which will have to be divided into sentence units. Both the tagged and the parsed texts will be concordanced.

10 Distribution

We agreed early on that distribution of computer disks and speech recordings will be entrusted to the Norwegian Computing Centre for the Humanities at Bergen, which is responsible for the International Computer Archive of Modern English (ICAME).

11 Prospects

ICE is a splendid example of international cooperation in English language research. The project will undoubtedly provide valuable information on the use of English in many countries, in most of which there have never been systematic studies, and it will provide the basis for international comparisons. It will stimulate insights into the sociolinguistics of English nationally and internationally, and offer data for sociolinguistic theory. The results of the project will have implications for the teaching of English and in some countries will be applied to language planning (cf. Greenbaum 1988b: 32–9). Phileas Fogg traversed the world in eighty days to win his wager. ICE encompasses the English-speaking world, but there is no deadline for its completion and no one is offering a wager. In any case, I must not press the analogy: Phileas Fogg never existed, whereas ICE has been conceived and is being propelled into existence.

References

Aarts, J. and van den Heuvel, T. (1985), 'Computational tools for the syntactic analysis of corpora', *Linguistics* 23, 303–35.

Garside, R., Leech, G. and Sampson, G. (eds) (1987), *The Computational Analysis of English: A Corpus-based Approach*, Longman.

Greenbaum, S. (1985), Commentator 1, in Quirk and Greenbaum (eds) 1985: 31–32.

Greenbaum, S. (1988a), 'A proposal for an international computerized corpus of English', *World English* 7, 315.

Greenbaum, S. (1988b), *Good English and the Grammarian*, Longman.

Kachru, B. (1985), 'The English language in a global context', in Quirk and Widdowson (eds) 1985: 11–30.

Leech, G. and Garside, R. (1991), 'Running a grammar factory: The production of syntactically analysed corpora or "treebanks"', in Johansson and Stenström (eds) 1991.

Oostdijk, N. (1988a), 'A corpus for studying linguistic variation', *ICAME Journal* 12, 3–14.

Schmied, J. (1989), 'Text categorization according to use and user and the International Corpus of English', *Computer Corpora des Englischen in Forschung, Lehre und Anwendungen* (*CCE Newsletter*) 3, 1–11.

24

CORPUS DESIGN CRITERIA

Sue Atkins, Jeremy Clear and Nicholas Ostler

Source: *Literary and Linguistic Computing*, 7:1 (1992), 1–16.

Abstract

'Corpus Design Criteria' begins (Section 1) by defining the object to be created, a corpus, and the constituents of it, texts themselves, noting briefly the pragmatic constraints on the sort of documents which will actually be available, spoken as well as written.

It then (Section 2) reviews the practical stages in the process of establishing a corpus, from selection of sources through to mark-up, assigning annotations to the texts assembled. This is followed by a consideration of copyright problems (Section 3).

Section 4 points out the major difficulties in defining the population of texts that the corpus will sample, contrasting the sets of texts received versus those produced by a target group, and internal (linguistic) versus external (social) means of defining such groups.

The next three sections look at the sets of markers which can be useful at different levels. Section 7 begins at the highest level, considering the different types of corpus there may be. Section 6 is intermediate, considering how to distinguish the different types of text occurring within a corpus. Then, for the intra-text level, Section 7 reviews considerations governing mark-up, distinguishing those markers useful for written and spoken texts. Of these three sections, Section 6 is the most fully explicit, listing twenty-nine significant attributes assignable to a text.

Sections 8 and 9 turn away from the corpus design itself, to focus on its social context and function, both of the corpus design process, and of the corpus when implemented: to what extent are there now accepted standards relevant to the criteria reviewed in preceding sections? And what are the major classes of potential users and uses for corpora, both now and in the future?

Introduction

There has been over the past few years a tremendous growth in interest and activity in the area of corpus building and analysis. European, USA, and Japanese efforts in the development of NLP and IT are converging on the recognition of the importance of some sort of corpus-based research as part of the infrastructure for the development of advanced language processing applications. Statistical processing of text corpora has been demonstrated as a viable approach to some of the traditional hard problems of computational linguistics, machine translation, and knowledge engineering.

Our aim in this paper is to identify the principal aspects of corpus creation and the major decisions to be made when creating an electronic text corpus, for whatever purpose; and to discuss in detail the criteria that must inform some of these decisions and that are relevant to the establishment of basic standards for corpus design. We offer this paper as a step towards an in-depth study of corpus design criteria, with the object of defining the minimal set necessary to foster the creation of high-quality compatible corpora of different languages, for different purposes, in different locations, and using different types of software and hardware.

To this end, we attempt to identify the principal features in corpus design, and to note others which must not be forgotten but which need not be addressed in the initial stages of corpus building. Our aim is not to make a comprehensive and watertight listing of everything it is possible to decide, for we believe that this would be totally counter-productive: rather, a corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing, living language. Our aim is simply to pick out the principal decision points, and to describe the options open to the corpus-builder at each of these points.

We specifically exclude from this paper all consideration of acoustic corpora, which falls outside our brief; however, this is not to imply that research into spoken language is any less important; simply that the needs of the two groups are different, making it possible to consider these two aspects of NLP separately. What we say in this paper relates to the needs of speech research only in so far as any general text corpus must include as high a proportion as possible of transcribed spoken text, which is included in our understanding of the terms 'text' or 'lexical material'.

1 Defining text collections and a unit of text

We distinguish four types of text collection,[1] which we find helpful and urge the community to accept.

Archive: a repository of readable electronic texts not linked in any coordinated way, e.g. the Oxford Text Archive.

Electronic text library (or ETL, Fr. 'textothèque'): a collection of electronic texts in standardized format with certain conventions relating to content, etc., but without rigorous selectional constraints.

Corpus: a subset of an ETL, built according to explicit design criteria for a specific purpose, e.g. the Corpus Révolutionnaire (Bibliothèque Beaubourg, Paris), the Cobuild Corpus, the Longman/Lancaster corpus, the Oxford Pilot corpus.

Subcorpus: a subset of a corpus, either a static component of a complex corpus or a dynamic selection from a corpus during on-line analysis.

This paper is concerned with ETLs, corpora, and subcorpora but for the sake of brevity we use the word corpus to refer to all three types of collection.

1.1 Texts

A corpus which is designed to constitute a representative sample of a defined language type will be concerned with the sampling of *texts*. For the purposes of studying spoken language in transcription (not speech *per se*) it is convenient to use the term 'text' to include transcribed speech. The use of the word to describe a unit of text, informally considered to be integral in some way, raises some issues of definition for the corpus builder. If, for example, it is decided that the corpus should include 'full texts', or if sampling is to be carried out by some random selection of texts, then it will be important to consider what is meant by a text. Since the notion of text is derived strongly from models of written language, and is more tenuously applied to speech events, we should consider the definition separately for these two modes.

1.1.1 Written texts

The printed monograph or work of narrative fiction is the model for the notion of a 'text'. It manifests several apparently criterial characteristics:

- it is discursive and typically at least several pages long;
- it is integral;
- it is the conscious product of a unified authorial effort;
- it is stylistically homogeneous.

'Texts' are often assumed to be a series of coherent sentences and paragraphs. By integral, we mean that a printed book usually has a beginning, middle, and end and is considered to be complete in itself. Even though a book may have more than one author, their collaboration usually counts as authorial unity (no particular parts of the book are ascribed to any one individual) and the result is usually a unit of stylistically consistent language.

The important aspects in these criterial features for the corpus builder are those relating to stylistics and text linguistics. The designer of the corpus will wish to neutralize as far as possible the effects of sampling bias and the stylistic idiosyncrasies of one particular author can be reduced in significance if texts by many different authors are included. The need to control stylistic parameters leads to the concern with a unified authorial effort and consistent style. Similarly, if the corpus is to provide the basis for studies of cohesion, discourse analysis, and text linguistics—all linguistic patterning beyond the sentence or paragraph—then the integrity of the samples as textual units ought to be taken into consideration.

Novels fit the above-mentioned schema very neatly and are prototypical 'texts', but there are many types of written language which are more problematic. Listed below are some examples of language units which are likely to be incorporated in a general corpus and which illustrate typical deviations from the model instance.

Small ads in newspapers where the corpus builder might prefer to make a *collection* of these small ads and treat this as one text.
An article in a newspaper or magazine: it may be convenient to treat one issue of a newspaper (and single issues of other periodicals) as one text.
An academic Festschrift, learned journal, etc., where the bibliographic data applies to the whole book but the papers differ linguistically to such an extent that they might be best treated as discrete texts.
A poem. It is often more convenient to gather many short poems by the same author into collections and to treat each collection as a text.
Published correspondence in which the letters are the product of two authors, but they constitute a single discourse. The critical apparatus, introduction, and editorial material will be yet another author's intervention. The way this problem is handled will depend on the requirements of the corpus.

