather than 'perfect'. While it is tempting to attempt to create a corpus which is perfect and a thing of beauty, the important thing is for the corpus to be 'fit for purpose'. It is also worth bearing in mind that most of the techniques of corpus analysis require the identification of repeated patterns. While errors will skew results and may, if serious, hide certain important patterns, you may also be able to rely on a tendency for repeated patterns to shine through despite a certain error rate. In any case, the extent of quality control checks should be documented.

It will also be easier to stop if your project plan is scalable. If your workplan requires everything to be dependent on a final processing stage which can only take place if all

previous stages are completed 100% successfully, then there is a high risk of failure. At best, the corpus building process may drag on for a long time beyond the projected end date, with all the problems associated with carrying on without the necessary funding and support. If on the other hand, the project has been designed with a more robust and scalable plan, then there is a much greater chance of successful completion of the project. Such a plan might involve complete production of sub-sections at various stages, with a design that will still work if less than 100% of the texts are successfully collected and processed.

What are my rights and responsibilities?

Corpora are usually made of texts written by different people, and the authors or owners of these texts have intellectual property rights. In addition, the fact that intellectual work has gone into the sampling selection, markup and annotation of texts means that corpus creators have rights over the corpus as a collection. The project to create the corpus will probably have a funder, the work will usually be done within an academic institution which may claim ownership over the products of research. Several people will have been involved. The rights of these stakeholders can potentially restrict the use, reuse, sharing and long-term preservation of the corpus.

The relevant laws in the UK forbid the copying of published materials without the permission of the rights holder. The fact that a text is available freely on the web does not mean that it 'in the public domain' and you can put it in your corpus. On the contrary, publication on the web confers the right of ownership on the creator, and makes copying illegal, even if this is only for your private use. In practice such rights and prohibitions need to be tested in court, and it is usually the case that the corpus developer has to assess the probability of being sued rather than being able to obtain a clear statement of the legal position regarding the use of a text in a corpus. It may be that increased visibility of a widely distributed corpus might increase the likelihood of legal action in defence of copyright. In any case, it is advisable for these issues to be explored and clarified at the planning stage of the project, to ensure that you do not spend time constructing a corpus which cannot then be used legally.

Any agreements which were entered into with funders, copyright holders, publishers, data developers, archives, research assistants and other stakeholders need to be considered and documented. As an example of a responsibility to a funder, if your corpus development project is funded by the Arts and Humanities Research Council (AHRC) in the UK, you will normally be expected to deposit the completed resource with the Arts and Humanities Data Service (AHDS). Measures need to be taken to make sure that the documentation of these issues will continue to be available, preferably in an electronic form which is associated with the corpus. Ethical considerations may be relevant, especially if your corpus is the product of

linguistic field-work. It may be useful to conduct a stakeholder analysis, an established business management technique. This analysis would attempt to consider the points of view of the various parties who have an interest in the corpus. It can be useful to highlight potential conflicts, in legal and ethical questions, and may help the development of a plan to ensure that the necessary steps are taken.

It is also useful to document any ways in which the rights associated with any of the materials are going to change. Are some texts likely to come out of copyright soon? If so, which ones and when? Are your rights in some materials likely to expire? For example, have you made use of journal texts or images which you only have rights to for a fixed period of time? These issues need to be discussed with an archivist, and any relevant information included in the metadata. It is likely that future changes in the legal status of the corpus texts can only be dealt with effectively by an archive with the relevant procedures in place.

How is the corpus stored?

First, it is necessary to have some backup procedures during the data collection and data development stage of your corpus building project. While your own ad hoc procedures can be useful for providing extra copies and having them easily to hand, it may be best to make use of professional backup facilities such as those which should be offered by your the computing service at your institution.

Once the corpus is completed, then it is necessary to archive it. It is perhaps useful to explain here the distinction which is usually made by information professionals between backup and archiving. Backup means taking a periodic copy of a file store. Archiving means the transfer of information of public value into a separate repository where it is to be held indefinitely, or for an agreed period of time. It is likely that you will need backup solutions during the lifetime of your project, and you will need to find an archiving solution when the resource is completed. It is however useful to plan the archiving from the start, so it is a good idea to talk to the archivists and make sure that the resource can be provided in an appropriate format, and also so that you can include the time and effort necessary for depositing the corpus in the archive in the project workplan.

In terms of the technical solutions for backup and archiving, there are important issues to do with media, location, metadata and management. Storage media are susceptible to the breakdown and the loss of data. The possibilities of fire, theft and damage need to be considered. It is necessary to consider how the media and files are labelled, and how the documentation is associated with the relevant resource. These technical issues are not covered in detail here, as they are subject to constant change due to technical innovation, development of standards and changes in practice. It is best to consult the AHDS, or other

information professionals, for up-to-date advice which takes into account the latest developments.

