EAGLES

Expert Advisory Group on
Language Engineering Standards

EAGLES

Preliminary Recommendations on Text Typology

EAGLES Document EAG–TCWG–TTYP/P

Version of Jun, 1996

Contents

1 Contributors

This document, as with all EAGLES documents, is the result of cooperative work involving a number of individuals, contributing in different ways.

1.1 Authors

J. McH. Sinclair
Corpus Linguistics Group
School of English
Birmingham University
Westmere Mews
50 Edgbaston Park Road
Birmingham — Tel.: +44.121.414.6789
United Kingdom — Fax.: +44.121.414.3288
B15 2TT — E-mail: J.Sinclair@bham.ac.uk

J. Ball
c/o Corpus Research
School of English
Birmingham University
Westmere Mews
50 Edgbaston Park Road
Birmingham
United Kingdom
B15 2TT

1.2 Editors and Assistants

N. Calzolari	
Co-Chief Editor, EAGLES	
ILC–CNR	
Via della Faggiola 32	Tel.: +39.50.56.04.81
Pisa I-56126	Fax.: +39.50.58.90.55
Italy	E-mail: eagles@ilc.pi.cnr.it

J. M^{c}Naught	
Co-Chief Editor, EAGLES	
Centre for Computational Linguistics	
UMIST	
Manchester	Tel.: +44.161.200.3098
United Kingdom	Fax.: +44.200.3099
M60 1QD	E-mail: jock@ccl.umist.ac.uk

T. Paskiewicz	
EAGLES Editorial Assistant	
Centre for Computational Linguistics	
UMIST	
Manchester	Tel.: +44.161.200.3097
United Kingdom	Fax.: +44.200.3099
M60 1QD	E-mail: teresap@ccl.umist.ac.uk

T. Ayazi	
EAGLES Secretariat	
ILC–CNR	
Via della Faggiola 32	Tel.: +39.50.56.04.81
Pisa I-56126	Fax.: +39.50.58.90.55
Italy	E-mail: eagles@ilc.pi.cnr.it

A. Enea	
EAGLES Webmaster	
ILC–CNR	
Via della Faggiola 32	Tel.: +39.50.56.04.81
Pisa I-56126	Fax.: +39.50.58.90.55
Italy	E-mail: enea@ilc.pi.cnr.it

2 Introduction

This Report and its recommendations should be read in conjunction with EAGLES interim recommendations on Corpus Typology (EAGLES, 1996a). The categories in the Appendices have been prepared from the published accounts of practice and the few speculative discussions on the subject. The starting point is the NERC Report (Calzolari et al., 1995). Preliminary versions of this Report were circulated to members of the EAGLES Working Group on Corpora, and the Report was presented to a wider circle of colleagues at an EAGLES workshop in Madrid in January 1996. Feedback from these activities and from correspondents has been incorporated in this version.

In policy discussions about the format of these recommendations, it was made clear to the various groups of contributors that the desired format for classification is a hierarchy of attributes, each with a set of values, because this is a convenient format for database organisation and machine processing. This format has been adopted wherever possible. However there is a number of circumstances where it is not an appropriate model for the patterns of language in communication, and there recourse has been made to other formats. In some cases, the lack of a full and finite set of values simply reflects a lack of experience in applying such a typological classification; in a short time they are likely to be clarified. In other cases, the set of values is dependent on what the texts say about themselves (see the discussion of *reflexive* under section 3 below) and therefore is not under the control of people who wish to classify; in other cases the governing criteria are *internal* (see section 5), based on linguistic choices, and sometimes of quite a different order compared with external classifications. While these internal patterns are often reflected dimly in external distinctions, it would be a grave error of judgement to settle for a set of values that has a large arbitrary element, which in turn is likely to obscure the internal ones.

At times the attribute/value model is just irrelevant — it would in fact be most unlikely that such a simplistic model would always be adequate for the description of human language.

The focus of this typology is written language. However, it is envisaged that some quantities of transcribed spoken material will normally form part of reference corpora, and so basic provision is made for the classification of spoken data in the classification. Towards the end of this report there is a short note on spoken texts(see section 8.

2.1 Text Encoding Initiative

The Text Encoding Initiative guidelines (Sperberg-McQueen & Burnard (eds), 1993) suggest a classification of texts based purely on external criteria (i.e. non-linguistic criteria). The categories and subcategories specified are both exhaustive and exhausting to contemplate, and in practical terms, that is in terms of person hours needed to encode each text as well as computer resources needed to store the headers, the standards do not seem feasible for the size of corpora advocated in EAGLES (see EAGLES (1996a). At the minimum level of encoding according to TEI, we record the details of how the texts were made available in machine readable form (method, persons responsible, date and so on), details of publication (including date, address, copyright status, etc.) and a bibliographic description (information about the author such as age, sex, mother tongue, nationality and so on, and details of how the text is published — series, journal and so on). It is also recommended that the size of the file and statements on edition and series are recorded. Within these categories all non-linguistic information about the text can be stored.

The list is endless, and there is no requirement of linguistic relevance to help us distinguish the external features which are likely to be of value for users, and those which are only of archival interest. Almost any conceivable feature may come into relevance on a future occasion, but that is not sufficient reason for investing in the cost of ascertaining and recording such a feature in respect of each text in the corpus.

In any case, experience has already shown that such precategorisation is almost certainly going to be inadequate, for two reasons. One is just the record of experience — customers always want features that are not readily available; the other is that the reasons that bring into focus some feature of the context of situation of a text will involve a reconceptualisation of the relationship between the language and the situation, and that will marginalise any previous classification in the area.

Consider the matter of ‘political correctness’, which is typical of a parameter of classification of language that was not recognised at all a few years ago, but might now be the object of serious studies. Various kinds of offensive language have traditionally been recorded, and blasphemy, for example, has a long history, as has pornography. Racist language has more recently been recognised and used in classification, and sexist language also. But political correctness requires a reconceptualisation of the relationship between writer, text, reader and third party that has specifically not informed the earlier classifications; therein lies its interest. Until we have the concepts that give an adequate reason for the classification, the classification is not worth doing.

So it is not good policy to second-guess the future. Instead we should pay close attention to those external features that are known or strongly suspected to influence linguistic choices systematically.

In parallel to this report, an EAGLES working group on Text Representation makes practical recommendations on how TEI policies can be adapted to corpora (EAGLES, 1996b).

2.2 Internal and external criteria

The earliest electronic corpora were designed using external criteria — reference to institutionalised types of text, or features of the nonlinguistic environment or society in which the texts occurred. More recently, some internal criteria — differentiating features of the language of the texts — have been offered by researchers. This work suggests that a thorough classification of texts, an adequate typology, will eventually consist of a balanced combination of the two types of criteria.

Many internal and external criteria reflect each other. A text that showed a high average sentence length would be likely to be of one kind of book or magazine rather than another. For some parameters of classification the external evidence is primary — whether a text is printed in a book or a newspaper, whether it is in written or spoken form, etc. No matter what kind of language is found in the document or transcript, these real-world distinctions are clear-cut, and relate to the experience of end-users of the typology, who will not all be experts in linguistics.

So if it turns out that some texts which are classified separately from an external perspective share some of the same linguistic features, that is not in itself an adequate reason for classifying them together, though it raises an interesting question. Much research on language variety, genre etc. has worked on the assumption that external distinctions are reflected internally, and indeed it is ultimately that assumption that makes external criteria relevant at all — for if they had no influence on the language used, they would be of no interest to linguists or others in the language industries.

Perhaps the major indication of whether a corpus criterion should be expressed through internal or external criteria is whether the criterion itself is expressed intensively or extensively. This is a basic distinctiion in lexicography; an intensive criterion is expressed by something like a generalisation which is shared by all the members of a group, whereas an extensive criterion is merely a list of the members. Where the group is a fairly small and clearly delimited one, like ‘quality newspapers’ in the UK, there is little to choose between listing them and phrasing a generalisation that will identify just the same newspapers. But when there is a large number of members of a group — and possibly an indefinitely large number — then the techniques of extensive definition do not work.

Hence, where we find, as in the classification of topic by external criteria through listing possible topics, an example of extensive listing without a principled limit or known dimension, there is a strong case to be made for abandoning this approach and looking instead for a generalised means of classifying the texts using internal criteria and intensive definition.

There is a halfway house, where the classification is based on what the text says about itself. If on the title page of a text it is stated that the text is a ‘manual of lithography’, it is fairly safe to classify it as a manual, and to record its topic as lithography. Such a classification is called reflexive (see section 3).

In a perfect world these claims would have to be verified with respect to the criteria for manuals, and the place of lithography in relation to other arts and crafts, and the printing trades. For the time being we accept the classification at its face value; we do not define a manual carefully, and we do not attempt to divide the world around lithography into a coherent set of labels.

2.3 Topic

Topic is one of the central controversial areas of text typology. No existing external classification seems to be satisfactory. It can be analysed in several different ways, and the practice so far in many corpus designs has been to erect a makeshift and broad-mesh framework, within which the texts are disposed into undefined or inadequately defined categories. That is to say, an extensive categorisation is imposed, representing topic as capable of being dealt with by essentially external criteria. However, we believe that topic should be conceptualised principally as an internal matter, to do with things like the vocabulary choices in a text, rather than an external matter, where the Universe is endlessly chopped up into subcategories. Because of the importance of this point, we anticipate our position and illustrate the line of argument that is developing in the handling of the area. We make the following claim:

> *In the classification of topic, the internal evidence is primary.*

That is to say, it will lead to a better classification of topic if the internal evidence, such as the vocabulary clustering, is developed first of all, and the external evidence is added at a stage of greater detail.

This matter is taken up in detail later in this report. For the present, we put forward some notes towards substantiating the above claim. Let us first of all consider the topic variation in documents and conversations.

A newspaper covers a wide range of topics, with some stability from issue to issue, but also a great deal of variation. Unless they are hand-annotated (and this report envisages corpora that are much too large for hand-annotation) the electronic versions of most newspaper texts are not explicit about topic.

This position excludes cases where incoming newspapers to a corpus, say on a compact disk, are pre-analysed by experts, and divided into their constituent stories, reports, news items etc., to each of which is assigned a topic or a small list of key words. Such texts are hybrids between printed and electronic material (see 4.2 below).

Newspapers in their printed form are usually divided into sections, e.g. the sports section. This is a kind of self-classification for which the term used in this report is *reflexive* (see section 3). If we make use of self-classification, for topic consistency it might be helpful to see the structure of a newspaper in two dimensions — the issue and the contents, as in table 1. The vertical dimension of table 1 is likely to show more consistency of topic than the horizontal.

	news	sports	women	science
issue 1				
issue 2				
issue 3				

Table 1: Structure of a newspaper

Similarly, a novel covers a variety of topics, but without even the segmentation of a newspaper. It comes in chapters, usually substantial, and there is no requirement on the author to have consistency of topic within a chapter. Popular novels (spy stories and the like) frequently divide each chapter into a number of short episodes, each on a different topic.

An impromptu conversation is seamless, and moves from topic to topic normally in a series of steps that specifically obscure the changing of topic. The procedure is very subtle (Hazadiah, 1993), and there is no immediate prospect of automating the analysis of discourse in such a way that it would be capable of topic segmentation.