1.1.2 Spoken texts

The difficulty and high cost of recording and transcribing natural speech events lead the corpus linguist to adopt a more open strategy in collecting spoken language. Practical considerations may make it more acceptable to the corpus builder to capture *fragments* of speech than fragments of writing. A stretch of speech can be thought of as forming a text, rather than a fragment, if one of the following conditions applies:

- the speech unit starts when the participants come together and ends when they part (e.g. telephone conversation);
- the speech has an obvious opening and closing (e.g. a lecture).

Some examples of units of speech which might be considered to be texts are:

An informal face-to-face conversation.
A telephone conversation.
A lecture.
A meeting.
An interview.
A debate.

2 Stages in corpus building

The planning stage of corpus building will aim to arrive at specifications in two related areas: the linguistic design of the corpus and the planning of the project as a whole.

The administrative planning will be concerned with budget, timing, and implementation of corpus design, and with the costs and the stages of the corpus building work. The principal stages are:

Specifications and design.
Hardware and software.
Data capture and mark-up.
Corpus processing.
Corpus growth and feedback.

2.1 Specifications and design

The linguistic design will need to establish at the least what type of corpus is being constructed (see Section 5, Corpus Typology), taking into account *inter alia* the sizes of the text samples to be included, the range of language varieties (synchronic), and the time period (diachronic) to be sampled, whether to include writing and speech and the approximate level of encoding detail to be recorded in electronic form.

The design will involve consultation with representatives of the anticipated users of the corpus and also perhaps with other specialists. For a corpus which aims to represent general language synchronically, a sociolinguist may be called upon. Decisions concerning the sampling strategy may involve a statistician and statistical expertise will almost certainly be valuable to some extent throughout the corpus building project. Linguistic expertise will also be required to ensure that all decisions concerning the design, balance, encoding, and processing of the corpus are appropriate to the linguistic aims of the corpus project. It is because of the particular *linguistic* interest in a large body of computerized text that a language corpus is quite different from any of the very large on-line information databases which are available (commercially and for research purposes) around the world.

As a result of the consultation process a preliminary (or even definitive) selection of text sources will be made. The selection of sources might be based on a systematic analysis of the target population or on a random selection method. Alternatively, the need for large volumes of data may lead one to adopt a more opportunistic approach to the collection of text.

2.2 *Hardware and software*

Estimates will be required of the hardware and software needs of the corpus project. These will depend to a large extent on the size of the corpus and the amount of processing and manipulation which is to be carried out on it. Merely to store a corpus will require relatively modest computing equipment, but to set up an indexing and retrieval system for that corpus and to allow for further analytic processing will demand much greater disk storage and processing power.

2.3 *Data capture and mark-up*

Data capture is also time-consuming and the costs incurred are unavoidable and determined to a large extent by the amount of text to be captured. Printed material can be scanned by OCR devices with appropriate software. This method of conversion from print to electronic form is becoming more cost effective each year as OCR technology improves. Alternatively, text can be keyboarded manually. For transcription of audio recordings and for the capture of degraded or complex printed material keyboarding is the only option.

Text may also be acquired in machine-readable form (either having been captured by someone else or having originated on computer). This eliminates the costs of data capture but may require time-consuming parsing, reformatting, and restructuring to bring the data into line with the corpus conventions for encoding and mark-up.

The texts in the corpus will probably (though not necessarily) be marked up with embedded codes to signal elements of structure and to record features of the original text source. This is an important issue and one which will have implications for the scheduling and costing of a corpus project (see Section 7, Mark-up). There is a growing consensus that SGML provides a suitable basis for a standard mark-up scheme for texts held on computer, and the Text Encoding Initiative (TEI) published in July 1990 its draft guidelines for the encoding and interchange of machine-readable texts.

Whatever method of data capture is adopted, the text will require some degree of validation and error correction to ensure that it is reasonably accurate and consistent with the encoding conventions for the corpus.

2.4 *Corpus processing*

The mere existence of a large corpus will not satisfy demand for linguistic data. A set of general tools for processing the corpus will be essential. Many such tools already exist and are in use, but they are often designed to meet very specific local needs and there is work to be done on agreeing on standard formats for data derived from a corpus, for word-class tagging, parse tree notations, semantic labelling, etc. KWIC concordances, word frequency lists, collocation statistics will be basic requirements and software to perform these basic tasks can be made widely available as part of the corpus package. Tagging software and parsers will be more difficult to implement and their design more contentious, but these tools should also be available.

2.4.1 *Basic tools*

These might include:

Word frequency: Software to produce lists of word types and their frequency in the corpus and perhaps also some statistical profile of the relation of types to tokens in the corpus, indications of the distribution of types across the text categories, and graphical displays to summarize these lists in a form which can be assimilated by the user of the corpus.
Concordancing: Text retrieval and indexing software (of the kind already provided by such packages as WordCruncher, OCP, PAT, TACT, Free Text Browser, SGML Search, etc.) with features appropriate for linguistic analysis.
Interactive searching: Flexibility in search and display/presentation.

2.4.2 *Advanced text handling*

In order to allow more sophisticated statistical analyses to be carried out on a large corpus, it may be useful to implement a number of more advanced text processing tools which can automatically process linguistic information in a corpus. Such software might include:

Lemmatization to relate a particular inflected form of a word to its base form or lemma, thereby enabling the production of frequency and distribution figures which are less sensitive to the incidence of surface strings.
Part-of-speech labelling (sometimes called 'tagging') to assign a word class or part-of-speech label to every word. This allows simple syntactic searches to be carried out.
Parsing to assign a fully labelled syntactic tree or bracketing of constituents to sentences of the corpus.

Collocation to compute the statistical association of word forms in the text.
Sense disambiguation ('homograph separation') to distinguish which of several possible senses of a given word is being employed at each occurrence. Early work in this area shows promising results by methods based on look-up in a machine-readable dictionary or on identifying collocational patterns.
Link to lexical database to integrate the instances of words or phrases in a corpus with a structured lexical database which records detailed morphological, syntactic, lexical, and semantic information about words.

2.5 Corpus growth and feedback

In order to approach a 'balanced' corpus, it is practical to adopt a method of successive approximations. First, the corpus builder attempts to create a representative corpus. Then this corpus is used and analysed and its strengths and weaknesses identified and reported. In the light of this experience and feedback the corpus is enhanced by the addition or deletion of material and the cycle is repeated continually.

There is a need therefore for an appropriate mechanism to allow users of the corpus to relay their findings, comments, warnings, and so on back to the corpus development team. This might be done electronically over a network or via e-mail or through regular, scheduled meetings and discussions between the users and those responsible for the maintenance and development of the corpus. Statistical expertise is likely to be valuable in assessing and testing methods for improving the balance of the corpus to suit the needs of the users.

3 Corpora and copyright

One of the serious constraints on the development of large text corpora and their widespread use is national and international copyright legislation. In most cases, copyright permission will need to be obtained for texts to be computerized. However, the copyright status of any text is often unclear, as is the degree of privilege in this area attaching to university research.

It is necessary and sensible to protect, through copyright laws, the rights of authors and publishers in texts that they create. In response to the rapid development of computer technology in the areas of networking, DTP, personal computing and electronic publishing, copyright legislation is being extended and revised to cover what are perceived to be new threats, as well as opportunities, for the writing and publishing industries. The effect for the corpus builder is that it is quite likely that any text (or sample of text) which is to be computerized and included in a corpus will be under copyright protection and that permission will have to be obtained for its use. The following considerations are relevant to copyright and corpora:

Is the text protected by copyright? National legislation varies but in general texts can pass out of copyright after a certain time period. However, in the US and UK copyright may reside in a particular *edition* of a text (the arrangement and typesetting) even though the original text itself is out of copyright. The computerization of speech transcriptions may also require permission—particularly if the speech is recorded from radio or TV broadcasts.

Will payments be offered? Within the publishing world, copyright permissions are usually obtained on payment of a fee, a royalty, or some combination. It is clear that payment of even modest fees will make the compilation of a corpus of contemporary texts from a large number of different sources so expensive that only a very few organizations will be able to justify the costs. If no fee is to be paid then the copyright holders must be assured that the compilation of a language corpus is no threat whatsoever to the revenue-earning potential of the text and that no direct commercial exploitation is to be made of the corpus.

What use is to be made of the corpus? If the corpus is to be used for commercial purposes then these will probably need to be clearly stated and defined to the copyright holder. Similarly, if the corpus builder plans to copy and distribute the corpus to other people or to make it accessible by others, this will need careful agreement with the copyright holder.

How many different sources are to be collected? The use of a renewable, single source, such as the AP Wire Service, requires permission from only one source with perhaps only one fee. If the corpus is to contain many hundreds or thousands of text samples, the administrative and clerical task of identifying copyright holders and obtaining permission can be very substantial.