Where is the corpus archived?

You are likely to need to store the data locally during the data development phase, and you will undoubtedly want to continue to do this so that you can use it. However you may opt to pass on the job of archiving, cataloguing, distributing and preserving your corpus to an organisation which offers professional archival services, such as the AHDS.

The fact that the corpus is archived elsewhere does not mean you lose rights over your resource. An archive will not normally acquire any exclusive rights over the corpus. The creator and other rights holders do not lose any of their rights. The normal arrangement is for the resource creator to retain ownership, and to grant the archive permission to keep a copy, and, possibly, to distribute the resource. The arrangement should be non-exclusive, meaning that this does not prevent the corpus creator from depositing it elsewhere, and it should be possible to dissolve the agreement. You should check the licensing agreement for these and other issues which are relevant to you if you deposit your corpus in an archive. It would also normally be necessary to take a look at the terms under users may be able to download the corpus, and check that this does not come into conflict with any of your rights or responsibilities.

As long as the agreement is non-exclusive, you can continue to distribute the corpus yourself, develop it and exploit it in other ways.

Who will have access to the corpus?

There are several factors which sometimes influence corpus builders not to make resources more widely available. Some are listed below:

- to avoid copyright and other rights issues;
- to ensure that the creator has the first, or even exclusive, opportunity to exploit the resource and publish research or further resources based on it;
- to retain the option to sell the rights on a commercial basis;
- because of the danger of uncontrolled commercial exploitation or pirating;
- because it is too much trouble to administer distribution.

Avoiding legal issues

It should be noted that the first reason above is not a sound one from the legal point of view. As noted above, copying texts and putting them in a corpus can constitute a breach of copyright, whether or not the corpus is then distributed.

Getting the first chance to use the data

While it may be desirable for the creator to have the first opportunity to publish results based on the corpus, it is also desirable that any published results be replicable, which means that the corpus on which the research is based needs to be made available to other researchers. In any case, the creator will normally have a head start over other researchers, with a research agenda already in place and underway as soon as the corpus is completed. Delaying the deposit of the corpus in an archive runs the risk of the data becoming corrupted, or of versions of the resource becoming confused. In some cases delay leads to the deposit never happening, as priorities and circumstances change.

Releasing the corpus commercially

If commercial exploitation of the corpus is an option, the creator must weigh up the options. While a commercial deal may please your employer, and bring some financial reward, there are some good arguments for open access. The more widely available the corpus is, the more widely known it is, and the more publicity the creator will receive. A community of researchers who work on the corpus will come into being, creating a higher profile for research based on the resource, including your own. Feedback will be obtained on the usefulness of the resource, and errors can be corrected. Others are more likely to share their resources with you if you share yours. Funders are more likely to give you more funding if you have a good record of ensuring that resources which you have created are properly archived and distributed. The funders generally perceive better value for money in creating resources that are reusable. Failure in this respect could seriously weaken a proposal for further funding. Further project funding may be more lucrative and prestigious than what can be obtained from commercial exploitation of the data. In any case, commercial publication and open access are not necessarily mutually exclusive. It may, for example, be possible to sell copies of a corpus bundled with access software, while also making the raw corpus data freely available.

Concern about unrestricted access and piracy

Concern about piracy is not a good reason not to deposit a corpus. It is likely to be easier to control access and defend the rights of stakeholders if the corpus is distributed through an archive. A reliable archive will have a rights management policy, and have the means to take action to defend rights that are violated. The corpus creator is unlikely to want to get involved in these issues, even with local institutional support.

It's all too much trouble

It is not necessarily as much trouble as you might think. It should be noted the AHDS normally offers a free service to academics in the UK to archive, catalogue, distribute and preserve corpora, and so the expense and work of the administration of granting access does not need to be borne by the corpus builder or their institution.

Open access: conclusion

It is for the corpus developer to weigh up these issues and decide whether they want to enlist the help of an archive to distribute the corpus. In the short term they may be able to manage distribution of a resource, but it is unlikely to be viable in the long term. If the situation regarding access is not clearly defined and well-documented then this could seriously affect the future viability of the resource. The developer could thus fail to meet the expectations of their funders, users and other stakeholders. Managing access and dealing with rights issues can be time consuming and complex.

In the event that it is not possible to distribute the resource for some valid reason, it is still be good practice to deposit a copy of the corpus in an archive for long-term preservation purposes. Such an arrangement can normally be negotiated with the AHDS.

How will users find the corpus?

Depositing your corpus in a trusted archive should help ensure that best practice is followed in ensuring the security, availability and long-term preservation of the corpus. It should also help users to find the resource, since an effective archive will make its catalogue records visible to potential users. They will participate in sharing of resource descriptions, through open archives initiatives and institutional and subject portal projects. Such initiatives are currently growing in importance. One of particular relevance to the field of corpus linguistics is the Open Language Archives Community (OLAC, <http://www.language-archives.org/>). All of the major archives of language resources have come together in OLAC in order to enable users to go to one place to search for corpora and other resources held in different archives and repositories. The creation of this community is also helping the development of standards in the description of resources.

