Expert Advisory Group on
Language Engineering Standards

EAGLES

Preliminary recommendations on Corpus Typology

EAGLES Document EAG–TCWG–CTYP/P

Version of May, 1996

Contents

1 Author

J. Sinclair
School of English
University of Birmingham
Edgbaston
Birmingham Tel.: +44.21.414.56.88
United Kingdom Fax.: +44.21.414.32.88
B15 2TT E-mail: J.Sinclair@bham.ac.uk

2 Introduction

Electronically-held corpora are new things, and little has yet emerged of a consensus as to what counts as a corpus and how corpora should be classified. The most important areas are:

(a) the minimum conditions for any collection of language to be considered as a corpus;
(b) the separation of corpora which record a language in ordinary use from corpora which record more specialised kinds of language behavior.

Both of these are contentious areas. If the profession finds that the criteria recommended here are adequate for current needs, then considerable progress will have been made, for there are many collections of language called corpora which do not meet these conditions, and there are some corpora available which record special and artificial language behavior, but do not point this out to the undoctrinated.

Furthermore, the discipline of corpus linguistics is developing rapidly and norms and assumptions are revised at frequent intervals. Categories have to be particularly flexible to meet such unstable conditions.

Hence the classifications in this paper go as far as is prudent at the present time. They offer a sound and resonably replicable way of classifying corpora, with clearly delimited categories wherever possible, and informed suggestions elsewhere. The paper has been reviewed by many experts in the field, who are in broad agreement that to present a more rigorous classification would be intellectually unsound and would be ignored by the majority of workers in the field. The present paper has a chance of acceptance because it raises the relevant issues and offers usable classifications.

For nearly twenty of those thirty years, the original targets of the Brown corpus (Kucera & Francis, 1967) were taken to be the standard:

(a) one million words

(b) divided roughly evenly into genres

(c) 500 samples

(d) 2000 words in each

(e) written published sources

This is still a much used reference point, although the circumstances that led the Brown designers to make those choices are quite unlike those of today.

It is more helpful to extrapolate from the original design the principles that lay behind the specific decisions:

(a) The corpus should be as large as could possibly be envisaged with the technology of the time. Brown's one million words was just that and its appearance was like a miracle — so many words at one's command. However, by the mid-seventies, the targets had gone up by an order of magnitude, the Birmingham Collection of English Text, for example, ending up with 20 million words in 1985. Now, in the mid-nineties, they are up another order of magnitude, with the Bank of English, for example (see below), being close to 200 million words in length.

(b) It should include samples from a broad range of material in order to attain some sort of representativeness.

(c) There should be an intermediate classification into genres between the corpus in total and the individual samples.

(d) The samples should be of an even size.

(e) The corpus as a whole should have a declared provenance.

Point (d) above — that samples should be of an even size — is controversial and will not be adopted in these proposals — see below for further consideration. The restriction of the Brown corpus to written material — still frequently copied in later work — is regarded as unfortunate for a model although understandable in its historical context. Indeed, the first alternative models to the Brown were European collections of transcribed speech, such as the Edinburgh-Birmingham corpus of the early sixties (Jones & Sinclair, 1974). For the importance of spoken corpora and their special contribution to corpus work, see below. It is noticeable that there is still considerable reluctance among corpus designers to include spoken material; at the planning stage of the Network of European Reference Corpora (NERC) in 1990 it was almost abandoned and there are signs now in the EU that spoken and written corpora may be developed separately and that the confusion between corpora for speech and corpora for language has returned.

The early corpus designers worked with slow computers (from our viewpoint) that were oriented towards numeric processing, and whose software had great difficulty with characters. Texts were assembled as large trays of cards, and retrieval programs were done on an overnight batch basis. All material had to be laboriously keyboarded on crude input devices.

In the last decade, there has been an unprecedented revolution in the availability of text in machine readable form, the emergence of a new written form — e-mail — which only exists in that form, and the invention of scanners to aid the input of certain types of text material. The processing speed of machines and the amount of storage has risen dramatically and costs have fallen as dramatically, so that modest PC users can have access to substantial corpora, while major users manipulate hundreds of millions of words on line. The balance of problems has switched from bottlenecks in acquiring corpus material to handling floods of it from a variety of unco-ordinated sources. In anticipation of this change, the notion of *monitor corpora* is under development to reconceptualise a corpus as a flow of data rather than an unchanging archive (see below section 8).