We have, then, identified three major sources of language corpus data for which external topic analysis is likely to be problematic and unhelpful. In the case of some classes of document the relationship to topic may be simpler (specialist magazines, formal reports, etc), but even when what is written seems to be confined to a restricted topic area, it would be foolish to be optimistic.

Hence, although general classification systems have been worked out (Dewey Decimal Classification, etc. — see Internal criteria5), they are not reflected simply in text structure, No doubt very short texts, and very short stretches of text — fragments — will be so classifiable, but not the majority of documents or conversations.

The reasons are obvious and need only be mentioned here. One is that many communication events have as their origin the need to mix topics — to talk about the influence of *X* on *Y*, etc. One answer to this might be multiple classification, but the likely result of taking that step would be a network of criss-crossing topics that would be useless for practical purposes.

Another reason is that many communicative events have a structure that is quite insensitive to topic — as for example a daily newspaper or news broadcast. The topics are whatever is felt to be newsworthy, their sequencing reflects their perceived importance, with occasional moves towards coherent grouping, and some deliberate contrasts.

Yet another reason for the unpredictable nature of topic is the social need for maintaining interest and attention by refocusing the topic frequently in conversation. This feature, combined with the social requirement to obscure changes in topic, makes any simple idea of topic unlikely to be usable in classification, whether external or internal. As we shall see below, the likely internal topic-related patterns are not just a list of one-word subject labels.

The Subject typology given in appendix is culled from as many published sources as we could find, following the NERC study, and it shows the unsuitability of trying to build on received practice (because it is so inconsistent) or of trying to arrange a hierarchy of simple topic labels. The memoirs of a retired medical missionary who had an important collection of military paintings, particularly canvases showing the details of early ordnance and regimental uniform; who delighted in the languages and buildings that he had met in his travels, and paid close attention to the level of scientific sophistication in the agriculture of the regions, and the problems of distance from major centres — such a document, which is by no means fantastical, would be listed under more than a dozen topics of those given in appendix.

The discussion of topic is returned to later in this report (see section 5); in particular how a means of analysis might be developed considering topic as an internal property of texts. For the primary purpose of this paper — to offer a workable classification scheme for today — topic has to be replaced by several lists of external criteria.

The sectionalisation of newspapers has already been identified as reflecting some aspects of topic; to this we can the stated subject-matter of magazines and periodicals, which in the case of the more specialised ones is helpful. Also we can use the fact that a society institutionalises a number of topic-related classifications; of particular value to text typology are lists of recognised professions and educational courses. These are matters of social policy and expediency at a particular stage of development, and have quite a different status from general attempts to classify the cosmos; they will vary from culture to culture and language to language, but they constitute a set of parameters of classification that has scientific validity and is workable in average corpus projects.

The self-classification of reports, textbooks, etc., has already been mentioned as a valuable feature of the reflexivity of language. A typology based on such criteria will be untidy, but they contribute to a defensible assignment of topic that will suffice until it can be replaced with something better.

2.4 Genre

The classification of texts into different genres seems to have been mostly achieved through external criteria rather than through internal, linguistic criteria. Bhatia (1993) claims that the nature of genre can be specified through external criteria. The particular external criteria that are usually specified consists of information concerning the speaker/writer and audience, as well as the relations between these two groups, the intended goals of the author, the historical, socio-cultural, philosophic and occupational details. Bhatia would further determine the genre of a text by relating the particular text to other texts of the same genre and by identifying topic, again externally.

Whether the text is written or spoken is a basic distinction, though Sager (1980) identifies types of text within the language of science and technology which can be both, such as the report, the memo and the schedule where the medium would not necessarily determine the text type. Types are distinguished instead on grounds of the status of the participants and their knowledge or authority on the subject of the text. These modes of expression are also based on the level of pre-planning, that is whether the text is spontaneous or prepared.

Sager also claims that such matters as informative and evaluative intentions provide criteria for subdivision of texts. This is the basic classification of the written part of the British National Corpus — which is primarily divided into the categories of informative and imaginative writing — the intention being the fundamental division between text types. Informative writings include the fields of sciences, arts, commerce and finance, belief and thought, leisure, natural and pure science, social science and world affairs. Imaginative writing is literary and creative works. These types are then further categorised according to other criteria, such as medium, date of publication, topic and so on.

An early distinction between informative and other text — whether the other is said to be evaluative or imaginative — is of doubtful validity because it perpetuates the illusion that many texts have as their principal aim the transfer of information. Oversimplifications regarding the concept of information abound in informatics — in fact the science of informatics is based on the oversimplification that information can somehow be decoupled from the rest of communication. In contrast, Hunston (1989) has shown that the blandest and — apparently — most objective scientific report is seething underneath with evaluation.

The category 'factual' is far wider than 'informative', but often the two seem to be merged. The distinction between fact and fiction is relevant, but it is a distinction that cuts across others, and concerns the relation of the utterances to objective reality, and not in the first instance to their purposes. In contrast, 'informative' is a term that relates to purpose.

Hence, in the typology proposed, the category of information is one of the categories of outcome, and is restricted to reference works that appear to have no practical relevance beyond being repositories of data.

Halliday & Martin (1993) do not offer a formal text classification system. They identify the areas of field, mode and tenor through internal criteria. Since language varies with situation (which is presumably the fundamental reason for text classification at all), we can identify particular specialised languages through the specification of certain internal linguistic criteria — that is, the frequency of lexico-grammatical features. Halliday & Martin associate scientific text with passives and nominalisations and so on. However, here it would seem that some sort of initial external classification is necessary, to begin with at least, in order to specify the characteristics of a particular genre, here a 'scientific text'. The internal linguistic criteria of the text is analysed subsequent to the initial selection based on external criteria. The linguistic criteria are subsequently upheld as particular to the genre of the 'scientific text'. This classification begins with external classification and subsequently focuses on linguistic criteria. If the linguistic criteria are then related back to the external classification and the categories adjusted accordingly, a sort of cyclical process ensues until a level of stability is established. Such a process would additionally check that there were no other prominent varieties that showed similar linguistic features — a point that can be readily overlooked when an investigation starts with externally-defined texts.

Biber (1989), on the other hand, rejects typologies which have been determined by initially isolating an important external difference among texts and have subsequently attempted to identify the linguistic features associated with that difference. Biber therefore establishes a typology of texts based on internal linguistic criteria only and then interprets the results with reference to external 'functions'. Biber's internal criteria are taken from published studies of language variation; therefore they can be assumed to be relevant. The resultant typology ultimately distinguishes eight text types based on five 'dimensions' identified automatically through cluster analysis, as we will see below.

This classification of texts based purely on internal criteria does not give prominence to the sociological environment of the text, thus obscuring the relationship between the linguistic and non-linguistic criteria. Atkins et al. (1992) believe that:

> *it is impossible to 'balance' a corpus on the basis of extra-linguistic features alone*

but similarly that:

> *a corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation.*

It seems most feasible that texts be selected initially on external criteria, then once the particular linguistic features of a 'text type' have been established through analysis of internal criteria these can be used in selection and classification of texts. Biber later suggests that a cyclical process of refinement between internal and external criteria is necessary (Biber, 1993).

3 External criteria

Both external and internal criteria are therefore to be considered in the classification of texts. A typology of texts for inclusion in corpora cannot be based entirely on external or internal criteria. It is not feasible to classify texts entirely on internal evidence, nor is it desirable if we are to learn about the relation between linguistic and non-linguistic features. Similarly, in a classification system based purely on external evidence we would not necessarily group linguistically similar texts together but may impose divisions on texts which may not reflect linguistic differences.
Many internal and external criteria are likely to be broadly correlated, and it is not always straightforward to decide which phenomena are best treated by one or the other. To complicate the issue, there is an in-between category, which we term *reflexive* (Lyons, 1977). This is where a text talks about itself, and proposes its own classification. Reflexivity is a property of all languages, and is the basis of much of what is conventionally regarded as external criteria. The title page of a novel, for example, will contain the name of the author, and the date may appear on it. These will be accepted as external 'facts', unless challenged. It may well also say "A Novel", which although more contentious is probably acceptable as a genre label. The existence of such proposals in the text itself is not incontrovertible evidence of the accuracy of the classification, but for the control of large corpora there is no practical alternative.
We therefore divide external criteria into two varieties:

1. *Circumstantial* where the evidence comes from outside the text;

2. *Reflexive* where the evidence comes from statements within the text.

Reflexive evidence should wherever possible be cross-checked against circumstantial evidence.
The example of 'novel' above brings into focus another useful group of terms. In establishing a typology for texts in computer corpora, we should not forget categories of classification that are already in use. It is expedient to reuse terms that already are accepted as categorisers in literary criticism, rhetoric, etc., even when they cut across the hierarchical organisation of this typology. For example, a novel is a (E.2.1.2.1.1. — see 4.2) book and its audience is the general public (E.3.1.*.2.1. — see 4.3); it is literary-recreational (E.3.2.4. — see 4.3). Terms such as novel are asterisked to indicate that they are *terms externally defined* (TEDs).

3.1 Received categories

In the NERC investigation that forms the starting-point of this paper, active centres of corpus work were asked about the categories that they used in order to organise and balance their corpora. Several categories were found to be much use, and they are dealt with here as *received categories*.

3.1.1 Literary Genre

A distinction is proposed within the category recreation between fact and fiction rather than between literary and non-literary. 'Faction' is regarded as a variety of fiction, and, to avoid controversy, religious works are given a separate category of their own. Within fiction the distinction between poetry and prose is reasonably reliable, and between dramatic and non-dramatic text. In prose, there are sagas, cycles, trilogies and various other kinds of serial linkages, which we propose to ignore for a corpus of today's dimensions. The novel, novella and short story are three well-recognised sizes of prose fiction. Novels are sometimes classified further, into historical, psychological, humour, science fiction, romance, thriller, crime, spy, etc., with alarming possibilities of double classification — spy thriller, historical romance, etc. This is where some use may be made of the category of reflexive — if a novel declares itself to be such-and-such, then it would be reasonable and economical to record this; but if it does not, then needless labour would have to be expended on deciding, and endless controversies could arise.
Biography and autobiography are factual, along with other categories such as academic works, textbooks, and practical guides, but their purpose is recreational rather than instructional.

3.1.2 Medium

The distinction between books, magazines, etc., is almost universally accepted, and followed here. It cross-cuts the literary genres to some extent, because while a novel is almost certainly a book, a poem is not one, normally — yet a book of poetry is familiar enough. Dickens' novels came out as part works, and so do some modern stories. An obituary in a newspaper is a kind of biography at the level of a short journalistic piece, etc. However, it seems in our normal understanding of classification, literary texts can be kept together, at the risk of duplication of some criteria, rather than mixing them up with nonliterary material according to other criteria.

3.1.3 Style

This is always a hotly disputed area, and there is no agreed set of values. Aspects of the audience of a work — the readership — can affect the language, and the relationship between writer and reader. Style is typical of the kind of classification that is almost certainly driven by internal criteria rather than external, and the unsatisfactory state of the classifications so far (e.g. Joos, 1961) points this up. Style will be treated in detail below under internal criteria.