Many more initiatives within institutions and different communities to share information about resources are likely to appear in the coming years, in the shape of portals, virtual learning and research environments, institutional archives and online library and information systems. These are all likely to be built on the open standards which are used by archives and other trusted repositories. Depositing your resource with an archive means that they will catalogue

your resource according to appropriate standards and thus make it possible for the existence and availability of the corpus to be discovered via these mechanisms.

What file format should my corpus text files be in for archiving?

One piece of important general advice for file formats for digital preservation is to avoid ties to proprietary formats. If your corpus is made up of files in a format for a commercial word-processing program, such as Microsoft Word, then they cannot be processed by most corpus analysis tools. What is more, the format may not be supported indefinitely into the future, and there will come a time when users won't be able to read the files any more. XML is usually considered to be a more appropriate file format for long-term preservation, because it is an open international standard defined by the World Wide Web Consortium (W3C), it is not tied to a particular applications or platforms and it uses Unicode (another open standard) for encoding the text. However, it should not be thought that simply saving files as XML is a panacea for all archiving and preservation problems. It is perfectly possible to use XML to make a corpus which is in an appropriate form for long-term preservation, but it is also very easy to make a corpus using XML which is NOT viable in the near, let alone distant, future. Simply automatically converting a file from a word-processing format to XML does not magically make it into a good electronic resource. Recommendations of preferred file formats, encoding schemes and software options can obscure more important factors. Open standards like XML are preferred because they make it possible to encode the intellectual content of the resource and the metadata in a consistent and unambiguous way. While there are reasons why XML, and Unicode, are desirable, and likely to become more firmly entrenched and widely used for language corpora, it is often trivial to migrate from other formats and standards, including proprietary ones, as long as good practice has been followed in the creation of the electronic text in whatever format. There should be no short-term problems with converting a text file created and edited in MS Word in which the various relevant textual phenomena have been dealt with in a principled and consistent way.

There is however a particular issue with text corpora, which means that the type of text encoding is especially important. To use a corpus, the text needs to be searchable, preferably with generic tools. This means that binary encoding formats, such as PDF, RTF and Word are inappropriate, and 'plain text' or Unicode (with or without markup) are preferable. There is unfortunately a conflict here between the needs of corpus linguists and those working in the archiving, preservation and digital library worlds. The latter are generally more concerned with ensuring that the content of text documents and other types of data are preserved, sometimes including the 'look and feel' of a text, rather than preserving the 'searchability' of the texts. For this reason, proposals from the digital preservation

professionals for an open standard for PDF for preservation purposes (PDF-Archive), or any other kind of binary format, are not appropriate for language corpora (see <http://www.aiim.org/standards.asp?ID=25013>). Indeed, it would be a hindrance to linguists hoping to use electronic archives as the basis for research if the archives were to adopt binary formats for preservation.

While it may be convenient to use one file format for all stages of the life-cycle of a corpus, it may well be the case that the best preservation formats are not the best formats for the data development stage, or for using with the relevant analysis tools. In this case, a separate preservation version of the resource may be created. But it is important to bear this in mind while developing the corpus and to make sure that information necessary for accompanying the preservation version is not lost. For electronic text, this means avoiding the insertion of annotation or processing instructions in such a way that the original text and its structure are not recoverable. In the case of audio data, this means capturing, storing and depositing the best possible quality, in an uncompressed audio stream, and then converting to a more convenient lower quality, compressed sound for analysis and distribution, if necessary.

A further discussion of issues in digital preservation of electronic resources in humanities disciplines can be found in [Smith \(2004\)](#).

3. Conclusion

Unambiguous, rigorous, consistent and well-documented practices in data development are usually more important than the technologies used. There are preferred options for file formats, encoding, markup, annotation and documentation, but these will change over time. For the latest recommendations, consult the Arts and Humanities Data Service (<http://www.ahds.ac.uk/>) at the planning stage of your project, and build into your workplan adequate time and resources for the preparation of the corpus for distribution, archiving and preservation.

The general advice here is for conformance to open standards in corpus creation and documentation, but it is acknowledged that there is more than one way to do this. It is hoped that these are the messages of this entire guide.

Appendix: How to build a corpus (John Sinclair, Tuscan Word Centre © John Sinclair 2004)

Introduction

The job of corpus building divides itself into two stages, design and implementation, but these cannot be completely separated, for reasons which are largely practical.

One is the cost. Nowadays most corpora are put together from text that is already digitised; the cost of putting into electronic form text which only exists on paper is very much greater than the cost of merely copying, downloading and gathering data that is already digitised; so there has to be a compelling reason for using any of the more laborious methods which were used to capture data in the days before electronic text.