4 Population and sampling

In building a natural language corpus one would like ideally to adhere to the theoretical principles of statistic sampling and inference.[2] Unfortunately, the standard approaches to statistical sampling are hardly applicable to building a language corpus. First, it is very difficult (often impossible) to delimit the total population in any rigorous way. Textbooks on statistical methods almost always focus on clearly defined populations. Secondly, even if the population could be delimited, because of the sheer size of the population and given current and foreseeable resources, it will always be possible to demonstrate that some feature of the population is not adequately represented in the sample. Thirdly, there is no obvious unit of language (words? sentences? texts?) which is to be sampled and which can be used to define the population. We may sample words or sentences or 'texts' among other things. Despite these difficulties, some practical basis for progress can be established. An approach suggested by Woods, Fletcher, and Hughes[3] is to accept the results of each study as though any sampling had been carried out in the theoretically 'correct' way, to attempt to foresee possible objections. In corpus linguistics such a pragmatic approach seems the only course of action. Moreover, there is a tendency to overstate the possibility and effects of experimental error: indeed, good scientific estimation of the possibility and scale of experimental error in statistics of natural language corpora is seldom carried out at all.

All samples are *biased* in some way. Indeed the sampling problem is precisely that a corpus is inevitably biased in some respects. The corpus users must continually evaluate the results drawn from their studies and should be encouraged to report them (see Subsection 2.5).

The difficulty of drawing firm conclusions when the number of observed instances is few underlines the methodological point made by Woods, Fletcher, and Hughes: that researchers should question how the sample was obtained and assess whether this is likely to have a bearing on the validity of the conclusions reached.

4.1 Defining the population

When a corpus is being set up as a sample with the intention that observation of the sample will allow us to make generalizations about language, then the relationship between the sample and the target population is very important. The more highly specialized the language to be sampled in the corpus, the fewer will be the problems in defining the texts to be sampled. For a general-language corpus, however, there is a primary decision to be made about whether to sample the language that people hear and read (their *reception*) or the language that they speak and write (their *production*).

Defining the population in terms of language reception assigns tremendous weight to a tiny proportion of the writers and speakers whose language output is received by a very wide audience through the media. However, most linguists would reject the suggestion that the language of the daily tabloid newspapers (though they may have a very wide reception) can be taken to represent the language production of any individual member of the speech community.

The corpus builder has to remain aware of the reception and production aspects, and though texts which have a wide reception are by definition easier to come by, if the corpus is to be a true reflection of native speaker usage, then every effort must be made to include as much production material as possible. For a large proportion of the language community, writing (certainly any extended composition) is a rare language activity. Judged on either of these scales, private conversation merits inclusion as a significant component of a representative general language corpus. Judged in terms of production, personal and business correspondence and other informal written communications form a valuable contribution to the corpus.

To summarize, we can define the language to be sampled in terms of language production (many producers each with few receivers) and language reception (few producers but each with many receivers). Production is likely to be greatly influenced by reception, but technically only production defines the language variety under investigation. However, collection of a representative sample of total language production is not feasible. The compiler of a general language corpus will have to evaluate text samples on the basis of *both* reception and production.

4.2 Describing the population

A distinction between external and internal criteria is of particular importance for constructing a corpus for linguistic analysis. The internal criteria are those which are essentially *linguistic*: for example, to classify a text as formal/informal is to classify it according to its linguistic characteristics (lexis/diction and syntax). External criteria are those which are essentially *nonlinguistic*. Section 6 contains a list of attributes which we consider relevant to the description of the language population from which corpus texts are to be sampled. These attributes, however, are all founded upon extra-linguistic features of texts (external evidence). Of course, the internal criteria are not independent of the external ones and the interrelation between them is one of the areas of study for which a corpus is of primary value. In general, external criteria can be determined without reading the text in question, thereby ensuring that no linguistic judgements are being made. The initial selection of texts for inclusion in a corpus will inevitably be based on external evidence primarily. Once the text is captured and subject to analysis there will be a range of linguistic features of the text which will contribute to its characterization in terms of internal evidence.[4] A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation. A corpus selected entirely on external criteria would be liable to miss significant variation among texts since its categories are not motivated by textual (but by contextual) factors.

5 Corpus typology

A corpus is a body of text assembled according to explicit design criteria (see Section 6 below) for a specific purpose, and therefore the rich variety of corpora reflects the diversity of their designers' objectives.

It is worth mentioning in parenthesis that the text typology discussed in Section 6 is, in many cases, also relevant to corpus typology, in that corpora may be classified according to text types if they consist solely of texts of one single type. Thus, if the corpus is created for the purpose of studying one single MODE, then one may have a SPOKEN or a WRITTEN corpus; similarly, if only one MEDIUM is of interest, one may have A BOOK or a NEWSPAPER or a CLASSROOM LESSON corpus.

In this section, however, our purpose is to outline certain contrastive parameters of corpus typology *per se*:

1 Types: FULL TEXT SAMPLE MONITOR
Notes: For Full Text: each text in the corpus is unabridged.
For Sample: sample size to be defined, also location of sample within full text and method of selection of samples.
For Monitor: texts scanned on continuing basis, 'filtered' to extract data for database, but not permanently archived. (Clear (1988), Sinclair (1982).)

2 Types: CLOSED OPEN-ENDED

3 Types: SYNCHRONIC DIACHRONIC
Notes: A specific period must be designated for a synchronic corpus; this requires research into how low that period may be if the corpus is to be considered synchronic.

4 Types: GENERAL TERMINOLOGICAL
Notes: Terminologists must define conditions which must obtain if a corpus is to be valid for terminological use, this in terms no doubt of the text typology (see Section 6 below).

5 Types: MONOLINGUAL BILINGUAL PLURILINGUAL

6 Types: *LANGUAGE(S) OF CORPUS*
Notes: English, Russian, Japanese, . . .

7 Types: SINGLE PARALLEL-2 PARALLEL-3 . . .
Notes: Are all the texts in the corpus stand-alone or part of a parallel pair/trio, etc., of translated texts? (This applies only to bi- or plurilingual corpora.)

8 Types: CENTRAL SHELL
Notes: The central corpus is a selected body of texts, of manageable size, big enough for normal purposes. The shell, which may be the remainder of the ETL, is available for access when necessary.

9 Types: CORE PERIPHERY
Notes: These concepts are discussed by Leitner[5] in relation to the ICE. The 'core' contains text types common to all varieties of English, and therefore present in all the subcorpora; while the 'periphery' contains text types specific to some subcorpora only.

6 Text typology

There is much talk of a 'balanced corpus' as a *sine qua non* of corpus analysis work: by 'balanced corpus' is meant (apparently) a corpus so finely tuned that it offers a manageably small scale model of the linguistic material which the corpus builders wish to study. At present, corpus 'balance' relies heavily on intuition, although work on text typology is highly relevant. It is not our purpose to lay down a methodology for 'balancing' a corpus, for we believe that this is not something that can be done in advance of the corpus-building process, if it can be done at all. Controlling the 'balance' of a corpus is something which may be undertaken only after the corpus (or at least an initial provisional corpus) has been built; it depends on feedback from the corpus users, who as they study the data will come to appreciate the strengths of the corpus and be aware of its specific weaknesses.

In our ten years' experience of analysing corpus material for lexicographical purposes, we have found any corpus—however 'unbalanced'—to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts. It would be shortsighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as 'unreliable' or 'irrelevant' simply because the corpus used cannot be proved to be 'balanced'. It should also be noted that recording attributes of texts is very labour-intensive and an overambitious plan could turn out to be unsustainable. Experience teaches us that it is better to aim to record initially an essential set of attributes and values which may later be expanded if resources permit.

The significant variables considered here, in the context of corpus design criteria, are all extra-linguistic. We believe, however, that it is impossible to 'balance' a corpus on the basis of extra-linguistic features alone. Diagnosis of imbalance must come from an analysis of internal evidence. All that the corpus-builder can do is to try not to skew a corpus too much in any direction. Balancing it, or at least reducing the skew, is something which comes along much later, and will demand information on both linguistic and extra-linguistic features in the corpus.

When creating a corpus for a specific purpose, the corpus designer must be able to make principled choices at all the major decision points. Information on features fundamental to corpus design must therefore be recorded on each of the texts in the electronic text library from which the corpus is to be selected. We suggest that the concept of a set of features is relevant to an ETL, while that of a taxonomy proper relates to a corpus created for a specific purpose. See Kučera and Francis (1964), Johansson, Leech, and Goodluck (1978), Leitner (1990), Oostdijk (1988a,b), Engwall (forthcoming) as examples of the earlier treatment of text typology.

Looking at features relevant to the typology of texts, we identify extra-linguistic variables which are of interest to anyone establishing an ETL or a corpus; we propose that these variables may be considered criterial attributes in the context of ETL/corpus design, and we try to indicate the minimum level of detail which seems to us essential to record. We exemplify the set (usually open-ended) of values of these attributes, and for some consider the type of criteria that may be applied in the process of identifying the features of a text in an ETL or a corpus. Naturally, the text attribute features proposed below are not all independent of each other.