3.1.4 Mode

This normally divides into *spoken* and *written*. We have added a third, *electronic*, to emphasise that language transmitted in electronic media is not quite the same as the older established modes. There is not universal agreement about how exactly the special status of text originating as electronic files should be related to the two long-established modes, and the point should be followed up in order to achieve a consensus if possible. Certainly electronic text shows internal features which are different from both traditional written and traditional spoken text, and the social role of such text is also new and different; for these reasons a third mode is at least temporarily posited here.

The spoken mode is included in this typology and is dealt with at greater length below; (see also Leech et al. (eds), 1995 for an up-to-date account of issues involving spoken corpora).

This typology follows the recommendations of the EAGLES Recommendations on Corpus Typology(EAGLES, 1996a) in order to avoid confusing the modes. As Firth (1957) said, written language has "the implication of utterance", and equally, any policeman knows that anything you say can be written down. So the only way of keeping the two apart is to classify according to the actual mode of the material when selected for inclusion in the corpus. Hence descriptions like 'written-to-be-spoken' are not to be found below; they are too close to the operations of the Intentional Fallacy.

Spoken language is often thought to be less formal than written, more impromptu. This may be statistically true in a broad general perpective, but there is plenty of formal spoken language around, on the radio and television, in legal and official circles and in diplomacy. Equally there is still a lot of very informal correspondence and despite the invention of e-mail.

Sinclair makes this point with a figure, reproduced here, of audience size related to genres of spoken and written material (Leech et al. (eds), 1995). The formality level of a piece of language probably owes more to such factors as audience size and occasion than it does to the mode.

```
spoken                             BIG                         written

      radio/tv                  1,000,000s              newspapers
        local radio               100,000s            magazines/books
          rallies                  1,000s           notices
            lectures                100s          local publications
              classrooms             10s        workplace records
                discussions          5s       circulation lists
                  interviews         3s     working groups
                    conversation     2s   private letters

                                  LITTLE
```

Following the EAGLES recommendations on corpus typology, there is a conscious attempt in this report to avoid mixed categories, because the classification rapidly becomes meaningless and arbitrary if mixing is permitted. It has already been pointed out that anything that is said can be written down, and anything written can be read out. Either speech or writing can be converted to electronic form with ease. More and more texts are available in more than one form, for example correspondence which is composed on a word-processor and then printed out for transmission by ordinary post, or newspaper archives put on CD-ROM.

However, texts are rarely composed for transmission simultaneously in more than one mode, and where a text is found to exist in two or all three modes it is usually easy to identify the primary mode. A conversation, recorded, transcribed and keyed into a computer exists in all three modes, but if it is chosen for the corpus because it is a conversation, and if researchers consider that its priimary mode is spoken, then it fills a place in the spoken dimension of the corpus. The written transcription can usually be discarded if it contains no more information than the electronic version, but the original sound recording will normally be preserved, because no transcription is an adequate substitute (Leech et al. (eds), 1995).

Computers are already able to hold texts in one or more modes, and in nonlinguistic formats as well (e.g. the sound track of a conversation, or a digitised facsimile of the written page), and to correlate the different versions; in the near future it will become commonplace to hold alternative versions, so the corpus design will have to be ready with methods for multiple classification of texts.

For a meaningful classification we have to go back to the design of the corpus and the place in that design that a given text is chosen to fill. The corpus builder must be clear enough in his or her mind which is the primary mode, and distinguish that from any others.

All texts in a corpus end up in electronic form, but that is not the reason for establishing the category of electronic mode in this typology. The reason is that a large body of text material is now appearing only in electronic form, or apparently primarily in electronic form; the communicative features of this new means of transmission lead to distinctive choices of form and style, that are not found in either spoken or written texts.

There are, and there always have been, controversial categories between the spoken and written forms — for example a playscript. Although written down, and often printed, no-one disputes that the intention of the playwright is eventually to hear it spoken in performance. However, a successful play will be produced many times, with many alterations to the text; each performance of each production is a distinct artistic experience, and can be recorded and archived for good reason. The original text of the play, however, is not affected by these spoken performances (unless the author alters the text, and thus puts the text through different editions).

Closely related to the playscript is dialogue in the novel — a representation of spoken language, but one which is not intended in the first instance to be presented in the spoken a mode; if, as so often nowadays, the author has his or her eye on a screen or TV adaptation, a new text will be prepared, probably by someone else, and the dialogue may change substantially. Some authors, e.g. Steinbeck, write novels that have most of the characteristics of playscripts, while others, from Faulkner to Runyon, write novels whose language is written to represent the speech of the narrator.

These are complex categories, and capable of great variation; many authors experiment with forms, and there are now electronic novels, where the reader participates in the choice of text he or she reads (there was at least one paper novel written on the ‘programmed learning’ model, where the reader had choices to make at various points). Provision is made in the typology for ‘special’ categories, which are already institutionalised and separately defined, and these can be used to maintain the clarity of classification without requiring the possibility of general mixed categories.

4 Typology

The following classification has been worked out to include most of the information about a text that is likely to be:

- Relevant to the linguistic structure of the text
- Obtainable without disproportionate work.

It is by no means exhaustive; indeed it tends in the other direction. In practical terms, even a very large corpus in today's terms will be unwieldy to use if a large number of parameters of classification are applied. The reason for this is that for a typological analysis to be useful, each binary parameter doubles the number of the eventual *components* of the corpus — sets of texts which all have the same characteristics (see EAGLES (1996a)). Each component has to contain several texts in order to be reliable; so if there are even only 5 binary parameters, there will be 2^5 components, i.e. 32. If the average size of a component is 0.5 million words, then the corpus must be at least 16 million words in size.
The question of relevance is specific to each text, and these categories are not intended to be made mandatory; for example, the identification of the author of a text is relevant towards the literary end of the spectrum but probably not at the bureaucratic end. Nevertheless, a text which seems to be composed in American English, but by a British author and publisher, may require a specific probe into its origins before it can be placed in a corpus without misleading users (this can happen, for example, if a book is first published in the USA, then reprinted in a UK edition from the original plates).
Also, it may in some cases be quite impossible to determine the author, or to chart accurately all the people who have made some contribution to the composition of the text. Scholars are specially trained in this work and may spend a great length of time establishing the authorship of a single text — for literary reasons, perhaps, or forensic ones. So, in the normal use of this typology, it is expected that assignment to the authorship category or to any other will be done immediately or not at all.
The typology can be elaborated for the requirements of a particular application. Entries which were not made in the original establishment of the corpus can be added, and additional parameters can be introduced alongside those advocated here.

E.1. origin — matters concerning the origin of the text that are thought to affect its structure or content.

E.2. state — matters concerning the appearance of the text, its layout and relation to non-textual matter, at the point when it is selected for the corpus.

E.3. aims — matters concerning the reason for making the text and the intended effect it is expected to have.

The study of internal parameters is much less advanced than that of external parameters. The main internal parameters are thought to be:

I.1. topic — the subject matter, knowledge domain(s) of the text.

I.2. style — the patterns of language that are thought to correlate with external parameters.

Sometimes an aspect of the classification is reflexive, in that it contains a statement of its origin, its audience, etc. If this is used as the basis for external classification, it is recommended that *(R)* is placed after the classification. A declaration in the text of a document or transcription is not necessarily accepted as correct, but should always be recorded, and any discrepancy between the reflexive classification and the one chosen for the typology should be justified.

4.1 E.1. origin

The main parameters dealing with the origin of a text are:

E.1.1. people — the various people whose work helped to shape the text.

The roles that are relevant include:

- **E.1.1.1. author** — the person who writes the text, whose original work it is.
- **E.1.1.2. editor** — anyone who alters or advises alteration in the text once it is produced, or who prepares it for a change of form.
- **E.1.1.3. publisher**] — the person identified as legally responsible for the act of publication.
- **E.1.1.4. rights holder** — anyone with a legal right in the published work.
- **E.1.1.5. translator** — a person who translates a text into another language.
- **E.1.1.6. adapter** — a person who alters a text in order to make it suitable for another artistic genre (e.g. screenplay, TV series, illustrated comic book).

Each of these roles may be played by two or more people, or two or more may be conflated in a single person. For each person involved it may be relevant to know:

- **E.1.1.*.1. age** — at the time of composition; between the ages of 16 and 60 broad indications of decade will do; works by children and Senior Citizens should be marked as such since they are known to differ linguistically from mainstream adult text.
- **E.1.1.*.2. sex**
- **E.1.1.*.3. language background** — native language, followed by any other language learned in infancy. Other languages that are relevant to the text should also be noted.
- **E.1.1.*.4. domicile** — at the time of composition, if relevant and discoverable.
- **E.1.2. processes** — production processes that are thought to have had an influence on the text. This is a broad and vague category that is not likely to be used much, since there is only rarely a backwash effect of production processes on the text.
- **E.1.3. circumstances** — any material or other circumstances that are thought to be relevant to the structure or contents of the text. Another broad category, with no fixed values.
- **E.1.4. timing** — matters of dating and timing that are relevant to the construction of the text. Some indication of the year of composition is normally considered to be necessary.

4.2 E.2. state

Texts are selected from the stream of natural speech and writing, and at the point of selection there are a number of external factors which may be relevant.

E.2.1. mode — mode of transmission; whether spoken, written or electronic, essentially.

- **E.2.1.1. spoken** — the language, when selected, is in spoken form, i.e. primarily as a sound wave. Material in this category requires to be transcribed from a sound wave recording before mark-up.
  - **E.2.1.1.1. participant awareness.** — to what extent the participants are aware their speech is being recorded.
    - **E.2.1.1.1.1. surreptitious** — the participants are completely unaware that they are being recorded.
    - **E.2.1.1.1.2. warned** — the participants know something of the recording but are not fully aware of the details.

E.2.1.1.1.3. aware — the participants are fully aware that what they say is being recorded for use in linguistic analyses.

E.2.1.1.2. venue — the location of the participants at time of recording.

E.2.1.1.2.1. studio — the recording takes place in the controlled environment of a recording studio.

E.2.1.1.2.2. on location — for example at home, work, travelling, leisure centre, etc.

E.2.1.1.2.3. telephone

E.2.1.2. written — the language, when selected, is in written form. This material requires to be scanned or re-keyed from a visual image before mark-up.

E.2.1.2.1. printed material — the language, when selected, is in printed format and must be keyboarded or scanned to produce an electronic version.

E.2.1.2.1.1. books — objects with an ISB number.

E.2.1.2.1.1.1. non-fiction

E.2.1.2.1.1.2. fiction

E.2.1.2.1.1.*.1. arts

E.2.1.2.1.1.*.2. economics

E.2.1.2.1.1.*.*. etc.

E.2.1.2.1.2. newspapers — normally daily; usually easy to distinguish from the next category; have a wide readership, contain up-to-date news.

E.2.1.2.1.2.1. general news — the main section of the newspaper where headline news and current affairs are reported.

E.2.1.2.1.2.2. business/economics — sections of the newspaper specifically marked as a separate business or economics section.