Sometimes, however, it is necessary, to do things the hard way; for a corpus of informal conversations, for example, or historical documents or handwritten or manuscript material. But in all such cases it is worth a serious search of various collections and archives, and perhaps a query on the professional lists, before undertaking the labour of entering new text.

Another reason for mixing principle and practice in corpus building is because some kinds of data are inherently difficult or even impossible to obtain, and a measure of compromise is often necessary; some authors categorically refuse to have their work stored in a corpus or insist on high fees; some types of interaction are extremely difficult to make records of; in many countries surreptitious recording is illegal;¹ some documents that use graphics are unscannable and have to be unpacked before being laboriously typed in to the corpus.

For languages that are used in substantial segments of the globe there will be found a very large amount of text material on the internet. Even for smaller languages there is often a remarkable amount and range of material. If the electronic resources available are not adequate then the least expensive alternative is scanning printed texts; however this is time-consuming and the output from the scanner needs to be edited at least superficially. See below on Perfectionism.

The worst option is to have to type in large amounts of textual material; this is still unavoidable with transcripts of spoken interaction, but requires a consumption of resources that drags a project, limits its size and reduces its importance. Keying in may be a viable option for individual texts which are not available in digital form and which are not easy to scan, but for a large text corpus, there are likely to be easier options.

The World Wide Web

While web pages are likely to be the most immediately accessible sources of material, they are by no means the only source, and some of the most valuable text material is merely indexed on a web page, requiring further searching. For example, many large document archives put up their catalogue on the web, and give opportunities for downloading in various formats. Here the web is playing the role of a portal. Other providers of text data may issue CDs, especially when there is a lot of data to be transferred. Sometimes payment is required, especially for material that is popular and under copyright; corpus builders should consider carefully the costs of such data and whether it is justified.

Also available on the internet are many — probably millions — of documents that are circulated by e-mail, either messages or attachments. By subscribing to appropriate lists, your collection of material can grow quickly.

The Web is truly bountiful, but it is important to appreciate that the idea of a corpus is much older than the Web, and it is based on "hard-copy" concepts, rather than cyber-objects like web "pages". A corpus expects documents (including transcripts) to be discrete, text to be linear and separable from non-text, and it expects documents to fall into recognisable sizings, similar to hard-copy documents. A normal corpus has no provision for hypertext, far less flashing text and animations. Hence all these familiar features of the Web are lost unless special provision is made to retain them. The procedural point (1) — see below — is relevant here; the documents in their original format should be carefully preserved; it is up to the corpus managers how far hypertext links are preserved as well in a "family" of documents, but, like all the other texts in a corpus, the Web document is ultimately removed from the environment of its natural occurrence.

Some projects are learning how to make multimedia archives within which spoken or written text is one of the data streams, and a more modern notion of a corpus may result from this research. Linguists need make no apology, however, for concentrating on the stream of speech or the alphanumeric stream; particularly in the early stages of a new discipline like corpus linguistics the multimedia environment can be so rich that it causes endless diversions, and the linguistic communications can get submerged.

At present it is important to know precisely what is actually copied or downloaded from a web page. This is not always obvious, and quite often it is not at all the document that is required. The "source" file, which contains all the mark-up, is easy to download but difficult to handle; "text-only" or "print-friendly" versions of a page can be helpful. In all cases it is essential to review what you have tried to capture to make sure that it is the target document — it may only be the address, or a message such as "page not found".

The cheerful anarchy of the Web thus places a burden of care on a user, and slows down the process of corpus building. The organisation and discipline has to be put in by the corpus builder. After initial trials it is a good idea to decide on a policy of acquisition and then stick to it as long as it remains practical; consistency is a great virtue as a corpus gets larger, and users of a corpus assume that there is a consistency of selection, processing and management of the texts in the corpus. Already we read a lot of apologies for the inadequacies of corpora, often inadequacies that could have been avoided.

Another tricky question is that of copyright — not the familiar copyright of publications, but the more nebulous issue of electronic copyright. In principle, under UK law, publication on the internet confers the rights on the author whether or not there is an explicit copyright statement. Every viewing of a web page on a screen includes an act of copying. If there is doubt, contacting the named copyright holder is advisable.

Perfectionism

So when you design a corpus it is probably best to write down what you would ideally like to have, in terms of the amount and the type of language, and then see what you can get; adjust your parameters as you go along, keeping a careful record of what is in the corpus, so that you can add and amend later, and if others use the corpus they know what is in it.

It is important to avoid perfectionism in corpus building. It is an inexact science, and no-one knows what an ideal corpus would be like. With good research on such matters as the penetration of documents in a community, our present guesswork can certainly be improved on, and even the influence of the spoken word relative to the written word may be estimated more securely than at present. Until then compilers make the best corpus they can in the circumstances, and their proper stance is to be detailed and honest about the contents. From their description of the corpus, the research community can judge how far to trust their results, and future users of the same corpus can estimate its reliability for their purposes.