6.1 Text attributes

This section consists of:

1 A list of attributes that may be recorded for every text introduced into the text collection.
2 A note regarding criteria for assignment of values. (Some criteria are noted as "to be defined". Our purpose in this discussion paper is to record the need for such definition; much detailed work will be required before the criteria can be established.)
3 An example of acceptable values and an indication of a reasonable level of depth to be recorded. (Values in capitals are, we believe, essential to note where they differ from a default value; those in roman type are highly desirable but not essential in the first instance—insistence on this degree of detail could be counter-productive.)
4 Where appropriate, some notes on the value-assignment criteria.

1 Attribute: MODE
Criterion: Mode of delivery of the original contents of the text: text transcribed from an audio/video recording is 'spoken'; text written as dialogue is 'written to be spoken'; poetry poses a problem (written to be read or spoken?).
Values: WRITTEN (default)
written-to-be-read (default)
written-to-be-spoken
SPOKEN
spoken-to-be-written

2 Attribute: TEXT ORIGIN
Criterion: Is the text a product of a collaborative effort?
Values: single
several
joint
corporate
mixed
unclassified (default)
Notes: A single text is always the product of one individual. A 'several' text contains sections attributable to different people, while a joint text is unified. A corporate text is a production of an unnamed collective (e.g. a company brochure). Mixed texts contain parts which are any or all of the above.

3 Attribute: PARTICIPATION
Criterion: number of people originating the text
Values: 1-PERSON (default)
multi-person

4 Attribute: CONSTITUTION
Criterion: Is the text single or composite? One integral text by one author is single; a newspaper, journal, collection of essays, or textbook, made up of many distinct small texts which could each be classified individually, is composite.
Values: SINGLE (default) COMPOSITE

5 Attribute: PREPAREDNESS
Criterion: Still to be defined in detail.
Values: PREPARED (default)
from-notes
spontaneous
Note: This is a cline and the points along it may be set according to the needs of the corpus.

This parameter subsumes the notion of timed versus non-timed (e.g. an essay written under examination conditions), which some corpus-builders have considered separately.

6 Attribute: MEDIUM
Criterion: Medium of original contents of the text; all written text, whether published or private, is TEXT; the others are mostly self-evident, though criteria for assigning values to problematic cases such as films (scripts or soundtracks) need clarification.
Values: (*Writing*)
PRINT (default)
BOOK
PERIODICAL
EPHEMERA
manuscript
(*Speech*)
distant
direct
TV
talk (lecture/speech, etc.)
radio
entertainment (theatre, etc.)
telephone
person-to-person

7 Attribute: STYLE
Criterion: Surface features of text or author's claims
Values: PROSE (default)
VERSE

blank verse
rhyme sonnet
. . .

8 Attribute: GENRE
Criterion: This set is open-ended, and culturally specific.
Values: (*Writing*)
NOVEL
SHORT STORY
PLAY
POEM
ESSAY
LETTER
business
personal

(*Speech*)
lecture
debate/discussion
speech
conversation
demonstration
classroom lesson

(*Written or Spoken*)
advertisement
regulation/law
article
advice column
horoscope
announcement

examinations
report
commentary
feature
advice programme

Notes: There are many problem areas (how long is a short story? how much of TV news programmes is report and how much commentary? etc.). Also, attaching values to the often tiny pieces of text that go to make up the whole of a newspaper or periodical is labour-intensive; but see Subsection 1.1 above. The values may need to be adjusted according to the cultural context.

9 Attribute: FACTUALITY
Criterion: To be defined: often deducible from authorial claims
Values: FACT faction FICTION (cline)
Notes: The truth or falsehood of the content is irrelevant to FACTUALITY.

10 Attribute: SETTING
Criterion: In what social context does the text belong?
Values: UNCLASSIFIED home
education
work
leisure
public affairs

11 Attribute: FUNCTION
Criterion: Illocutionary force.
Values: UNMARKED (default)
narrative
informative

expository
hortatory/persuasive
regulatory/instructional
reflective (=giving one's opinion) etc.
entertaining

12 Attribute: TOPIC
Criterion: 'This text is about X', or 'The subject of this text is (related to) X'
Values: GENERAL (default)
science
biology
chemistry
etc. etc.

music
orchestral
opera

animals
dogs

Note: It is necessary to draw up a list of major topics and subtopics in the literature. Library science provides a number of approaches to topic classification.

13 Attribute: TECHNICALITY
Criterion: Based on degree of specialist/technical knowledge of the author and target readership/audience.
Values: GENERAL (default)
(non-specialist author and target)
TECHNICAL
(specialist author and specialist target)
semi-technical
(specialist author, general target)
Note: These must be external variables, not linguistic style variables. This particular attribute is highly important for terminological corpora.

14 Attribute: DATE
Value: DATE OF PUBLICATION (default) or date of speech event
Note: For some corpora it will be desirable to note the date of first publication, in case of revised editions.

15 Attribute: TEXT STATUS
Criterion: Is the text in its first appearance? Or a reprint? Or a 'new' or 'revised' or 'updated' edition?
Values: ORIGINAL/REPRINT (default)
updated
revised, etc.
Notes: This attribute is of particular importance in terminological corpora but also of interest in literary corpora.

16 Attribute: LANGUAGE
Criterion: Usually self-evident, but points of dispute may arise over the status of dialects.
Value: English, French, Japanese . . .

17 Attribute: LANGUAGE LINKS
Criterion: Self-evident: is the text stand-alone or part of a parallel pair/trio, etc., of translated texts?
Values: SINGLE (default)
PARALLEL-2
PARALLEL-3 (etc.)

18 Attribute: LANGUAGE STATUS
Criterion: Is the text the source language or is it a translation?
Values: SOURCE (default) TRANSLATION

19 Attribute: METHODOLOGY FOLLOWED
Criterion: To be defined: depends on assessment or knowledge of whether this is concept-based, whether it is standardized or officially approved, what research principles were applied.
Values: To be determined: possibly on a graduated scale?
Notes: This attribute of importance in terminological corpora. The criteria and values must be discussed with a terminologist.

20 Attribute: AUTHORSHIP
Value: NAME OF AUTHOR(S)

21 Attribute: SEX OF AUTHOR(S)
Criterion: self-evident
Values: MALE FEMALE

22 Attribute: AGE OF AUTHOR(S)
Criterion: self-evident
Values: the actual age in years

23 Attribute: REGION OF AUTHOR(S)
Criterion: To be defined: this attribute relates to the regional type of the language of the author(s).
Values: STANDARD (default) (for those languages where applicable: other values will depend on specific language: e.g. for English—UK, USA, Canada . . . , or English, Scottish, Welsh . . .).
Notes: Is a variety of language to be described as (e.g.) 'Irish' if the AUTHOR(S) spent the first twenty years of life in Ireland and the last fifty in the US?, etc. This is a parameter which will be refined by internal evidence.

24 Attribute: NATIONALITY OF AUTHOR(S)
Criterion: Mainly self-evident.
Values: Actual nationality at time of writing/speaking.

25 Attribute: NATIVE LANGUAGE OF AUTHOR(S)
Criterion: To be defined. May depend on detailed biographical information being available about the author(s).
Values: actual language, if known
default = language of text
Notes: This attribute of particular importance in terminological corpora. Precise information about first language is in many cases unavailable or irrelevant.

26 Attribute: AUTHORITY OF AUTHOR(S)
Criterion: To be defined. This relates to credibility of author(s) and authority on subject-matter; determination of this will depend on qualifications, experience, etc.
Values: Probably a point on a cline, of a granularity to support corpus uses.
Notes: This attribute is of importance in terminological corpora.

27 Attribute: AGE OF INTENDED READERSHIP
Criterion: To be defined.
Values: ADULT (default)
CHILD
teenage
middle-aged
elderly
Notes: The values can be made more or less granular, depending upon the amount of information that is available.

28 Attribute: SIZE OF INTENDED READERSHIP
Criterion: To be defined.
Values: A scale:
an individual
a small group of people
a large but homogeneous group
a large group geographically defined
unrestricted and heterogeneous
Notes: This is independent of the actual readership figures. So, for example, a novel has an intended readership size which is large, even though it may sell only very few copies.

29 Attribute: FAMILIARITY WITH INTENDED READERSHIP
Criterion: To be defined.
Values: A scale:
very familiar (close friend or family)
familiar (colleague)
known by name (acquaintance)
known only in terms of lifestyle, profession, etc.
completely unknown
Notes: This attribute may seem to overlap substantially with the preceding one, since an unrestricted heterogeneous mass of people cannot practically be familiar to the author(s). Nevertheless, these attributes can vary independently to some extent.