E.2.1.2.1.2.3. sport — sections of the newspaper specifically marked as a separate sports section.

E.2.1.2.1.2.4. arts — sections of the newspaper specifically marked as section on the arts.

E.2.1.2.1.2.5. features — sections of the newspaper specifically marked as a separate features section.

E.2.1.2.1.3. magazines — usually weekly or monthly; of higher print quality than the newspaper and usually more expensive.

E.2.1.2.1.4. ephemera — leaflets, pamphlets, brochures, local flyers, junk mail.

E.2.1.2.1.5. correspondence — a sub-classification is under consideration, consisting of automatic and human, official and personal.

E.2.1.2.2. ‘typed’ material — all sorts of reports and documentation (includes word-processor output).

E.2.1.2.3. manuscript — actual handwritten texts.

E.2.1.3. electronic the language, when selected, is in electronic form (e-mail, bulletin boards, World Wide Web pages, etc.) That is to say, it requires conversion neither from aural nor visual image. It may be already marked up, and may only need to have the mark-up converted to meet the conventions of the corpus.

E.2.2. relation to the medium — if written, how it is laid out, the paper, print, etc.; if spoken the acoustic conditions, etc.

E.2.3. relation to non-linguistic communicative matter — diagrams, illustrations, other media that are coupled with the language in a communicative event.

E.2.4. appearance — there may be, for example in advertising leaflets, aspects of presentation that are unique in design and are important enough to have an effect on the language.

4.3 E.3. aims

E.3.1. audience — the person or group of people for whom the text is created.

E.3.1.1. the immediate audience — people forming part of the communicative event, with at least a theoretical opportunity to participate.

E.3.1.2. the wider audience — for example, a readership, viewership.

Cross-cutting with this are the following:

E.3.1.*.1. audience size — for example, under 5 / 5–20 / 21–50 / hundreds / thousands.

E.3.1.*.2. audience constituency

E.3.1.*.2.1. members of the general public

E.3.1.*.2.2. informed lay people

E.3.1.*.2.3. professional people

E.3.1.*.2.4. specialists

E.3.1.*.2.5. students, trainees

E.3.1.*.2.3.1-n. list of professions

E.3.1.*.2.4.1-n. list of specialists

E.3.1.*.2.5.1-n. list of study and training courses

E.3.1.*.3. author–audience relationship

E.3.1.*.3.1. distant — no personal acquaintance, and further separated by institutional roles that depersonalise.

E.3.1.*.3.2. neutral — no personal acquaintance, but both author and member of audience are considered as individuals.

E.3.1.*.3.3. close — personal acquaintance, or the assumption of it. (Note this is a place where the reflexive categorisation may conflict with the judgement of the researcher, e.g. in commercial circulars which are personalised by computer working from mailing lists)

E.3.2. outcome intended

E.3.2.1. information — an unlikely outcome, because texts are very rarely created merely for this purpose. Mainly reference compendia.

E.3.2.2. discussion — polemic, position statements, argument.

E3.2.3. recommendation. — reports, advice, legal and regulatory documents.

E.3.2.4. recreation

E.3.2.4.1. fiction — including faction.

E.3.2.4.1.1. novel/novella/short story

E.3.2.4.1.2. historical, thriller, etc.

E.3.2.4.2. non-fiction

E.3.2.4.2.1. biography

E.3.2.4.2.2. autobiography

E.3.2.4.2.3. letters — the published variety, not correspondence.

E.3.2.5. religion — holy books, prayer books, Order of Service. As well as a category of text on external criteria, religion is an important topic, which is allocated on internal criteria. Hence, unless care is taken this category may overlap with others in this hierarchy, and it is recommended that it is only used for texts that cannot be properly classified under other headings. To have a religious topic (e.g. a biography of a religious figure) is not enough for classification here.

E.3.2.6. instruction

E.3.2.6.1. academic works — these are written for a specialist audience (see E.3.1*.2.4.). It follows that the author is also a specialist.

E.3.2.6.2. textbooks — these are written for a student audience, and therefore the author has an appropriate reputation for expertise, whether academic or professional.

E.3.2.6.3. practical books — these are written to teach and guide the exercise of practical skills.

5 Internal criteria

Here we will develop the argument that two central parameters of the classification of texts are better described using internal, or text-linguistic, rather than external, or sociocultural, criteria.

The parameters are *topic* and *style*. Attempts to analyse them in order to provide a basis for classification are usually expressed in terms of external criteria and are unsatisfactory on several counts. They tend to be unsupported by scientific evidence, and therefore unreplicable; unsupported also by wide general agreement, so that we see several different versions co-existing without good criteria for preferring one to the others. There is no control over the granularity of the classification, and not sufficient dimensions for the very large amount of possible cross-classification. The prospect of topic and style being suitable for attribute/value analysis are extremely remote, and unattainable in the present state of our knowledge.

External criteria follow distinctions and classifications that are already available in the culture, both of documents and speech events. They may not use all of a pre-existing classification, and they may use facts in the culture that are not traditionally used to classify texts, but the defining characteristic is that the criteria for an external criterion are in the culture and not in the language.

For example, consider the category of ‘quality newspaper’. This is an informal term that is mutually exclusive with ‘tabloid’, but there may be other categories as well. In the UK, for example, price used to be an important distinction, but there is currently a price war going on and this is no longer a sound criterion — a few years ago quality newspapers had no news on their front page, but only advertisements and notices. Tabloid newspapers use a sheet size that is approximately half that of quality newspapers. The editorial in a tabloid consists of a few pithy sentences on an issue, while in the quality papers the editorial can be a learned essay of two or three thousand words setting out the editor’s view.

All these are external criteria. The point is that one does not look inside the papers at the kind of language they use in order to classify them, although this would be quite simple and rewarding. There is, for example, a well-known style of headline-writing in some of the tabloids, using short words and very few grammatical ones (*Nude Peer in Sex Storm*), which might form an internal criterion, but that is not considered in the tradition of classification.

6 Topic

Topic is the lexical aspect of internal analysis of a text. Externally the problem of classification is that there are too many possible methods, and no agreement or stability in societies or across them that can be built upon. There are semantic classifications, such as Roget's famous *Thesaurus* (Roget, 1962); there are bibliographical ones, such as the Dewey Decimal Classification. Educational systems divide knowledge up into a myriad of confusing hierarchies, and keep changing their minds. Hierarchies and other organisations of terminology can be used as topic identifiers, terminologists organise term banks on cross-cutting principles, and organisations like ISO play a part in the attempt to stabilise terminology, and thus provide a means of organising conceptual knowledge.

It is thus a gross oversimplification to attempt to produce a single model of topic that embraces everything that can be written or spoken about, neatly arranged in discrete boxes, such that each text can be placed in one box, with only a small percentage of overlap or doubt.

In Roget's *Thesaurus* (Roget, 1962), the English language is primarily divided into six global categories: 'Abstract Relations', 'Space', 'the Material World', 'Intellect', 'Volition' and 'Affections'. From here, there are sub-categories which further divide the world into more manageable areas, for example the category 'Affections' divides into 'Affections Generally', 'Personal', 'Sympathetic', 'Moral' and 'Religious'. At this level we may already find difficulty in accepting a fundamental difference between, for example moral and religious affections, especially if one's moral values are fully embedded in religion. The boundaries become even more vague the further we separate language into smaller branches. Under the sub-categories come further groupings of topics. Following the category 'Affections' and, say, the sub-category 'Personal' we find the groupings 'Passive', 'Discriminative', 'Prospective', 'Contemplative' and 'Extrinsic'. These further subdivide, finally, into the 1,000 topics by which we can conveniently categorise language.

At this level of categorisation, the boundaries between the proposed topics become very hard to distinguish. The amount of cross-reference also becomes immense. Following the same branch of 'Affections – Personal' further down, under 'Passive' we find such topics as 'Joy', 'Suffering', 'Pleasurableness', 'Painfulness', 'Content', 'Discontent' and so on. On this level, however, how can we justifiably differentiate between affection which comes under the topic 'Affections – Personal – Prospective – Dislike' and one which is 'Affections – Sympathetic – Social – Hatred'; or between language which is classed as 'Communication of Ideas – Means of Communicating Ideas – natural – conventional – letter' and another topic 'Communication of Ideas – Means of Communicating Ideas – written language – correspondence'. Is it desirable, in the classification of language into topic, to do so?

The boundaries between the topics are ultimately blurred, and we would argue that in the classification of topic for corpora, it is best done on a higher level, with few categories of topic which would alter according to the language data included.

There are numerous ways of classifying texts according to topic. Each corpus project has its own policies and criteria for classification, which, indeed, is the underlying impetus for this paper, to offer common guidelines for classification of texts. The fact that there are so many different approaches to the classification of text through topic, and that different classificatory topics are identified by different groups indicates that existing classification are not reliable. They do not come from the language, and they do not come from a generally agreed analysis. However they are arrived at, they are subjctive, and the subjective categorisation of language is bound to lead to diversities in the categories established since the resulting typology is only one view of language, among many with equal claims to be the basis of a typology.

The NERC report offers a summary of the classification systems used by major corpus project in Europe. Updated tables are attached as appendices. There are 35 categories in the subject typology report. Extensive though this may seem, this is only a consolidated summary of the common categories of the corpus projects. There are many different 'topics' by which texts are classified in different corpora. Some classificatory systems go into more detail than others. In the Danish Corpus we find an extensive list of topics from 'transportation' to 'music', 'business' to 'environment' (see Norling-Christensen, 1995, Annex 1). There is even a separate topic for the 'EU'. In this kind of classification system it will be difficult to classify a text according to one of the 66 topic areas offered; most texts will fall into more than one topic in this system. For example 'law' and 'crime' and 'society' are all listed as separate topics, where they are inextricably linked.