We should avoid claims of scientific coverage of a population, of arithmetically reliable sampling, of methods that guarantee a representative corpus. The art or science of corpus building is just not at that stage yet, and young researchers are being encouraged to ask questions of corpora which are much too sophisticated for the data to support. "It is better to be approximately right, than to be precisely wrong."[2](#)

We should also keep a distance from claims of accuracy of analysis by current software. Even what seems to be almost perfect accuracy is likely to be systematically inaccurate, in that whole classes of data are always misclassified. This problem arises because accuracy is in the mind of the analyst, and may not correspond with the distribution of patterns in the

corpus. Furthermore, in a corpus of, say, a hundred million words, 99% accuracy means that there are more than a million errors.

Indicative, not definitive

The results of corpus research so far are indicative of patterns and trends, of core structures and likely contributions to theory and description, but they are not yet definitive. It should become a major objective of serious corpus research to improve the procedures and criteria so that the reliability of the descriptive statements increases. However, this move to greater maturity of the discipline is not an admission of any limitations; we established above that corpora are ultimately finite and that this is a positive property of them, giving them descriptive clarity. A description based on an adequate theory and a very large and carefully built corpus, combined with flexible and theory-driven software will provide descriptions far above what we live with at present. The fact that, like any other conceivable description, they will not reflect the ultimate flexibility and creativity of language will be of interest to a small group of specialists, no doubt, but not to the mainstream of research.

Corpus-building software

If you ask Google for "corpus builder" — at the time of writing — you get a number of useful leads which can support corpus-building activities, for example, recognising a particular language in order to select only texts in that language. Software is also offered which will build a corpus for you, index it and allow searches of various kinds. More and more of this kind of product is likely to appear, and according to your purposes and resources you may look into suitable packages. Some are commercial ventures and sell at up to a thousand euros or so, and these normally allow a trial period, which is worth investigating. Study the small print carefully, looking for limitations of size, speed and flexibility, and make sure that the software will perform as you want it to.

Free or open-source software is often more specialised than the commercial products, but is more likely to be tricky to install and is not always friendly to use, so be prepared for some initial problems with this.

Procedure

The main considerations that affect the choice of texts in a corpus are given in the paper above under "Representativeness". Once a text has been chosen for a corpus, and the location of a copy in some usable format has been determined, then there are several recommended steps towards making a useful and understandable corpus.

1. First, make a **security copy** of the text, in exactly the format received. If it is not in electronic form, keep the hard copy version for later reference³,
2. Save the text in **plain text** format. Sometimes this is not straightforward, and in extreme cases a text may have to be rejected if its formatting cannot be standardised. But most text packages have a plain text option. This step is recommended even if your intended processing package will handle a mark-up language like HTML, XML, SGML, or word processor output. The issue is flexibility — conventions keep changing, new ideas come in and suddenly everything is old-fashioned. A corpus consisting of texts in a mixture of formats is impossible to handle. Conversion from one format to another is usually laborious and uncertain, no matter what the optimists say. Plain text, the rock-bottom linear sequence of letters, numbers and punctuation marks, is almost always an easy conversion, and that is the one to keep.
3. Provide an **identification** of the text at the beginning of it. The simplest identification, and the one that makes the least disruption to the text, is a short reference — just a serial number, for example — to an off-line database where relevant information about the text is stored. More elaborate identifications are called headers, and these can be elaborate structures where information about author, date, provenance etc. is added to the text. To keep the added material separate from the text material, a corpus with headers has to be coded in a mark-up format, such as one of those mentioned above. The mark-up allows the headers and other additions to be ignored when the corpus is searched.
4. Carry out any **pre-processing** of the text that is required by the search software. The proprietary software intended for small corpora on a Windows platform normally includes all necessary steps in processing. For large corpora using Linux-type platforms, search engines frequently specify some initial processing to make the most popular retrieval tasks quick and efficient, e.g. the compilation of a list of all the different word-forms in the text, to be used as a basis for wordlists, concordances and collocational profiles.
5. When the corpus is complete, at least so that you can get started with research on it, make a security copy on a CD; as you add to an initial corpus, make further CD copies so that you can always restore the corpus following a disk crash (see also the advice in [Chapter 6](#) on archiving and preservation). Check your working version from time to time because mysterious corruption can affect your files. Always check the integrity of the corpus if it gives you strange results.