6.2 Reference coding

The set of data which is to be recorded for each text sample included in a corpus may well become fairly large. It addition to the attribute values presented above, it may include discursive information about the editorial decisions taken when the text was captured and validated or other information which will help future users of the corpus understand the precise nature of each sample. This data can be conveniently stored separately from the actual textual material of the corpus, but this will require an adequate reference system to relate any word in the corpus back to its location in the original text and to the associated corpus information. If the corpus is made up entirely of published printed texts, then a reference system based upon standard bibliographic citations and page numbering will probably be appropriate. However, for transcribed speech, printed ephemera, manuscript material, private or business correspondence, and similar material some other method must be used to refer from the electronic form of the corpus to the original. In some projects this may not be necessary and a simple text-referencing system may be adequate. In others where, for example, audio and/or video recordings accompany the computer corpus texts, a sophisticated encoding will be necessary to allow the user to consult the original or determine exactly where a given stretch of discourse is located. For many corpus applications sentence number will be a suitable identifying unit within each text sample.

A useful accompaniment to a reference coding system in a corpus is the collection of original sources (either paper or audio recording) for the computerized texts, to which corpus users can refer if necessary.

7 Mark-up

Mark-up, here, means introducing into the text, by means of some conventional set of readable labels or tags, indicators of such text features as, for example, chapter, paragraph, and sentence boundaries, headings and titles, various types of hyphenation, printers' marks, hesitations, utterance boundaries, etc.

7.1 Methodological considerations

7.1.1 Converting written material

The Text Encoding Initiative has already published draft guidelines dealing with the mark-up of machine-readable texts for interchange, and proposing a mark-up which achieves a high level of detail and descriptive generality.[6] The cost of introducing a sophisticated mark-up is high in terms of manual effort, and the cost/benefit balance may be badly upset unless it can be justified by worthwhile gains in ease of processing, accuracy, and reusability. What is needed in order to build up a large and useful corpus is a level of mark-up which maximizes the utility value of the text without incurring unacceptable penalties in the cost and time required to capture the data.

There are two approaches which can be taken in marking up texts. First the *descriptive*, in which one tags the underlying structural features of a text (the sentences, paragraphs, sections, footnotes, etc.). These are usually signalled by spacing, punctuation, type font and size shifts, and so on, but there is no one-to-one correspondence between features and realization. Secondly, the *presentational*, in which one tags these typographical features themselves.

The two types of mark-up are not mutually exclusive categories; they describe the two ends of a scale. In some instances it will be important to 'record what's there on the page', since the researcher will be concerned to discover patterns of usage relating to features of punctuation and spelling. In other cases, the computational considerations relating to the processing of the text in indexing and free-text retrieval software are likely to take precedence and a descriptive approach will be preferred.

In the case of sources which need to be converted from printed form, mark-up will have to be introduced. This could be carried out during the process of OCR scanning, during keyboarding, or as a post-editing operation once the plain text has been captured. In the case of data which comes from other computer systems, there is often some explicit descriptive encoding of text structure, often arbitrarily intermingled with presentational codes, which can be parsed and converted automatically into SGML format. In both cases there is a potentially large hidden cost involved in introducing, converting, and standardizing mark-up by program, even though this appears at first to be significantly more cost effective than tedious and costly human editing. The programming work required to normalize the encoding and mark-up must often be repeated for each new text, because there is very little standardization in printing format or in word-processing packages.

When adding a new text to a corpus under construction, one has to make the decision as to whether it is an instance of a predefined general text type, or whether it deserves separate treatment with different mark-up. That is, the SGML mark-up will require the specification of a set of document type definitions (DTDs) which formally define the structure which is to be marked up for each document type. The advantage of creating more DTDs is that the mark-up will better reflect the structure and content of each of a diverse collection of text samples. The disadvantage in creating more DTDs is that the marking up of a large number of different texts from many different sources becomes increasingly complex, and the generalizing power of the mark-up is diminished. The increased complexity of the mark-up may not be justified, however, if most of the processing on the corpus is to be done over large aggregates of text rather than individual samples.

7.1.2 Transcription of speech

The TEI has not yet published any guidelines relating to SGML mark-up for speech in transcription. One reason for this may be that gathering and transcribing authentic speech is quite a different operation from handling written documents and is clearly a much more specialist area of activity, for linguists, lexicographers, and speech technologists.

This section does not address the problems of the design of acoustic corpora—corpora intended to assist in the analysis of the physical characteristics of speech—in which we have no specialist expertise. We assume here that spoken language is to be collected in large quantities in order to form the basis for quantitative studies of morphology, lexis, syntax, pragmatics, etc.

Many media organizations or research establishments will be able to supply paper or machine-readable transcriptions, which can be processed to bring them into conformance, as far as possible, with a standard mark-up. Transcriptions made by non-specialists will typically be in the form of quasi-written text. That is, there will be recognizable sentences and punctuation, with a high degree of normalization of false starts, hesitation, non-verbal signals, and other speech phenomena. This type of transcription converts spoken language into a form of idealized 'script' (like a screenplay or drama script) which conforms to many of the established conventions of written English.[7] Unless the corpus is intended to serve the needs of speech specialists, then the usefulness of a 'script' transcription is sufficient for a wide variety of linguistic studies. The advantages of transcribing in this way are:

- the cost and time of transcription are minimized
- the transcription is easily readable without any special training
- the transcription can be processed using established and widely available text processing software without substantial pre-editing.

7.2 Features for mark-up: written text

Non-ascii characters: The number of such characters which may need encoding could be quite large. These might be encoded at the time of data capture and for the sake of economy might be recorded using any local convention which ensures that the codes are unambiguous yet easily keyed and checked. They can be expanded into standard SGML entity references by a search-and-replace operation carried out later. In many cases texts which are received already in machine-readable form will include special codes for graphic shapes which are not specified as part of the ascii set. These will have to be identified and standardized.

Quotation: This is an important aspect of text encoding in a corpus. There are three types of quotation that need to be considered:

- direct speech
- block quotes
- other uses of quotation marks.

Direct speech is probably the most difficult of these to deal with satisfactorily. In order to tag or parse direct speech accurately, it would be necessary to distinguish the multiple levels of *subsidiary* and *primary* discourse, each with its own syntactic structure. Block quotes are much more easily handled. Either in print or in machine-readable form, most texts signal the start and end of block quotations (with indentation, typeface change, opening and closing quote marks, etc.). Such representational features need not be recorded, as long as the extent of the quotation is indicated. Other uses of quotation marks, to signal ironic or jocular uses, cited words, titles of books and films, etc., may be marked with SGML entity references for the opening and closing punctuation marks.

Lists: If lists are not marked up in any special way they give rise to a number of undesirable side-effects in text analysis. First, the *item labels* of a list may be roman or arabic numerals, letters, or other printer's mark, and these can be misinterpreted by text searching software as 'real' words. Secondly, lists are often not punctuated according to the normal conventions with respect to sentences. This is likely to confuse tagging and parsing software.

Headings: Headings can usually be interpreted as labels attaching to some structural unit: chapter, section, article, and so on. Or they could be treated as short interruptions in the main flow of the text. Unfortunately, real-life texts are much less tidy than our idealizations of them, and often texts are found in which apparent headings do not seem to be functioning as labels to any structural unit. Newspapers and magazines are particularly inconsistent in this respect.

Abbreviations, initials, and acronyms: A surprisingly high proportion of the word tokens of a large English-language corpus will be accounted for by abbreviations, initials, and acronyms, e.g. personal names, organizations, titles of address, postcodes, units of measurement, days of the week, month names, chemical elements, conventional Latin-derived abbreviations. Of these a substantial proportion could be identified automatically by pattern matching and tagged as contracted forms. Automatic identification and tagging of abbreviations and initials in a corpus may be supplemented by manual editing with the aim of eliminating most of the overlap between words and abbreviations.

Front and back matter: Books will usually include a certain amount of front matter (e.g. preface, foreword, contents, list of figures, acknowledgements) and back matter (e.g. index, appendices, bibliography). Some of these may be captured and included in the corpus, while others may be omitted.

Chapters and sections: These can easily be encoded with little extra effort and keying. Manual editing of the encoding of major text divisions will not be too time-consuming.

Proper names: The aim, in marking these, is to resolve by pre-editing the ambiguity between proper names and other homographic word forms. If carried out manually this is a major undertaking. Word-class tagging software can be used to identify a large percentage of proper names automatically.

Correspondence and addresses: The conventional paraphernalia which attach to the body of a letter (addresses, dates, 'cc' lists, salutation, document references, etc.) need to be handled in an appropriate way to ensure that the non-discursive material is identifiable automatically by processing software.

Pagination: Generally, it will not be possible in all cases in a large corpus to preserve the pagination and numbering, especially if texts are acquired in electronic form. It may, however, be important to users of the corpus to be able to refer to the actual page of the printed original of the text.