As is explained in the NERC Report:

> *Generally applied rather than corpus-specific texts have been distinguished.*
>
> (Calzolari et al., 1995)

Even though this table shows the 'generally-applied' categories of topic, there is great variation between classification techniques. Looking at the list of subject categories proposed, we find 'Leisure' as a separate category from, say 'Sport'; the category 'Science' and then, separately 'Physics', 'Biology', 'Chemistry'; 'Finance' and 'Economy'. This would cause problems in the practice of labelling texts under such topic headings. The boundaries between them are by no means fixed no clear. It would be down to one person, or a group of people to decide where to place most of the texts and this is not satisfactory procedure either from the point of view of resources or from the need for replicability of decisions.
It is recommended here that topic be determined through internal criteria to provide linguistic justification for the resulting categorisation of texts through topic. Guidelines will be based on results from objective computer-assisted analysis of the texts to be included in the corpus.
Here we will review work which has been done to date on the analysis of linguistic features of texts with computer software. This will give a preview of the procedures by which we can select texts and maintain corpora in the future. Some of the main research in this area has been done by Phillips and his work is reviewed here since it intimates the kind of analysis that is likely to be used in order to reinstate topic as a useful segment of corpus typology.
Phillips (1983) offers a rationale for the determination of the topic of a text through an objective, quantitative distributional methodology. What we here are referring to as 'topic' in the classification of texts for the purposees of the present typology, Phillips refers to as 'aboutness', i.e. 'the psychological perception of subject-matter'. He claims that the 'aboutness' of a text is due to the global patternings in the text, or the text's 'macrostructure'. In his thesis he analyses the macrostructure of texts by computational means, so that the results are derived from the text itself and not from external structures. He emphasises what we would urge here, that in any statement about subject, or 'aboutness', of a text, the basis must be an objective analysis of linguistic features. It would be a mistake at this stage of our awareness to try to map language direcly onto external structures. As we have seen in the NERC report, this proves impossible in practice.
In order to say something on the 'aboutness' of a text, the human reader has to understand that text. This opens the doors to subjective interpretation of the kind that we wish to dispel from the classification of texts for inclusion in corpora. An objective investigation of the linguistic features which make up the aboutness of a text is a real option now with the advent of computer corpora and sophisticated software for analysis. With this kind of objective analysis, Phillips points out that not only do all the results come from the text itself without outside interference, but that the procedure can be repeated for any number of language units (which, for Phillips, are chapters in a book, but could as well be any length of text that is long enough to show clusterings) for similar analysis.
Phillips further explains that the aboutness of a text cannot depend directly on the details of linguistic form, such as will be found in a lexico-grammatical analysis of it, since a reader will invariably be able to summarise a text and say what it is about without direct reference to the specific lexical items used in that text. Ultimately, however, the understanding of the text and evocation of topic can only be somehow derived from the actual text itself. The aboutness of a text would appear to be separate from its specific representation in language but at the same time it is something evoked in our consciousness only through the symbolism of language. Phillips therefore concludes that we need to look towards the global patternings in a text and it is through the analysis of these global patternings that we gain insight into the aboutness of the text. This global patterning is the macrostructure of the text and it is on this higher level, at the level of the macrostructure, that the text is analysed by sophisticated software for information concerning aboutness.
It is important, therefore, for the purposes of this report, to determine firstly the linguistic features which are relevant to the concept of topic or subject, and secondly to discuss the possiblities for the analysis of such features. Analyses at the level of macrostructure have also been carried out successfully in the Aviator project which was a computational continuation of the results of Phillips' work.
Phillips claims that:

> *[...] in any discussion of aboutness, 'situation' is an important notion and [that] it is likely that different relations between texts and their contexts could form the basis for a text typology.*

This Firthian objective firmly situates text within a higher level of constraint, i.e. the context of situation. Both texts and contexts of situation are far too often classified through external means and designated categorisations which reflect, more than anything, a subjective interpretation and classification of our envirmonment. If the 'aboutness' of a text is dependent on its relation to its context, we must determine the relevant features and analyse these features objectively, through computer-assisted analysis, to avoid unecessary distortions and impositions.
Phillips argues that whilst the aboutness of a text is not directly dependent on linguistic representation, language concepts are ultimately contructed by lexical items or associations of lexical items. It is his contention that the aboutness of a text can be determined by patterns on a higher level of analysis. The patterns in the text are what he refers to as the text's macrostructure and it is from this that we can gain insight into features of aboutness:

> *I contend that with all semantic, syntactic and lexical markers of a science text neutralised, the meaning of the text is not exhausted but that a distributional analysis of what remains will reveal the presence of global patternings, which I call macrostructure.*

To analyse the macrostructure of a text, we look at the collocational associations in the text. We then must determine which are characteristic of a certain style, topic or genre. Phillips' methodology is a distributional analysis of texts (since it is a distributional analysis we must therefore also advocate the inclusion of full texts into the corpus so that all linguistic features are represented). The first step in Phillips' methodology is to identify the frequency of occurrence of each node–collocate pair within a range of four words on either side of the node. This is calculated taking each word in the corpus as node and the words within the span of four words on either side as collocates. The frequencies of the collocates are then collated to the level of the lemma (although the justification for this decision might be debatable today, with the greater understanding of corpus patterns that we have and the much greater computer power that we can bring to arithmetic tasks). Particular associations between lemmas are identified across the chapters (here, of science books) and the difference in the macrostructure of each assessed. In effect, the macrostructure is defined through the analysis of the patterns of the intercollocation of collocates.
Phillips employs Ward's method of cluster analysis to observe attractions between lemmas. He extends the association to include associations which regularly have at least two lemmas in common. Once a cut-off has been set (so that all the data does not merge into one cluster), each node is assigned a 'similarity coefficient' based on its collocational environment. Nodes with similar 'similarity coefficients' are merged and the process repeated for the new cluster. We thereby analyse the similarity of networks of lemmas which form the macrostructure of the text. For two networks to be similar, they must both contain at least two of the same lemmas, and at least one lemma must be a member of the nuclear set of nodes in each network.
The analysis proved the existence of patterns of association within the chapters. Phillips claims the two major findings of his study to be firstly "*the discovery of the existence of syntagmatic lexical sets*"; and secondly that "*the sets identified in the analysis are meaningful*" if we interpret the syntagmatic organisation of words into lexical sets established by the analysis as the conceptual concerns of the text.
This analysis gives rise to the notion of lexical macrostructure and stresses the importance of the intercollocation of collocates. A semantic interpretation of the lexical evidence leads us to a notion of what the text is about.
Phillips' work formed the basis for the part of the AVIATOR project coordinated by Renouf which aimed to develop software for monitoring changes in language patterns as corpora develop over time. There were two main goals of the project, firstly to identify new word occurrences, collocations and word-combinations; and secondly, and most relevant to the present purpose, to identify clusters of words in a text that to some extent reflect its conceptual content. Since the aim was to monitor change, among other things, there was a constant stream of language data (from *The Times* national newspaper) over the three-year duration of the project.
The software developed as part of the Aviator project gives an example of the way in which the topic of a text can be retrieved automatically from the text iself. The methodology was much the same as Phillips'.

First there was established a stop-word list of the most frequent words in the corpus, whose patterns of occurrence appeared to be stable over several years. These words were then overlooked in favour of those that seemed more responsive to their general environment. A collocate bank of the remaining words was created, holding frequency information.

Clustering software then processed the most frequent nodes to create clusters. From these clusters we can then see the aboutness of a text. The usefulness of the machine analysis was evaluated by comparison with the results from humans given the task of summarising the same texts or saying what they were about. The software thus created lists of keywords and clusters of keywords which are indicative of the topic of the text.

On the basis of the results of the Aviator project, it seems feasible that the topic of texts can be determined through analysis of linguistic criteria with software such as that used by Phillips (1983) and Renouf et al. (1993). A typology of texts according to the topics emerging from such analyses could then be established.

It is not foreseen, however, that the types of topic identified through this methodology ever be of finite number, or that orderly hierarchies will emerge. Rather, the relation between one text and another is best visualised as a pattern of overlap, and the clusters of collocates as a sophisticated extension of the familiar 'keywords'(see below), on which most of today's content queries are based. The model of language underlying this research is dynamic and the categories of analysis are not preordained; so no existing list of possible topics can be precisely relevant; no doubt useful classificatory patterns will emerge when a large number of texts have been processed.

Since the NERC report it has been regularly recommended that corpus building should have in mind the eventual goal of adding the time-change dimension and moving towards the condition of monitor corpora (see EAGLES (1996a)). Such a move will require the development of new software tools, which will have to be fully automatic, very fast and sensitive to quite small changes in the language structure. If Phillips' computerised aboutness allows us to represent topic without human intervention, it will be possible to control additions to corpora with a new accuracy; there could also be important byproducts in information retrieval and knowledge retrieval.

The topic of a text — what it is about — is probably unique to the text. However, this is not in practice very helpful, since for a variety of purposes people need a means of associating texts together according to topic, and of differentiating them on the same lines.

This has been recognised in practice for many years, and it is now commonplace to classify texts using a crude system of keywords. Authors are routinely asked to supply with a manuscript a small number of words (including short phrases) under which a text an be indexed. It is more than likely that these are words and phrases that appear prominently in the text itself, and thus constitutes a rough-and-ready method of analysis with some reference to internal patterns.

Related to the keywords method is a much more sophisticated method of classification — the summary or abstract, also routinely supplied by authors. A summary is supposed to be a short text that brings out the main message of the full text; as such it is likely to use language which is characteristic of the full text, though it is by no means restricted to that language, and frequently uses words of more generalised reference in order to summarise efficiently.

In the days of corpora of unlimited size, including complete texts of widely varying lengths, it is necessary that internal criteria will be formal and automatic in nature. Human abstracting services are expensive and slow compared with machine processing speeds. Hence we should recommend that operations such as abstracting and providing keywords should be developed for the internal classification of topic.

Keywords provide a set of labels or titles for indication what is inside a text, and how it relates to other texts. Abstracting gives indications of the kind of sentences and arguments that are to be found inside. These are clearly the kind of thing that people want, and find reasonably easy to understand, though they are not readily related by machine to the original texts.

However, automatic abstracting is the aim of a number of research projects, and one type in particular is of interest — where the machine picks out and fits together sentences of the original text in order to make the abstract (Hoey, 1991: 113–4, 142, 160). In such systems the relationship between the text and its summary is simple and explicit.

Automatic selection of keywords is another obvious step, and here Phillips' notion of aboutness is of considerable practical importance. In principle, using aboutness could lead to a hierarchy of classification of texts that would at its 'highest' level produce acceptable keywords, and then a pyramid of more and more detailed lexical analysis. Texts could be compared automatically and the overlap between them could be presented in terms that would be easy for a human to understand.

The combination of the Aboutness Pyramid and an automatic abstract of a desired granularity should give the best available tools for navigating corpora and other archives. Although neither of these tools is yet fully available, research is vigorous in both areas. Aboutness has the added benefit that it is language-independent; it is not yet clear whether any language-independent abstracting systems will give good enough results to be adopted.

The main limitation of the application of Phillips' ideas is the minimum length of text that will contain sufficient instances of the important words to allow the clustering techniques to work effectively and to depress the effect of other dimensions of patterning, which are apparent locally but do not repeat often enough to be included in the specification of topic. Each text has its own distributional properties, and it is not possible at present to determine a suitable general figure. By way of guidance, it can be noted that Phillips gets satisfactory results from considering each chapter in a scientific textook as a separate text, and Aviator was able to work with a variety of more popular texts of a few thousand words in length. Hence the prognostication is that if a corpus is divided down to very short entries, like newspaper reports for example, the aboutness technique will not have enough evidence to go on; however such tiny texts will surely be classifiable much more directly if it is felt necessary.

Phillips' work relates primarily to the specialised language of science and technology, where we assume there is less variety in the vocabulary than in general text. Parallel to his work is that of Yang (1986), who devised methods for the detection of technical terms in open text. Since the strategy involves the distribution of terms across texts, it is eminently suitable for application to corpora, and although as published it focuses on the individual terms, it can by extension be used to classify texts according to the distribution of technical terms within them. More recent relevant work on automatic term recognition which can be likewise used as a basis for text classification is due to Frantzi & Ananiadou (1996) and Lauriston (1996), the latter offering a substantial bibliography.