Notes

- [1.](#) In those countries that tolerate surreptitious recording, there is still the ethical issue of privacy, and anyone handling data of this kind should consider (a) offering the participants the option of deleting the recording after the event, (b) physically removing from the tape any passages which could be used to identify the participants.
- [2.](#) This is Rule 8 (Use Common Sense) of the 9 Rules of Risk Management published as an advertisement by the RiskMetrics Group, cited in *The Economist* of April 17th 2004, page 29.
- [3.](#) It is well within present technology practice to make a facsimile of a printed page and to align it with an electronic version of the text that is printed on it; the user could then call up

the physical image of the page at any time to resolve any issues of interpretation. This would do away with the need to have formatting mark-up tags in the text stream, at least in cases where the text is derived from a printed original.

Bibliography

AHDS: Arts and Humanities Data Service. <http://www.ahds.ac.uk/>.

AHRC: Arts and Humanities Research Council. <http://www.ahrc.ac.uk/>.

Allen, J., and Core, M. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. <http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/>.

Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM). <http://www.comp.leeds.ac.uk/amalgam/amalgam/amalghome.htm>.

BAAL: British Association for Applied Linguistics. BAAL Recommendations on Good Practice in Applied Linguistics. <http://www.baal.org.uk/goodprac.htm>.

Baker, J. P. 1997. Consistency and accuracy in correcting automatically tagged data. In *Corpus annotation: Linguistic information from computer text corpora*, eds. Roger Garside, G. Leech and A. McEnery, 243-250. London: Longman

Baker, P., Hardie, A., McEnery, A., Xiao, R., Bontcheva, K., Cunningham, H., Gaizauskas, R., Hamza, O., Maynard, D., Tablan, V., Ursu, C., Jayaram, B., and Leisher, M. 2004. Corpus linguistics and South Asian languages: Corpus creation and tool development. *Literary and Linguistic Computing* 19:509-524

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education

Burnard, L. 1995. *Users' reference guide to the British National Corpus*. Oxford: Oxford University Computing Services

Burnard, L. 1999. Using SGML for linguistic analysis: the case of the BNC. In *Markup languages theory and practice*, 31-51. Cambridge, Mass: MIT Press

Burnard, L., and Dodd, T. 2003. Xara: an XML aware tool for corpus searching. <http://www.oucs.ox.ac.uk/rts/xaira/Talks/cl2003.html>.

Carletta, J. 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics* 22

Carletta, J., McKelvie, D., and Isard, A. 2002. Supporting linguistic annotation using XML and stylesheets. In *Corpus linguistics: readings in a widening discipline*, eds. G. Sampson and D. McCarthy. London & New York: Continuum Interpretations

CLAWS part-of-speech tagger for English. UCREL.
<http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/>.

Clear, J. 1992. Corpus sampling. In *New directions in English language corpora*, ed. G Leitner, 21-31. Berlin: Mouton de Gruyter

COLT: Corpus of London Teenager. Department of English, University of Bergen.
<http://torvald.aksis.uib.no/colt/>.

Cook, G. 1995. Theoretical issues: transcribing the untranscribable. In *Spoken English on Computer*, eds. G. Leech, G. Myers and J. Thomas, 35-53. Harlow: Longman

Dunlop, D. 1995. Practical considerations in the use of TEI headers in large corpora. In *Text encoding initiative: background and context*, eds. Nancy Ide and Jean Veronis, 242. Dordrecht; London: Kluwer Academic

Edwards, J. 1993. Principles and contrasting systems of discourse transcription. In *Talking Data: Transcription and coding in discourse research*, eds. J. Edwards and M. Lampert, 3-32. Hillsdale, NJ: Lawrence Erlbaum Associates

Edwards, J., and Lampert, M. 1993. *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Garside, R., Leech, G. N., and McEnery, T. 1997. *Corpus annotation: linguistic information from computer text corpora*. London: Longman

GATE - general architecture for text engineering. <http://gate.ac.uk>.

Gibaldi, J. 1998. *MLA Style manual and Guide to Scholarly Publishing*. New York: Modern Language Association

Gibbon, D., Moore, R., and Winski, R. 1998. *Handbook of standards and resources for spoken language systems.vol. 1: spoken language systems and corpus design*. Berlin: Mouton de Gruyter

Gillam, R. 2003. *Unicode demystified*. Boston: Addison-Wesley

Goundry, N. 2001. Why Unicode won't work on the Internet: Linguistic, political, and technical limitations. <http://www.hastingsresearch.com/net/04-unicode-limitations.shtml>.