7.3 Features for mark-up: spoken text

Speaker change: The basic structure of speech transcription should be a sequence of speaker *turns*. Each turn should begin with an encoding identifying the speaker wherever possible. The use of a minimal encoding will reduce the amount of keyboarding required at the data capture stage.

Syntax: It will be a matter for each corpus project to decide whether, and to what extent, punctuation and mark-up reflects a notion of normal syntax. Word class taggers and parsers are currently oriented almost exclusively towards analysing written language[8] and it may be important to preserve as far as possible the syntactic units of clause and sentence. Other projects may prioritize prosodic analysis of the text and use a mark-up which encodes tone units or other segmentation strategies.

Accent, dialect, and standard orthography: It is important for the corpus builder to decide to what extent the transcription is to represent the sounds of speech (e.g. accent) and to adopt a transcription encoding which is adequate for the representation. If a plain prose-style transcription is adopted, the transcription should be consistent in the use of orthographic irregularities and clear specifications should be given for the use of enclitic forms, variant spellings, and the use of non-standard spelling forms which might be used by the transcriber to reflect the sounds of speech (e.g. *don't, can't, yeah, dunno*). A closed set of permissible forms can be given to the transcribers for guidance. Similarly, one must consider whether dialect forms (e.g. *cannae, you wuz, me = my*) are to be preserved in the transcription or regularized to conform to an agreed standard. Since it is probable that a large natural language corpus will be valuable for the study of dialect forms (in syntax and lexis) the standardization should probably not be applied so strictly that these distinctions are lost.

Interruption and overlapping speech: The writing system does not have very well-established conventions for representing this feature of speech. Interruption is a feature which can without difficulty be represented in the linear stream of writing. It merely requires the insertion of a code or tag which indicates that the preceding turn is incomplete because of the intrusion of the following turn. Interruptions are sometimes more messy, however, and the interrupted speaker may choose to continue regardless of the rival turn: the result will be overlapping speech, which cannot be so easily encoded in linear written form. A detailed mark-up would encode for each overlapping segment of speech

- the start and end points
- an identifier
- the speaker
- the point at which this segment begins overlapping with another
- the point at which this segment ends overlapping with another
- the continuity of each speaker's turn.

The mark-up could become very dense and specialist training and skill could be required to carry out transcription if precise details of overlapping speech are to be recorded faithfully. A sophisticated mark-up would require substantially more time and effort than a simplified encoding, and the cost of staff training, quality control, and the additional time required for analysis and keying should be carefully estimated before final decisions are made.

Pauses: Pauses may be voiced or silent, long or short. It is very simple for a transcriber to record the voiced pauses and they can be encoded in several ways. SGML entity references might serve this purpose quite well. Silent pauses are more problematic, because a transcriber who has only an audio recording of the speech event cannot be sure what other activities or interference might be the cause of a silence on the tape. Unless the recordings are analysed in detail in relation to the physical action of the speech event, the encoding of silent pauses may be misleading and unhelpful. Voiced pauses can be encoded using two codes; one for a short noise and another for a long one.

Functional and non-functional sounds: Some speech sounds have a clear discourse function. The recognized functions may be simplified to a small set each of which could be represented as a standard code. This places responsibility on the transcriber to interpret speech sounds and assign them to appropriate functions. Alternatively, one could devise a large set of orthographic representations to cover a full range of speech sounds and encode the sound rather than its discourse function. If the corpus is primarily for grammatical and lexical studies, then the precise recording of functional speech sounds will not greatly enhance its value. Laughter, coughs, grunts, and other sounds which may occur in recording can be simply encoded.

Inaudible segments: Often, especially if recordings are made in real-life situations, extraneous noise or poor recording conditions will make it impossible for the transcriber to hear exactly what is said. Lacunae should be marked with an indication of the extent of the inaudible segment (measured roughly in, say, seconds, syllables, or 'beats' of speech rhythm).

Spelling: Some words may not be familiar to the transcriber and cannot be spelled with certainty. Proper names and technical terminology are likely to be especially difficult for a transcriber. An attempted spelling can be made at the time of transcription and a marker inserted to allow correction during post-editing, or at least to indicate to the corpus user that the spelling is doubtful.

Numbers, abbreviations, symbols: The use of alternative written forms for such words as Mister (Mr), pounds (£), and two hundred and thirteen (213) needs to be considered. Since there is no sense in which the original form can be faithfully reproduced, a policy of normalization might be followed to make subsequent processing easier.

8 Progress to date with standards for corpora

8.1 Terminology

Standards in respect of corpora mainly concern the compatibility of the kinds of annotation used for texts. These we shall call 'encoding standards'. However, there is also in principle an issue as to the comparability of different corpora, including perhaps judgements on their suitability for different tasks. These we shall call 'evaluation standards'.

8.2 Encoding standards

There is now a fair, and growing, number of corpora becoming available which attempt to provide basic grammatical tagging of the texts they contain. It is not difficult to find differences in the regimes adopted, and hence to recognize the scope for arbitrary incompatibility in what is not a very controversial area (at least by comparison with syntactic parsing, or semantic tagging and analysis).

Any particular comparison speedily becomes lost in a thicket of details. However, it is possible to distinguish a number of different axes on which disagreements are possible, and perhaps this is the first step towards resolving them.

We may ask then:

- which levels of constituent are taggable (from character—e.g. punctuation marks—and morpheme through word to phrase, sentence, and higher units);
- whether the tags are atomic in form, or whether they are effectively complexes of features;
- whether tags are assigned to some singleton classes of words, or these are left to represent themselves (e.g. *the*);
- whether there are formal constraints on assignment of tags (especially, whether non-contiguous items can share a tag);

- whether in some cases the tagging system allows free variation in tag to be assigned—(e.g. is 'male frog' N+N or A+N or both?).

All these questions have been answered differently by the compilers of corpora currently available; no doubt there is scope for more variation too. In many cases decisions have been provoked by particular applications and research goals; in others by physical constraints of the amount of data to be processed within a given time, or the technical facilities available.

As corpora are more and more seen as a fundamental research tool for multiple uses, however, it becomes worthwhile to attempt explicitly to provide a common framework. The objective is, initially, as far as possible to define existing annotation schemes in terms of this framework; ultimately, to add substantive details to it so as to create an all-purpose notation scheme, with resources adequate for all the major applications. Since new applications will no doubt arise in future beyond those currently foreseen, it will also have to have clear rules for extension.

The Text Encoding Initiative (TEI) of the Associations for Computers and the Humanities (ACH), for Computational Linguistics (ACL), and for Literary and Linguistic Computing (ALLC) is the only attempt known to the authors deliberately to propose a common standard in this field, It is notable, however, that its guidelines as so far issued (Sperberg-McQueen and Burnard, 16 July 1990) (which are an extension of SGML, the Standard Generalized Mark-up Language for electronic text formatting) do not go beyond the high-level goal of providing a syntax of labelled bracketings. Hence editors attempting to provide a set of texts in accord with them are still compelled to make a large number of decisions of their own. (See, for example, Liberman's commentary on his redaction of texts for the ACL Data Collection Initiative.)

It is asking a great deal to expect a standard notation to provide a usable common framework for the actual categories used in grammatical tagging, even supposing agreement is reachable on the more formal considerations mentioned above. However, at this point it becomes possible, in principle, to benefit from other related standardization work, namely, that undertaken to identify and promote reusable lexical resources, and to set up compatible formats for machine-readable dictionaries.

The grammatical tags used in corpus annotation will need largely to convey the same information that is assigned to lemmata in dictionaries. Evidently, where corpora are to be processed by NLP devices that access machine-readable dictionaries, it will be simplest if the tags are identical with the dictionaries' grammatical categories.

Five recent or current projects in Europe address this problem.

ESPRIT ACQUILEX (finishing in mid-1991) addressed itself to machine-readable dictionaries as currently available (e.g. as typesetting tapes) for a number of European languages (namely, English, Dutch, Italian), monolingual and bilingual, and aimed to define a useful common notation for their grammatical categories.

EUROTRA-7 was a feasibility study, rather than a concrete project, which ended in April 1991: it surveyed existing terminological and lexical resources of all kinds, and assessed the feasibility of standardizing them. It included both monolingual and bilingual lexical resources, and considered possible architectures for a system that would make them available for potential reuse.

GENELEX, beginning in 1990, aims at defining a 'generic lexicon' which can form the basis for future, more application-specific, dictionaries. This is conceived as a large lexical database, with a specific data model for morphological, syntactic, and semantic information. It is distinctive in focusing first of all on the major Romance languages (French, Spanish, and Italian).

MULTILEX, running from 1990 to 1993, has as one of its main tasks to define standards for representation of lexical entries, monolingual, multilingual, and terminological. These are expected by mid-1992.