To summarise this discussion, we can look forward to being able to analyse texts in such a way that their topics will be retrieved automatically. This involves greater sophistication in the development of internal criteria than is widely available at the present time, but research towards it should be vigorously pursued. Two texts processed for aboutness can be compared at at least two levels of generalisation above the sentences of the texts themselves; those keywords or clusters that occur in both form an overlap in topic that can be evaluated by the investigator, and those that differentiate the texts provide urther helpful guidance. The Aboutness Pyramid is potentially a fully practical device for indexing texts in a corpus and comparing any two texts.

In the meantime, it is possible to use external criteria in order to arrive at classifications of subject-matter which, precisely because of their correlation with public and social distinctions, may well be superior for some purposes to the inherently less tidy internal picture of topic.

7 Style

Style is a notorious term, because it is used in so many different ways by researchers from several disciplines, and has popular meanings as well. It is used here to mean the way texts are internally differentiated other than by topic; mainly by the choice of the presence or absence of some of a large range of structural and lexical features. Some features are mutually exclusive (e.g. verbs in the active or passive mood), and some are preferential, e.g. politeness markers and mitigators.

As with topic, there are no institutionalised schemata to which we can turn. Although a great deal is talked about style, and there are several parameters of organisation proposed in the literature, there are no agreed standards for any one parameter. For example, most students of language believe that a parameter of formality is required, and terms like ‘formal’, ‘informal’, ‘colloquial’, etc., are freely used and not always well defined. In one of the most influential proposals (in its time), Joos (1961) set five levels of ‘frozen’, ‘formal’, ‘informal’, ‘colloquial’, ‘intimate’. However useful this was, Joos might just as easily have proposed four, six, seven or twenty, because the motivation for five levels was mainly convenience.

Halliday et al. (1964) distinguishes between registers according to three dimensions, ‘field of discourse’, ‘mode of discourse’ and ‘style of discourse’ (in later work he changed the label ‘style’ to ‘tenor’). Style is used here to refer to the relations among the participants. Halliday suggests a primary distinction into ‘colloquial’ and ‘polite’, a primary distinction which is adopted in many of today’s dictionaries. He also claims that the styles of discourse must be treated as a cline with categories such as ‘casual’, ‘intimate’ and ‘deferential’.

In most of today’s dictionaries there is a classification of certain words according to style. The most common distinction is between formal and informal language. Even within these broad categories, one dictionary may classify a word a formal where another may not. Since such decisions rest with the lexicographer, it is not surprising that inconsistencies exist between interpretations of the style of a word. ‘Formal’ is defined as a label which

> *[. . .] usually means that this is a word which is most likely to be used in highly formal writing or in formal speech read aloud.*
>
> (Summers et al., 1993)

Outside lexicography, categories such as ‘formal’ and ‘informal’ are usually defined by the type of context of situation in which the language is found, for example formal language will be found in forms, pronouncements, official documents and the like.

Some dictionaries (e.g. Collins Robert (Atkins et al. (eds), 1987)) have grades of formality or informality so that there are sub-groups of style within the formal and informal labels. For example we find a class for a word which

> *while not forming a part of standard language, is used by all educated speakers in a relaxed situation but would not be used in a formal essay or letter, or on an occasion when the speaker wishes to impress.*

Further down the cline, is a word which

> *indicates that the expression is used by some but not all educated speakers in a very relaxed situation.*

What would then qualify as a ‘relaxed situation’ and is it expected to be the same for all these ‘educated speakers’?

In Webster’s Dictionary (Woolf et al., 1976) the style labels are ‘slang’, ‘nonstandard’ and ‘substandard’, each defined by its relation to a standard or norm. Most of the dictionaries provide style labels for words which are not though to be part of standard language. Some dictionaries go as far as to provide labels for the particular sub-language in which a word is predominately used. Sansoni (Macchi (ed), 1988) offer a whole range of style labels which situate words into different sub-languages, for example *linguaggio burocratico,* *commerciale, cinematografico, giornalistico, infantile, marinaro, scolastico, universitario* and so on; the list is substantial.

'Literary' style is also a popular class of style in dictionaries. Alongside this we also find 'poetic' style but it is not clear how the two relate. Is poetic style a sub-class of literary style? Do we assume both can be considered under Webster's class of 'nonstandard'. Do we then class 'slang' as a style within 'substandard' with 'informal' as a kind of bridge between the two? In which case where would we place the category 'offensive' which is also a common style label. Is this on the opposite pole from 'polite'? If so how does it relate to Halliday's term 'colloquial', which is the alternative to 'polite'?

There seem to be so many different style labels used in dictionaries that it is difficult to see how they relate to one another in the proposed cline between the colloquial and polite, informal and formal. There is a whole mass of categories which accompany the informal/formal distinction, though the structure or hierarchical value of each is unclear. Other style categories used such as slang, colloquial, common, polite, rare, popular are not often defined by the dictionaries that use them, as if it is assumed that they have self-explanatory boundaries.

In practical lexicography, where compilers face decisions about style classification on a daily basis, it is clear that there is absolutely no natural consensus about the formality of an expression. People are all different, for a start, and likely to have different baselines from which they make their judgements. They have different attitudes to propriety in language use, and they have different language experience, where the influence of their ages and the regions of their upbringing play a part. They are also unsure of the way in which the terminology should be applied. Is it good or bad or neutral to be colloquial? Where does slang fit in?

There is also an underlying issue that never resolves itself — some choices made in writing are considered informal, but the same choices in speech are just neutral. Is there a systematic displacement of the spoken language with respect to the written so that its realisations are always a notch or two down in formality from writing? Definitive answers to such questions require an alignment of internal and external criteria for which we will have to wait for some time, but the questions indicate the lack of clarity that characterises the description of style.

Style is the other criterion for which we recommend an automatic analysis based on linguistic criteria, and we recognise that no suitable package is yet available. There are signs of progress in the statistical analyses of texts such as those already being used in literary and forensic linguistics for questions of authorship. What is important for the purposes of a text typology is that the criteria for categorisation according to style be established objectively.

Biber (1988) offers a methodology for the objective grouping of variations in English texts through statisitcal analyses. The analysis is based on the identification of different clusters of linguistic features across a range of written and spoken English texts. Using a technique called factor analysis, he successfully identifies the linguistic characteristics of texts by an objective analysis of the language data. Biber's fundamental claim is that the frequent co-occurrence of a group of linguistic features in texts is an indication of an underlying function shared by those features. Therefore we can identify which linguistic features consistently group together to perform a particular communicative function, which features co-occur and which features are mutually exclusive. We can interpret these results to establish a correlation between variations in linguistic groups and function. Not only that, but we can use the analysis to objectively define a set of texts which belong to each variation in the English language, with the intention of using this kind of classification technique to categorise new texts for inclusion in a corpus. We envisage, however, that the categories would be continually refined and not fixed, to cater for the steady flow of language into and out of the corpus.

Biber, who sets out to identify fundamental variations between speech and writing, works with texts taken from the LOB (Lancaster–Oslo–Bergen) corpus of written English and the London–Lund Corpus of spoken English. These texts have previously been categorised into various genres for the corpus, based, we assume, on external criteria. Categories in the LOB corpus include such genres as 'press reportage', 'editorials', 'press reviews', 'religion, 'skills and hobbies', 'popular lore', 'biographies' and so on; in the London–Lund corpus categories are 'face-to-face communication', 'telephone conversation', 'planned speeches', 'broadcasts', and so on.

Biber differentiates between 'genre' which he uses to describe categorisation according to external criteria, and 'text type' to refer to groupings of texts that are similar with respect to their linguistic form, irrespective of genre categories.

The first step in Biber's methodology for definition of text type through internal linguistic criteria is to review any previous research on linguistic features which will identify potentially important features. Here the aim is not to establish which are the more important linguistic features (since this should be done objectively through statistical analysis of the data) but to provide as wide a range as possible of possible significant linguistic features. Biber identifies 67 linguistic features of English text which he includes in the analysis (Biber, 1988: 86–87 offers a full list). Having chosen the texts to work with, the investigator must then obtain frequency counts of all 67 linguistic features in each of the texts used in the analysis. These figures must then all be normalised for a certain text length (here 1,000 words) to ensure comparability. The mean, minimum and maximum frequencies, the range between these frequencies and the standard deviation are all calculated before the factor analysis begins. Features with very low frequency of occurrence are discarded as insignificant.

The aim of the factor analysis is to identify groups of linguistic features which co-vary. To say they co-vary does not necessarily mean that they co-occur but that there is a definite correlation (or inverse correlation) between their frequency counts in the texts. This means that it is equally important to learn that two features co-occur as it is to learn that two features are mutually exclusive, or that the presence of one will point to the absence of the other.

In this way Biber establishes factors, which are sets made up of different linguistic criteria. The factor will include both negative and positive 'weightings' which indicate linguistic features which co-occur (positive weightings) or whose presence marks the absence of the other (negative weightings). The most significant features are those which have the largest weighting irrespective of whether this is shows attraction or repulsion. Biber ultimately establishes 7 factors which are representative of the texts in the study and are mutually exclusive.

We now turn to the interpretation of these factors by which have been identified from the texts and which will, in turn, be used to identify styles of text. Up to this point identification of the features, calculation of their frequencies and the extent of their co-variation has been objective and automatic. The factors which have been established are then interpreted by Biber to show the communicative functions to be associated with each factor. The reader is referred to Biber (1988: 104–114) for a full interpretation of the factors identified. For convenience we show here some examples of some of the kind of results obtained.

He finds, for example, that in the first factor there is a high weighting of nouns, private verbs, present tense, pronouns, WH–questions and so on. This is what the statistics tell us. Biber then interprets these patterns of linguistic features in terms of the style of the text. A high density of nouns (the primary bearers of referential meaning in a text), for example, indicates a great density of information. Longer word length suggests more specific, specialised meanings, and the type/token ratio also points to a high density of information as well as very precise lexical choices resulting in an exact presentation of informational content. An interactive style is indicated by a high weighting of present tense forms indicating a verbal rather than a nominal style, also of pronouns and WH–questions. The interpretation continues but there is no need to go into such detail here. The important point is the value of such analysis in the identification of style in text.

Biber relates the factor to the text and the text is then given a factor score. This enables us to relate the linguistic groupings of the first factor, for example, to two separate communicative parameters, namely that the primary purpose of the writer/speaker is informational while adopting an interactive, affective and involved style; and secondly that the production circumstances are such as to allow careful editing possibilities, allowing precise lexical choices and an integrated textual structure.

The dimensions identified by Biber are then related to the genres which have already been pre-established for classification of texts in the corpora used. It becomes obvious that these genres are not coherent in terms of their linguistic features. Some dimensions are equally applicable to various genres, while within one genre there may be a wide range of variation. As example, Biber shows that within 'academic prose' there is a wide range of variation, as is there within the genre 'conversation'. This would indicate that we can not take these genres as representative of written and spoken English respectively, nor can we take the analysis of any one text to be representative of the genre. The study also shows that there is no one, absolute difference between speech and writing, but that there are, instead, several dimensions of variation which are manifested in both.

The dimensions identified in this study succeed in defining a set of relations among texts which can be used for an overall text typology. As Biber points out, since the texts in the study cover a wide range of discourse types in English as well as the linguistic features of many communicative functions, the dimensions of variation that he establishes provide parameters of linguistic variation which he claims exist among English texts as a whole.