Granger, S. 1998. *Learner English on computer*. London: Longman

Granger, S., Hung, J., and Petch-Tyson, S. eds. 2002. *Computer learner corpora, second language acquisition, and foreign language teaching*. Amsterdam: John Benjamins

- Grice, M., Grice, M., Leech, G., Weisser, M., and Wilson, A. 2000. Representation and annotation of dialogue. In *Handbook of multimodal and spoken dialogue systems: Resources, terminology and product evaluation*, eds. D. Gibbon, I. Mertins and R. K. Moore, 1-101. Boston: Kluwer
- Halliday, M. 1993. Quantitative studies and probabilities in grammar. In *Data, description discourse*, ed. Michael Hoey, 1-25. London: Harper Collins
- Halteren, H. v. ed. 1999. *Syntactic wordclass tagging. Text, speech, and language technology*; 9. Dordrecht; Boston: Kluwer Academic Publishers
- Hirst, D. 1991. Intonation models: towards a third generation. In *Actes du XIleme Congres International des Sciences phonetiques. 19-24 aout 1991. Aix-en-Provence, France*, 305-310. Aix-en-Povence: Universite de Provence, Service des Publications
- Hofland, K., and Johansson, S. 1982. *Word frequencies in British and American English*. London: Longman
- Hofland, K. c. 1999. ICAME CD-ROM. HIT Centre, University of Bergen. <http://www.hit.uib.no/icame/cd>.
- Ide, N. 1996. Corpus encoding standard. Version 1.5. Expert Advisory Group on Language Engineering Standards (EAGLES). <http://www.cs.vassar.edu/CES/>.
- James, G., Davison, R., Cheung, A., and Deerwater, S. 1994. *English in computer science: a corpus-based lexical analysis*. Hong Kong: Hong Kong University of Science and Technology and Longman Asia
- Johansson, S., Atwell, E., Garside, R., and Leech, G. 1986. The tagged LOB corpus: Users' manual. Norwegian Computing Centre for the Humanities. <http://khnt.hit.uib.no/icame/manuals/lobman/INDEX.HTM>.
- Johansson, S. 1995. The approach of the Text Encoding Initiative to the encoding of spoken discourse. In *Spoken English on Computer*, eds. G. Leech, G. Myers and J. Thomas, 82-98. Harlow: Longman
- Karlsson, F., Voutilainen, A., Heikkilä, J., and Antilla, A. 1995. *Constraint grammar: a language-independent system for parsing unrestricted text*. Berlin & New York: Mouton de Gruyter
- Kipp, M. Anvil. <http://www.dfki.uni-sb.de/~kipp/anvil/>.
- Knowles, G., Williams, B., and Taylor, L. 1996. *A corpus of formal British English speech: the Lancaster/IBM Spoken English Corpus*. London: Longman

Korpela, J. 2001. A tutorial on character code issues. <http://www.cs.tut.fi/~jkorpela/chars.html>.

Lamport, L. 1986. *Latex: a document preparation system*. Reading, Mass.: Addison-Wesley

Leech, G., and Wilson, A. 1994. EAGLES morphosyntactic annotation. EAGLES report EAGSCSG/IR-T3.1. Pisa: Istituto di Linguistica Computazionale

Leech, G., Barnett, R., and Kahrel, P. 1995a. Guidelines for the standardization of syntactic annotation of corpora. In *EAGLES Document EAG-TCWG-SASG/1.8*.

Leech, G., Myers, G., and Thomas, J. eds. 1995b. *Spoken English on computer*. Harlow: Longman.

Leech, G., and Wilson, A. 1999. Standards for Tagsets. In *Syntactic Wordclass Tagging*, ed. Hans van Halteren, 55-80. Dordrecht.: Kluwer Academic.

Leech, G., and Weisser, M. 2003. Generic Speech Act Annotation for Task-Oriented Dialogue. In *Proceedings of the Corpus Linguistics 2003 Conference*, eds. D. Archer, P. Rayson, A. Wilson and A. McEnery. Lancaster: UCREL Technical Papers.

Lickley, R. HCRC Disfluency coding manual. <http://www.ling.ed.ac.uk/~robin/maptask/disfluency-coding.html>.

Marcus, M., Santorini, B., and Marcinkiewicz, M. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19:313-330.

Mengel, A., Dybkjaer, L., Garrido, J. M., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C. 2000. MATE Deliverable D 2.1. MATE Dialogue Annotation Guidelines. <http://www.andreasmengel.de/pubs/mdag.pdf>.

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

MICASE: Michigan Corpus of Academic Spoken English. <http://www.hti.umich.edu/m/micase/>.

Morton, A. 1986. Once. A test of authorship based on words which are not repeated in the sample. *Literary and Linguistic Computing* 1:1-8.

Pickering, B., Williams, B., and Knowles, G. 1996. Analysis of transcriber differences in the SEC. In *Working with Speech*, eds. G. Knowles, A. Wichmann and P. Alderson. London: Longman.

Perez-Parent, M. 2002. Collection, handling, and analysis of classroom recordings data: using the original acoustic signal as the primary source of evidence. *Reading Working Papers in Linguistics* 6:245-254. http://www.rdg.ac.uk/app_ling/wp6/perezparent.pdf.