The EUROPEAN CORPORA NETWORK, beginning towards the ends of 1991, is aimed directly at setting a common groundwork for the development of textual corpora throughout Europe. As such, it will be recommending aspects of a common annotation schema, i.e. a tag-set, and these should become available in the latter half of 1993.

In general, one can expect advances in developing the tag-set on a standard basis to occur step by step with standardization of dictionary coding; and it is largely to the lexicographer that corpus-builders will have to look for expertise in devising consistent but practical schemes.

It is clear that we are just taking the first concrete steps towards setting up the standards that will make the linguistic annotation of corpora compatible with arbitrary processors. At least the requirements, and to some extent the capabilities, of linguistic processors are clear, since there is a coherent natural language processing community world-wide, now supplemented with forward-looking lexicographers.

However, once the wider uses of corpora identified in Section 9 begin to become real, the need for more general standards will also become evident. These may require annotation codes for a wide variety of subject-matter, and will also involve many differing communities of users. It is to be hoped that, by that time, progress made in defining workable standards for linguistic processing of corpora will have provided insight into how best to extend coverage to quite different fields.

8.3 Evaluation standards

Evaluation standards are less of an issue in respect of text corpora than they are for many other constructs in natural language processing. There is in fact little danger of obfuscation for the major parameters that characterize a

corpus: its size (in numbers of running words) and gross characterizations of its content.

It is necessary to keep sight of both of these parameters when comparing and evaluating results derived from different text corpora. Mutual information figures are only comparable as between items derived from the same sizes of corpus. And evidently it is always at least possible that any statistical relationship observed is crucially dependent (however indirectly) on the choice of subject-matter that occurs in the corpus.

Having said this, however, it becomes clear that the major issues in this sort of standardization are still waiting for the resolution of current disputes. In order to get an uncontroversial view of what type of corpus evidence is required substantively to answer questions of a given type, we must first find agreement on whether 'balancing' of a corpus is ever possible, let alone ever necessary. And that itself is likely to have to await the accumulation of empirical results.

It would seem, then, that evaluation standards for corpora are still premature.

9 Potential users and uses

9.1 Generalities

The large files that constitute a corpus, an electronic text library, or a text archive have come together as if by chance. They are together because of the language they are written in, and it is very unlikely that this had anything to do with the author's motive in writing a text.

This means that users of these files are unlikely to be interested in the content of any particular text. Texts appear as specimens, and users will be interested in what they show about the class represented, not in any of the particular information the text was created to impart.

This makes corpus users rather untypical among the vast ranks of people who are interested in consulting computer files. In fact, corpus users can be divided into three types: those interested in the language of the texts, those interested in the content of texts (as representative of a larger collective), and those interested in the texts themselves as a convenient body of test material for electronic media.

9.2 Language specialists

This class will include the earliest users of corpora. At the moment the most active groups supporting corpus development are drawn from this class, though the picture is slowly changing, especially in Japan.

Lexicographers: They will use the corpus for information on the actual usage of the words they cover. It may be consulted directly for information on specific words; or it may be processed in various ways, in order to develop parts of a lexical database. The processing is likely to require any or all of the electronic-processing media surveyed in Subsection 9.4 below.

Language Workers: Translators, terminologists, and technical writers will be concerned with corpora for special purposes. Parallel corpora (of texts and their translations) are beginning to emerge, especially in the EC where large quantities of documentation is prepared in several languages. Specialized corpora representing technical areas can provide the basis for the enhancement of terminology databases and contribute to the success of efforts towards standardization of technical terminology.

Computational Linguists: At present these researchers separate into two camps, the 'self organizing' and the 'knowledge based'. The former attempts to use the statistical regularities to be found in mass text as a key to analysing and processing it; while the latter brings in the results of linguistic theory and logic as the foundation for its models of language.

The self-organizers use corpora first for initial tests for the presence or absence of regularities they expect to discover. Then, having created search and processing programs, they use corpora to 'train' them, i.e. to refine them through a repeated process of trial and error.

Adherents of the knowledge-based approach have only recently come to recognize the usefulness of corpora in their work, primarily to assess the level of coverage achieved by processes that they have designed *a priori*.

A major promise of corpus-based studies is to suggest how to integrate the work of these two groups. In particular, recent studies have suggested that self-organizing statistical techniques gain much in effectiveness when they act on the output of grammatically based analysis. If this promise is borne out, then we can expect the self-organizing techniques to contribute much more to the semantic, rather than the grammatical or syntactic, analysis of texts: i.e. as a means of analysing their content, rather than their linguistic form.

At the moment, such studies largely restrict themselves to showing how statistical analysis (using mutual information, etc.) can give results which seem intuitively justified in any case. The next stage is to use the techniques so as to reveal semantic regularities hitherto unsuspected; and also to begin to build up general mechanisms for the automatic content analysis of texts. The former of these will use corpus material as its main data source; the latter will need copious corpora for testing purposes during development—though of course ultimately it will be applied to particular texts for which a real analysis is required.

Theoretical Linguists: They view corpora as a mass exhibit *in extenso* of the facts of a language, yielding data on the relative frequency of phenomena of all kinds, and in particular providing a check on the evidence of their own, or their informants' intuitions.

Applied Linguists: In teaching a second language, corpora provide a substantial resource for extracting and studying authentic language data with the authority of attested use. This data might be presented directly to students for classwork or may underpin the preparation of teaching materials of every kind. Increasingly, computer corpora and a range of software packages and tools are available as part of the language teacher's resource pool. Subcorpora of restricted topic areas may be particularly appropriate for use in teaching Languages for Specific Purposes.

9.3 Content specialists

Corpora, as they grow so as to include large subsections classified by date, subject-matter, region, age-group, or whatever, will become interesting as a data source on the classes of people who created the texts as well on the texts' language. Such people will include historians, literary critics, and sociologists, perhaps also advertisers and pollsters. At the moment, these groups are limited in their analysis of corpora to looking at the incidence of fixed words or phrases. But as the language specialists make progress, we can expect corpora to become increasingly useful to those who want a means of handling mass text files, so as to draw out trends and overall analyses.

Historians will be able to track the development of opinions and ideas through study of the words and phrases that refer to them. An example might be a historical study of the use of noun-phrases referring to machines as the subject of action verb-phrases, as indirect evidence on how our mechanistic conception of people and animals has established itself. They will also be able to use dated and analysed corpora to discover implicit time-stamps and place-stamps, which they can then apply to identify documents whose origin is unknown.

Literary critics have already made signal use of corpus-based research under the heading of stylometrics. Statistical analysis of word use is already crucial in determining ascription of dubious work to a known author, and such techniques can only become more effective as linguists discover the significant features which are present at a higher level than individual words.

Besides acting as a mass training ground for these techniques, corpora will also be useful as sources for statistical information on the differences of style characterizing different groups, whether by age group, period, country of origin, or whatever. Whole areas of literary analysis which are now inescapably subjective will begin to allow some objective test—though this is unlikely to make them any less controversial!

Sociologists will be able to make similar uses of the corpus resources, though here the parameters of interest will be different: not period, author, or genre, but class, race, creed: whatever, indeed, the corpus architects may have chosen as significant labels for texts, and any combination of these labels. We can expect statistical confirmation of the whole host of perceived nuances by which people's language betrays their origin, but also the discovery of a host of others, hitherto too subtle or abstract to be picked up.

9.4 Media Specialists

Corpora will be the indispensable test bed for all the text-processing functions that software developers devise in the coming years. Inevitably, the success of these new developments can only follow on substantial progress in the research of the language specialists we have already considered. However, even now, it has been recognized by the US Defense Advanced Research Projects Agency that a large and common source of textual data is a *sine qua non* of progress both by the researchers and the developers who follow in their wake.

These text-processing systems will be quite various, and there may come a time when human authors are no longer the only source of readable text, nor human readers the only users of it. At the moment, attention is concentrated above all on three types of application: Information Retrieval Systems, Machine Translation, and Speech Processing.

Information retrieval systems are themselves extremely varied. They include mechanisms to extract information that fits a given format from bodies of text (which may be fixed or may be dynamically accumulating as real time messages), then using it to build up a knowledge base; mechanisms to find enough information in items (say, messages or articles) to decide or an appropriate addressee (i.e. message-routing, document-clipping); mechanisms to find the information for an index, or more ambitiously, to summarize the important content.

Machine translation (or machine-aided translation) is a computer application whose attraction, where it is feasible, is self-evident. Besides their use as a test bed, corpora here are beginning to make significant contributions to the actual capabilities of systems. This is because of the increasing availability of bilingual corpora (e.g. in the proceedings of institutions that are legally required to be available in multiple languages: literary classics and best-sellers may also often be available in this form, though the standards of equivalence in translation may be more lax). These enable a self-organizing approach to supplement the traditional knowledge-based one. Speech processing in general is benefiting more and more from the development of its own speech corpora. These compilations, though, are still quite distinct from text corpora, being extended acoustic analyses of speech wave-forms; hence they are not intrinsically based on character or word analysis as is a text corpus. Furthermore, the amount of language represented in such a corpus is several orders of magnitude smaller than a typical corpus of texts. As such, they are beyond the scope of this paper.