The implications of such research are many. The methodology outlined here has already been applied to various other kinds of research, namely the comparison of dialects within English and identification of linguistic variation between British and American English through analysis of lexical and syntactic features; for stylistic comparisons not only between specific authors, but of the historical evolution of English written texts. Particularly relevant in the discussion of style is the stylistic research done by Biber & Finnegan (1988) who use the technique outlined here in conjunction with cluster analysis to identify eight styles of stance in English texts, e.g 'Cautious', 'Secluded from Dispute', and so on. Within a multi-lingual environment, the consequences of this kind of analysis are important since this approach to textual analysis can be used for other languages, if similar features for classification can be identified, and therefore we can provide comparable text types within a multilingual environment.

Biber works with a considerable number of linguistic features; by contrast Nakamura has developed techniques for examining the classificatory role of individual features, using Hyashi's Quantificational Method, Type Three. In a series of papers (Nakamura, 1986; Nakamura, 1987; Nakamura, 1992; Nakamura, 1993) he has shown how texts, genres and corpora can be differentiated according to the incidence of chosen features. The basis of Nakamura's method is the means by which a large number of individual observations or classifications of language can be grouped together to show broad general tendencies. Although the statistical processing allows up to 14 different parameters of grouping (called 'axes'), in practice the top three usually account for the vast bulk of the data; also since the results are easiest to interpret when presented in the form of a diagram, three dimensions is the most complex structure that can be shown with ease. Nakamura's diagrams show graphically the relative distances of linguistic items from each other, or the relative distances from each other of texts or corpora containing these items.

Nakamura's techniques can be brought in to classify in a wide range of circumstances. The distribution of pronouns in texts, or the distribution of grammatical tags, or the distribution of the vocabulary; there can be a lot of corpus material or a little — Nakamura works on corpora from 1 million words to 200 million, and can use a single simple criterion or a complex set of them. From the above discussion it can be concluded that the specification of style in terms of internal criteria is feasible, but that work has only just started. Biber's classifications depend on the relevance of earlier, individual studies, and have not yet incorporated and feedback from large corpus analysis, which promises to offer new dimensions. Nakamura's most recent work (Nakamura & Sinclair, 1995) investigates collocational patterns in multi-million word corpora, showing that his analytical procedure can cope with such a scale. But these are only shafts of light in a vast darkness. As Biber recommends, progress will best be made by frequent cross-checking between internal and external criteria so that each establishes a framework of relevance for the other.

7.1 Interim recommendation: style

There is a small number of parameters of style that are frequently invoked, and will serve as an interim classification. They are offered below without definitions, and the user should add his or her own precise use of the terms.

I.2. style

- **I.2.1. Formality**
  - **I.2.1.1. informal**
  - **I.2.1.2. formal**
- **I.2.2. Preparation**
  - **I.2.2.1. considered**
  - **I.2.2.2. impromptu**
- **I.2.3. Communicative grouping**
  - **I.2.3.1. conversational group**

I.2.3.2. speaker/writer and audience

I.2.3.3. remote audiences (e.g. radio, TV)

I.2.4. Direction

I.2.4.1. one-way

I.2.4.2. interactive

These categories largely reflect external types; as the research reported above is showing, these generally accepted labels are not clearly definable, and no doubt rest on bundles of detailed selections of grammar and phraseology that have still to be discovered. One newcomer to the categories of style is mentioned below.

7.2 Collocation

There is a lot of interest at present in the feature of collocation, the co-occurrence of words within a short span in a text. Collocation is both linguistically powerful and easy to identify in computational processes. It is notable that collocation is used as a criterion both for topic and style, which may help to explain its popularity.

In topic, the clustering of collocates helps in disambiguation of the individual words, and gives a more accurate identification of the topic of the text than simple keywords. In style, types of word conbination are clues to style types. More recently, collocation has been used to classify genres, showing that the same word is characteristically associated with certain collocates in particular types of writing and speaking.

8 Spoken language

Much of this section is drawn from the report of the EAGLES Corpus Subgroup on Spoken Language Corpora (EAGLES, 1996c).

Most general corpora include an element of transcribed spoken language — broadly or even orthographically transcribed. A substantial quantity of spoken data, particularly impromptu recordings of ordinary people talking together, is regarded as one of the richest sources of insights into language.

There are corpora that consist only of spoken material, just as some restricted corpora containonly written material. There are also some quite different corpora (sometimes called speech corpora to point up the distinction), which are compiled by researchers into the intricacies of phonetics.

The crucial distinction is between a corpus of spoken language that is suitable to be run in parallel with a corpus of written language, to provide general evidence of the grammar, lexis, phraseology and style of the language; and a corpus of spoken language, often called a speech corpus, which is put together to further the research of the speech community into the nature of phonetic substance.

A sequence of conferences and seminars (see, for example, Leech et al. (eds), 1995) has clarified the issues and shown that the interests of the two groups overlap and with the direction of research at present and the development of technology they are overlapping more and more. However, much of the speech corpus is too specialised for inclusion in a general corpus, and in the EAGLES typology of corpora would be called special corpora (EAGLES, 1996a).

In practical terms most corpus providers see the need for making an attempt at including some transcribed spoken language, or at least making provision for the inclusion of such material at a later stage. Hence in this paper there will be a brief presentation of the issues that should be kept in mind — what aspects of spoken texts need to be accommodated in the preparation of a corpus of written material.

8.1 Definitions

Corpus — A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.

Spoken language corpus — A corpus consisting of recordings of speech which are accessible in computer readable form, and which are transcribed orthographically, or into a recognised phonetic or phonemic notation.

8.2 Development Phases

Three phases can be distinguished in the development of a corpus of spoken language:

1. The *pre-recording phase*: In this phase the objectives of the corpus are articulated, and from that the specification of the types of recording can be derived, including the physical situations, the number and type of speakers, the communicative events and the topic areas to be covered.

2. The *recording phase*: In this phase the management of the recording is devised and controlled — the nature of any intervention, and the state of awareness of participants. The recording equipment is specified, and other technical details such as the placing of microphones.

3. the *post-recording phase*: In this phase the transcription is specified and carried out, and the transcription is processed for inclusion in the corpus

Note: In the assembly of data for the direct study of the speech signal, the above issues may well be replaced by others. The capture of genuine communicative events may not be important, whereas they are central for a general corpus of spoken language; the technical specifications may be much more elaborate; intervention and experiment are common, whereas they are excluded from material in the general corpus. The transcription and the analysis of the sound wave may be extremely detailed. The following types of speech data are regarded as special corpora from the perspective of general reference corpora, and are therefore not further considered here.

- read aloud isolated phonemes
- read aloud isolated words
- read aloud isolated sentences
- read aloud text fragments
- semi-spontaneous speech (e.g. numbers, alphanumerical expressions)
- spontaneous speech about a predetermined subject (e.g. retelling a story)
- factorial experiments (subjects performing in situations remote from normal communicative behaviour e.g. passing information about maps and routes (possibly across a visual barrier)).

8.3 Participants

Age and sex are already features for the written language. See above (E.1.1.*.1.): between the ages of 16 and 60 broad indications of decade will do; works by children and Senior Citizens should be marked as such since they are known to differ linguistically from mainstream adult text.
Demographic coverage is important for pronunciation study but marginal in a normative reference corpus. It can be attached to E.1.1.*.3. language background.

8.4 Note on equipment

The current standard is the DAT recorder, which makes a digital audio recording which can be processed directly by computer. Both unidirectional and omnidirectional microphones should be used, and spaced to distinguish overlapping speech. Microphone arrays are expected to become widely used.

8.5 Transcription conventions

The NERC report specifies a level of transcription suitable for large-scale data entry. Speakers are minimally identified and there are simple conventions for unclear passages, hesitations and pauses, speaker overlap, laughter etc. For convenience this is copied in appendix to this paper.

8.6 Storage and retrieval

Transcriptions should normally be held as simple text files, with annotations (as above) clearly identified in ISO 8879 (SGML) compatible notation. Software that will align the digitally recorded sound wave with an orthographic transcription is available in prototype for several languages, and it is expected that it will shortly be on the market, and extended gradually to all the languages of Europe.
Databases are not recommended for the storage of language corpus data because of the heavy overhead of entering the material, and the slow performance of retrieval software.

8.7 Legal Issues

Recording speech is potentially an invasion of privacy, and countries have enacted various types of legislation to control it. Subject to the possibility of national laws overriding these remarks, surreptitious recordings should contain nothing that might identify participants (except voiceprint, which is inescapable). Volunteers should be asked to sign as agreeing to participate.

9 References

Atkins, B.T., Duval, A., Milne, R.C., Cousin, P-H., Lewis, H.M.A., Sinclair, L., Birks, R.O. & Lamy, M-N. (eds) (1987) Collins Robert French–English English–French Dictionary. 2nd edition. Glasgow: Collins.

Atkins, S., Clear, J. & Ostler, N. (1992) Corpus design criteria. *Journal of Literary and Linguistic Computing* 7(1).

Bhatia, V.K. (1993) Analysing genre: Language use in professional settings. London: Longman.

Biber, D. (1988) Variation across speech and writing. Cambridge: CUP.

Biber, D. (1989) A typology of English texts. Amsterdam: Mouton de Gruyter.

Biber, D. (1993) Representativeness in corpus design. *Journal of Literary and Linguistic Computing* 8(4).

Biber, D. & Finnegan, E. (1988) Adverbial stance types in English. *Discourse Processes* 11: 1–34.

Calzolari, N., Baker, M. & Kruyt, T. (eds) (1995) Towards a network of european reference corpora. Report of the NERC Consortium Feasibility Study, coordinated by Antonio Zampolli. *Linguistica Computazionale* XI–XII. Pisa: Giardini.

EAGLES (1996a) Preliminary recommendations on corpus typology. EAG–TCWG–CTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

EAGLES (1996b) Recommendations on corpus encoding. EAG–TCWG–CES/R-F. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

EAGLES (1996c) Preliminary recommendations on spoken texts. EAG–TCWG–SPT/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

Firth, J.R. (1957) A synopsis of linguistic theory, 1930–1955. In Palmer, F.R. (ed) (1968) Selected papers of J.R. Firth 1952–9. Harlow: Longman.

Frantzi, K.T. & Ananiadou, S. (1996) A hybrid approach to term recognition. *Proc. Int. Conf. on Natural Language Processing and Industrial Applications* (NLP+AI 96), Université de Moncton, Moncton, NB, Canada. 93–98.

Halliday, M.A.K., McIntosh, A. & Strevens, P. (1964) The linguistic sciences and language teaching. London: Longman.

Halliday, M.A.K. & Martin, J.R. (1993) Writing science, literary and discursive power. London:Falmer.

Hazadiah, M.D. (1993) Topic as a dynamic element in spoken discourse in text and technology. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds) Text and Technology. Amsterdam: John Benjamin. 55–74.

Hoey, M. (1991) Patterns of lexis in text. Oxford: OUP.