- Pierrehumbert, J. 1980. The phonology and phonetics of English intonation. MIT.
- Roach, P., and Arnfield, S. 1995. Linking prosodic transcription to the time dimension. In *Spoken English on Computer*, eds. G. Leech, G. Myers and J. Thomas, 149-160. Harlow: Longman.
- Roe, P. 1977. *The notion of difficulty in scientific text*. University of Birmingham.
- Sampson, G. 1995. *English for the computer: the SUSANNE corpus and analytic scheme*. Oxford: Clarendon Press
- Scott, M. WordSmith Tools. <http://www.lexically.net/wordsmith/>.
- Searle, S. J. Unicode revisited. <http://tronweb.super-nova.co.jp/unicoderevisited.html>.
- Searle, S. J. 1999. A brief history of character codes in North America, Europe, and East Asia. <http://tronweb.super-nova.co.jp/characcodehist.html>.
- Semino, E., and Short, M. 2003. *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Narratives*. London: Routledge
- Short, M., Semino, E., and Culpeper, J. 1996. Using a corpus for stylistics research: speech and thought presentation. In *Using corpora for language research*, eds. J. Thomas and M. Short, 110-131. London: Longman
- Sinclair, J. 1982. Reflections on computer corpora in English language research. In *Computer corpora in English language research*, ed. Stig Johansson: 1-6. Bergen.
- Sinclair, J. 1989. Corpus creation. In *Language, learning and community*, eds. C Candlin and T McNamara, 25-33: NCELTR Macquarie University.
- Sinclair, J. ed. 1990. *Collins Cobuild English grammar*. London: Collins.
- Sinclair, J. 1991. *Corpus, concordance, collocation: Describing English language*. Oxford: Oxford University Press.
- Sinclair, J. 1995. From theory to practice. In *Spoken English on Computer*, eds. G. Leech, G. Myers and J. Thomas, 99-112. Harlow: Longman.
- Sinclair, J. 2001. Preface. In *Small corpus studies and ELT*, eds. Mohsen Ghadessy, Alex Henry and Robert L. Roseberry, vii-xv. Amsterdam/Philadelphia: John Benjamins.
- Sinclair, J. 2003. Corpora for lexicography. In *A practical guide to lexicography*, ed. P Van Sterkenberg. Amsterdam: John Benjamins.

Sinclair, J. 2004. Intuition and annotation - the discussion continues. In *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized corpora (ICAME 23). Göteborg 22-26 May 2002.*, eds. Karin Aijmer and Bengt Altenberg, 39-59. Amsterdam/New York: Rodopi. <http://www.ingentaconnect.com/content/rodopi/lang/2004/00000049/00000001/art00003>.

Smith, A. 2004. Preservation. In *A companion to Digital Humanities*, eds. S. Schreibman, R. Siemens and J. Unsworth, 576-591. Oxford: Blackwell.

Sperberg-McQueen, C. M., and Burnard, L. 1994. *Guidelines for electronic text encoding and interchange (TEI P3)*. Chicago & Oxford: ACH-ALLC-ACL Text Encoding Initiative.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, R., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. 2000. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics* 26:339-373.

Tapanainen, P., and Voutilainen, A. 1994. Tagging accurately - don't guess if you know. In *Proceedings of ANLP '94*, 47-52. Stuttgart.

Thompson, H., Anderson, A., and Bader, M. 1995. Publishing a spoken and written corpus on CD-ROM: the HCRC Map Task experience. In *Spoken English on Computer*, eds. G. Leech, G. Myers and J. Thomas, 168-182. Harlow: Longman.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work: Studies in corpus linguistics*, v. 6. Amsterdam: John Benjamins

UCREL: University Centre for Computer Corpus Research on Language. <http://www.comp.lancs.ac.uk/ucrel/>.

Unicode Consortium. 2003. *The Unicode standard, Version 4.0*. London: Addison-Wesley. <http://www.unicode.org/versions/Unicode4.0.0/>.

van den Heuvel, H., Boves, L., and Sanders, E. 2000. Validation of content and quality of existing SLR: overview and methodology. <http://www.spex.nl/validationcentre/d11v21.doc>.

Voutilainen, A., and Järvinen, T. 1995. Specifying a shallow grammatical representation for parsing purposes. In *Proceedings from the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 210-214: Association for Computational Linguistics.

Wells, J. C., Barry, W., Grice, M., Fourcin, A., and Gibbon, D. 1992. Standard computer-compatible transcription. Esprit project 2589 (SAM). In *Doc. no. SAM-UCL-037*. London: Phonetics and Linguistics Department, UCL.

Whistler, K. Why Unicode will work on the Internet. <http://slashdot.org/features/01/06/06/0132203.shtml>.

Working Group on Romanization Systems. United Nations Group of Experts on Geographical Names (UNGEGN). <http://www.eki.ee/wgrs/>.

Zipf, G. K. 1935. *The psychobiology of language*. New York: Houghton Mifflin.