Nevertheless, text corpora derived from speech will be increasingly of use to the developers of speech processors. Such corpora will be literal transcriptions of spoken language. Part of their advantage for speech analysts lies in their potentially much greater scale: textual-type annotation is much quicker and easier to add than the type of annotations required for a speech corpus. But this does not in itself advance the capabilities of speech processing. Where it will help is in making it possible to identify some of the higher-level regularities correlating with parameters of a given type of speech (e.g. proneness to hesitation in certain contexts or preferred sentence structure): these in turn can be used to tune the speech processor.

9.5 Types of use

By way of summary, we can discern two major classes of use for corpora amidst all this potential variety: the corpus as a large-scale, but heavily diluted source of data, which new techniques are enabling us to sift; and the corpus as a test bed, composed of representative if mostly undifferentiated material, good for testing or training an automatic device under development.

Although we have described a corpus as undifferentiated, this is meant only in the sense that the precise content of the component texts is not to the point. For the corpus to satisfy as a useful test bed, it must be highly calibrated, with as many of the external parameters of its constituent texts stated as possible. The different users can then either discover the linguistic correlates of these external parameters, or else use them to judge which parts of the corpus are appropriate as a test for their processing device.

Notes

[1] The first three types are those identified by Bernard Quémada.
[2] For a more detailed discussion of this topic see Clear (forthcoming), Corpus Sampling, *Proceedings of 11th ICAME Conference*, Berlin, 1990.
[3] *Statistics in Language Studies* (1986), p. 55.
[4] Biber and Finegan have published widely on this subject: e.g. Biber, D. (1989). A Typology of English Texts, *Linguistics*, 27.
[5] In Section 4.2 of Corpus Design—Problems and Suggested Solutions, ICE working paper, May 1990.
[6] Sperberg-McQueen and Burnard, 1990.
[7] Such a transcription schema is described and its merits discussed by Du Bois (forthcoming).
[8] Svartvik and Eeg-Olofson have developed a tagger and parser tailored to analysing speech in the London-Lund corpus.

Acknowledgements

We should like to acknowledge gratefully the help we have received in preparing this paper; help from Ken Church, Patrick Hanks, Steffen Leo Hansen, Ken M. Lauridsen, Geoffrey Leech, Christine Leonhardt, Ole Norling-Christensen, Hanne Ruus, Don Walker and Antonio Zampolli.

References

Aarts, J. and Meijs, W. (eds) (1984). *Corpus Linguistics: Recent Advances in the Use of Computer Corpora in English Language Research*. Rodopi, Amsterdam.

—— (eds) (1986). *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora*. Rodopi, Amsterdam.

Atkins. B. T. (1987). Semantic ID Tags: Corpus Evidence for Dictionary Senses. In *The Uses of Large Text Databases: Proceedings of the 3rd Annual Conference of the UW Centre for the New Oxford English Dictionary*. University of Waterloo, Canada, 17–36.

Biber, D. (1989). A Typology of English Texts, *Linguistics*, 27: 3–43.

Church, K., Gale, W., Hanks, P. and Hindle, D. M. (1990). Using Statistics in Lexical Analysis. In Uri Zernik (ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon*. Lawrence Erlbaum, Hillsdale, New Jersey.

Clear, J. (1988). Trawling the Language: Monitor Corpora. In M. Snell-Hornby (ed.), *ZüriLEX '86 Proceedings:* Papers read at the EURALEX International Congress. Francke Verlag, Tübingen, 383–9.

—— (forthcoming). Corpus Sampling, *Proceedings of 11th ICAME Conference*. Berlin, 1990.

Du Bois, J. W. (forthcoming). Transcription Design Principles for Spoken Discourse Research, *IprA Papers in Pragmatics*.

Engwall, G. (forthcoming). Chance or Choice: Criteria for Corpus Selection. In B. T. Atkins and A. Zampolli (eds), *Computational Approaches to the Lexicon*, Oxford University Press, Oxford.

Francis, W. N. (1979). Problems of Assembling and Computerizing Large Corpora. In H. Bergenholz and B. Shäder (eds), *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora*. Scriptor Verlag, Königstein, 110–23.

—— (1980). A Tagged Corpus—Problems and Prospects. In S. Greenbaum, G. Leech, and J. Svartvik (eds), *Studies in English Linguistics, for Randolph Quirk*. Longman, London and New York, 192–209.

Garside, R., Leech, G. and Sampson, G. (1987). *The Computational Analysis of English*. Longman, London.

Halliday, M. (1966). Lexis as a Linguistic Level. In C. Bazell, J. Catford, M. Halliday and R. Robins (eds). *In Memory of J. R. Firth*. Longman, London.

Johansson, S. (1980). The LOB Corpus of British English Texts: Presentation and Comments, *ALLC Journal*, 1: 25–36.

—— (ed.) (1982). *Computer Corpora in English Language Research*. Norwegian Computing Centre for the Humanities, Bergen.

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

Källgren, G. (1990). The First Million is the Hardest to Get. *Proceedings of the 13th International Conference on Computational Linguistics (COLING)*. Association for Computational Linguistics, Helsinki.

Klavans, J. and Tzoukermann, E. (1990). Linking Bilingual Corpora and Machine Readable Dictionaries with the BICORD System, *Proceedings of the 6th Annual Conference of the Centre for the New OED and Electronic Text Research*. University of Waterloo, Canada.

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

Kyto, M. (1989). Introduction to the Use of the Helsinki Corpus of English Texts: Diachronic and Dialectal. In M. Ljung (ed.), *Proceedings of the Computer Conference in Stockholm*. Stockholm, Sweden.

—— Ihalainen, O. and Rissanen, M. (1988). *Corpus Linguistics, Hard and Soft: Proceedings of the 8th International Conference on English Language Research on Computerized Corpora*. Rodopi, Amsterdam.

Leitner, G. (1990). Corpus Design—Problems and Suggested Solutions. Working paper in: *ICE Newsletter 7*, May 1990. International Corpus of English, University College London.

—— and Schäfer, U. (1989). Reflections on Corpus Linguistics—the 9th ICAME Conference in Birmingham, England, *CCE Newsletter*, 3 (1): 2–16.

Liberman, M. (1989). Text on Tap: the ACL/DCI. *Proceedings of DARPA Speech and Natural Language Workshop*. Morgan Kaufman, New York.

Meijs, W. (ed.) (1987). *Corpus Linguistics and Beyond: Proceedings of the 7th International Conference on English Language Research on Computerized Corpora*. Rodopi, Amsterdam.

Oostdijk, N. (1988a). A Corpus for Studying Linguistic Variation, *ICAME Journal*, 12: 3–14.

—— (1988b). A Corpus Linguistic Approach to Linguistic Variation, *Literary and Linguistic Computing*, 3: 12–25.

Renouf, A. (1987). Corpus Development. In J. Sinclair (ed.), *Looking Up*. Collins, London, 1–41.

Quirk, R. and Svartvik, J. (1979). A Corpus of Modern English. In H. Bergenholz and B. Shäder (eds), *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora*. Scriptor Verlag, Königstein, 204–18.

Shastri, S. V. (1988). The Kolhapur Corpus and Work Done on its Basis So Far, *ICAME Journal*, 12: 15–26.

Sinclair, J. (1982). Reflections on Computer Corpora in English Language Research. In S. Johansson (ed.), *Computer Corpora in English Language Research*. Norwegian Computing Centre for the Humanities, Bergen, 1–6.

—— (ed.) (1987). *Looking Up*. Collins, London.

—— (1989). Corpus Creation. In Candlin and McNamara (eds), *Language, Learning and Community*. NCELTR Macquairie University, Sydney.

Sperberg-McQueen, C. M. and Burnard, Lou (ed.) (1990). *Guidelines for the Encoding and Interchange of Machine-Readable Texts—TEI P1*. ACH–ACL–ALLC.

Svartvik, J. (1990). *The London-Lund Corpus of Spoken English: Description and Research*. Lund Studies in English, 82. Lund University Press, Lund.

The Uses of Large Text Databases: Proceedings of the 3rd Annual Conference of the UW Centre for the New Oxford English Dictionary (1987). University of Waterloo, Canada.

Warwick, S. and Hajic, J. (1990). Searching on Tagged Corpora: Linguistically Motivated Concordance Analysis, *Electronic Text Research: Proceedings of the 6th Annual Conference of the Centre for the New OED and Electronic Text Research*. University of Waterloo, Canada.

Woods, A., Fletcher, P. and Hughes, A. (1986). *Statistics in Language Studies*. Cambridge University Press, Cambridge.