Hunston, S. (1989) Evaluation in experimental research article. PhD thesis, University of Birmingham, UK.

Joos, M. (1961) The five clocks. New York: Harcourt Brace.

Lauriston, A. (1996) Automatic term recognition: Performance of linguistic and statistical learning techniques. PhD thesis, UMIST, Manchester, UK.

Leech G., Myers, G. & Thomas, J. (1995) (eds) Spoken English on computer: Transcription, mark-up application. London: Longman.

Lyons, J. (1977) Semantics. Cambridge: CUP.

Macchi, J. (ed) (1988) The Collins Sansoni Italian dictionary. Firenze: Sansoni.

Nakamura, J. (1986) Classification of English texts by means of Hayashi's Quantification Method Type III. *Journal of Cultural and Social Science* 21: 71–86.

Nakamura, J. (1987) Notes on the use of Hayashi's Quantification Method Type III for classifying English texts. *Journal of Cultural and Social Science* 22: 127–145.

Nakamura, J. (1992) Hayashi's Quantification Method Type III: A tool for determining text typology in large corpora. An annex to a general report on annotation tools of the NERC Report. Unpublished manuscript.

Nakamura, J. (1993) Statistical methods and large corpora: A new tool for describing text types. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds) Text and Technology. Amsterdam: John Benjamin. 313–332.

Nakamura, J. & Sinclair, J. (1995) The world of woman in the Bank of English. *Journal of Literary and Linguistic Computing*: 2.

Norling-Christensen, O. (1995) Design and composition of reusable harmonised written language reference corpora for european languages. Draft report for PAROLE project.

Phillips, M.K. (1983) Lexical macrostructure in science text. PhD thesis, University of Birmingham, UK.

Renouf, A. (1993) Report on AVIATOR's clustering methodology and software. Unpublished paper.

Roget, P.M. (1962) Roget's Thesaurus of English words & phrases. London: Longmans.
Sager, J.C., Dungworth, D. & McDonald, P.F. (1980) English special languages: Principles and practice in science and technology. Wiesbaden: Brandstetter.
Sperberg-McQueen, C.M. & Burnard, L. (eds) (1993) Guidelines for electronic text encoding and interchange.
Summers, D. et al. (1993) Longman Language Activator. Harlow: Longman.
Woolf et al. (eds) (1976) Webster's New Collegiate Dictionary. Massachusetts: Merriam.
Yang, H. (1986) A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing* 1(2): 93–103.

10 Appendices

10.1 Guide to the tables

The following tables are largely taken from the NERC report (Calzolari et al., 1995) in which several corpus projects were reviewed with the aim of investigating which corpus design is the most feasible. They have been updated and other corpus projects which were not in the NERC report have been added e.g. British National Corpus, Survey of English. The tables are designed to show the common features of different classification systems in European corpora. The categories are those in the NERC report in which terminologies of different corpus projects were interpreted in order to find common features in the corpora. The categories in the tables are those categories which had common ground in many or most of the corpora listed. The tables were designed with generally applied rather than corpus-specific categories. These appendices have been compiled by J. Ball.
As in the NERC report, '+' is taken to mean "Explicitly included as a text type", '-' means "explicitly rejected as a text type" and 'A' refers to the features 'administrative data' (rather than text type). Cases of doubt are marked with '?'.

10.2 Text typology

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
LITERARY GENRE																														
poetry			-							+		-							+			-		+		-		+		
narrative				+								+	+	+		+	+							+	+	+	+	+		
(auto)biography					+							+		+		+										+		+		
novel/short story							+					+		+		+					+	+		+		+				
historical												+		+								+				+				
sciencefiction												+							+			+				-				
humour					+							+							+		+	+				+	+			
theatre/drama			+	+	-							-		+			+	+	+					+			-	+		
TOPIC																														
topic			-				+					+	+		+				+	+		+	+	+	+	+	-	+	+	
MEDIUM																												+		
books										+		+			+				+		+			+	+			+	+	+
letters/correspondence			+									+						+						+		+		+		+
newspapers	+		+	+	+		+		+	+		+		+	+		+		+		+	+		+	+	+		+	+	+
brochures/leaflets												+			+				+			+		+		+		+		
FICTION/NON-FICTION																														
fiction			+		+					+		+			+				+			+	+	+	+	+		+	+	+
non-fiction												+			+											+		+	+	+
STYLE																														
distance			+					+	+			+	+							+			+		+		+			
popular/solemn			+		+				+										+			+	+	+						
specialised/lay			+	+	+				+		+			+			+				+	+			+					
(=technical)																														
OTHERS																														
handbooks/textbooks	+		+									+		+	+						+					+		+		
translations		+																	-					+		-				

1. Bou Written + Spoken
2. Hoz unspec.
3. Svartik et al Written + Spoken
4. Juilland et al Written
5. Kucera et al Written
6. de Vriendt–de Man Spoken
7. Uit den Bogaart Written + Spoken
8. de Jong Spoken
9. Lara Written + Spoken
10. Allen Written
11. Altenberg Spoken
12. Birmingham Collection of English Texts Written
13. Birmingham Collection of English Texts Spoken
14. Gonzalez et al Written
15. Staphorsius Written
16. Feldweg Spoken
17. Morales Written
18. Collins et al Written + Spoken
19. Summers (selective component) Written
20. Crowdy Spoken
21. Bindi et al Written
22. Martin et al Written + Spoken
23. Werkgroep Taalbank Written + Spoken
24. Atkins et al Written + Spoken
25. Biber Written + Spoken
26. Malaga Written
27. Malaga Spoken
28. Bank of English Written and Spoken
29. British National Corpus
30. Survey of English

10.3 Subject typology

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Religion	+			+	+		+	+		+		+	+	+	+	+	+	+		+
Technics/-ology			+		+	+		+	+	+	+	+		+	+	+		+		
Law		+			+			+		+		+	+	+	+	+	+	+		
Sports		+			+			+		+		+	+	+	+	+	+	+		
Arts		+		+				+		+		+		+	+	+		+		+
Politics							+	+				+	+	+	+		+	+		+
History					+			+		+		+		+	+	+		+		
Medicine				+				+		+		+		+	+	+		+		
Philosophy					+			+		+		+		+	+	+				
Economy					+			+				+		+	+	+		+		+
Education					+			+		+		+		+	+			+		
Psychology					+			+		+		+		+	+			+		
Science		+	+	+			+						+	+			+			+
Sociology					+			+		+		+		+	+	+				+
Leisure				+						+		+	+	+	+	+				+
Civilisation					+			+		+		+					+	+		
Physics					+			+		+		+		+	+					
Biology					+			+				+		+	+			+		
Mathematics								+		+		+		+	+					
Household								+		+		+				+		+		
Travels					+			+				+				+		+		
Anthropology								+				+		+	+			+		
Military								+		+		+		+				+		
Media/communication								+				+		+	+			+		
Language								+				+		+	+	+				
Literature										+		+		+	+					
Architecture								+				+			+			+		
Fashion/clothes								+				+		+				+		
Computing								+				+				+		+		
Agriculture										+				+	+			+		
Geography										+					+	+		+		
Ecology/environment								+						+				+		
Traffic/transport												+		+		+				
Chemistry												+		+	+					
Finance												+		+		+				

1. Bou
2. Svartik et al.
3. Juilland et al.
4. Kucera et al.
5. Uit den Bogaart
6. Lara
7. Altenberg
8. Birmingham Collection of English Texts
9. Gonzalez et al.
10. Staphorius
11. Morales
12. Summers (selective component)
13. Crowdy
14. Bindi et al.
15. Martin et al.
16. Werkgroep Taalbank
17. Biber
18. Malaga
19. Bank of English
20. British National Corpus[a]

[a]Only non-fictional works classified into topic

10.4 Spoken typology

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Communicative Situation																
face to face	+			+	+	+			+		+		+	+		+
telephone	+			+	+				+		+		+	+		+
NUMBER OF PARTICIPANTS								A			+			+	+	
monologue	+								+			+		+	+	+
dialogue	+					+			+			+		+	+	+
GENRE MONOLOGUE																
lecture				+	+				+		+		+	+	+	
commentary	+		+		+		+		+				+	+	+	
speech				+			+		+		+			+	+	
sermon				+					+					+	+	
demonstration					+				+					+	+	
narration					+	+							+			
GENRE DIALOGUE																
conversation				+	+	+	+		+				+	+	+	+
interview		+	+		+	+			+				+	+	+	
discussion/debate				+	+				+				+	+		+
consultation									+					+	+	
proceedings									+					+	+	
chat shows									+				+		+	
meeting/gathering									+				+		+	
SETTING								A								
education				+	+				+		+		+		+	
commerce/business		+							+				+		+	
radio		+		+	+				+	+	+		+		+	
tv					+				+	+	+		+		+	
political and social organs.				+					+				+			
private/personal		+				+			+		+	+				
work											+					
institutional									+			+			+	
SPONTANEITY											+			+		
scripted								A				+		+		+
unscripted/preplanned	+							A	+			+		+		+
spontaneous/unprepared	+	+				+		A		+		+	+	+		+
TOPIC	+			+	A	+		A	+		+	+	?	+		
STYLE																
distance	+		+		+			A				+				
specialised/lay(technicality)									+			+				
PARTICIPANT DETAILS																
sex			+		?	+		+	+		?			+	+	
age			+		?	+		+			?			+	+	
ethnic group					?	+	A		A							
region			+		?	+		+	+		?			+	+	

In the above table, `A` means administrative data rather than text type.

1. Svartik et al.
2. de Vriendt de Man
3. de Jong
4. Altenberg
5. Birmingham Collection of English Texts Spoken
6. Feldweg
7. Collins et al.
8. Crowdy demographic
9. Crowdy context governed
10. Martin et al.
11. Atkins et al.
12. Biber
13. Malaga
14. Bank of English
15. British National Corpus
16. Survey of English

10.5 Regional and temporal aspects

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Regional aspects	+	+	+	+	+	+		+	+		+	+	+	+	+	+	+	+	+
Temporal aspects			+	+	+	+	+	+				+	+		+	+	+		+

	20	21	22	23	24	25	26	27	28	29
Regional aspects	+		+	+	+		+			+
Temporal aspects	+	+	+		+		+			+

1. Bou Written + Spoken
2. Hoz unspec.
3. Svartvik et al. Written and Spoken
4. Juilland et al. Written
5. Kucera et al. Written
6. de Vriendt–de Man Spoken
7. Uit den Bogaart Written and Spoken
8. de Jong Spoken
9. Lara Written and Spoken
10. Allen Written
11. Altenberg Spoken
12. Birmingham Collection of English Texts Written
13. Birmingham Collection of English Texts Spoken
14. Gonzalez et al. Written
15. Staphorsius Written
16. Feldweg Spoken
17. Morales Written
18. Collins et al. Written and Spoken
19. Summers Written
20. Crowdy Spoken
21. Bindi et al.
22. Martin et al. Written and Spoken
23. Werkgroup Taalbank Written and Spoken
24. Atkins et al. Written and Spoken
25. Biber Written and Spoken
26. Malaga Written
27. Malaga Spoken
28. Bank of English
29. British National Corpus