3 Definitions

3.1 *Corpus* and *computer corpus*

A *corpus* is 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.

Note that the non-committal word 'pieces' is used above, and not 'texts'. This is because of the question of sampling techniques used. If samples are to be all the same size, then they cannot all be texts. Most of them will be fragments of texts, arbitrarily detached from their contents.

A *computer corpus* is a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance.

3.1.1 Deprecated terms

Words such as *collection* and *archive* refer to sets of texts that do not need to be selected, or do not need to be ordered, or the selection and/or ordering do not need to be on linguistic criteria. They are therefore quite unlike corpora.

Citations are individual instances of words in use and collections of these also have no claims to be corpora. The precise conditions for a valid sample size for a corpus are indeed under discussion — see later — but no-one concerned seriously with corpora has attempted to gather a collection of ciations and announce it as a corpus. What has happened is that owners of previously-gathered citation collections have tried to use them as a bridge between traditional practice — particularly in lexicography — and corpus-based work.

It is unhelpful to confuse categories in this way, and important to assert minimal criteria for use of the word 'corpus'.

3.2 *Text*

A text is not defined here, because its definition will be the subject of future work. Here it is simply pointed out that the word is used for both spoken and written communications.

3.3 *Subcorpus*, *component* and *sublanguage*

A corpus can be divided into *subcorpora*. A *subcorpus* has all the properties of a corpus but happens to be part of a larger corpus. Corpora and subcorpora are divided into *components*. A *component* is not necessarily an adequate sample of a language and in that way it is distinct from a corpus and a subcorpus. It is a collection of pieces of language that are selected and ordered according to a set of linguistic criteria that serve to characterise its linguistic homogeneity. Whereas a corpus may illustrate heterogeneity, and also a subcorpus to some extent, the component illustrates a particular type of language. What are called *sublanguages* are components in this definition, but there are other restrictions on sublanguages which will be dealt with later.

3.4 Linguistic classification criteria

The following criteria for classification can be applied to corpora, subcorpora and components. Linguistic criteria may be:

External in that they concern the participants, the occasion, the social setting, the communicative function of the pieces of language, etc. These are the familiar categories of Kucera & Francis (1967), Hofland & Johansson (1982), Sinclair (1988), Atkins, Clear & Ostler (1992) and NERC (1994).

Internal in that they concern the recurrence of language patterns within the pieces of language. These are newer and of growing interest — see Biber (1988) and Nakamura (1993).

External criteria are largely mapped onto corpora from text typology, which is not dealt with fully here, being the subject of subsequent work. Internal criteria are the subject of later work also.

3.5 Characteristics

A corpus is assumed to have certain characteristics attached, with *default* values. Unless stated, these characteristics are attributed to anything called a corpus. A corpus which has one or more non-default values for these characteristics is termed a *special corpus*: its title should specify its deviations from the assumptions. See section 3.5.2 for further treatment of special corpora.

The assumed characteristics, with their defaults, are discussed in the following.

3.5.1 Quantity

The default value of **Quantity** is **large**. A corpus is assumed to contain a large number of words. The whole point of assembling a corpus is to gather data in quantity. The size of corpora continues to increase rapidly, and it would not be sensible to recommend any set of figures. Furthermore, the advent of *monitor corpora* (see below) changes the basis of size calculation from a total amount to a rate of flow. The size of a corpus is simply the sum of the sizes of its components. Questions of size are best dealt with by reference to a component.

In practice, the size of a component tends to reflect the ease or difficulty of acquiring the material. In turn, this factor may be loosely related to the availability of the material to the public and therefore to its relative importance as influential language, as against material which is difficult to get, perhaps because it is of small circulation. Such a relationship, however, does not hold with reference to the spoken language, where the most influential and pervasive material is informal and impromptu conversation, which is not normally recorded.

A more practical correlation to pursue is that between the size of a component and the number of people who are exposed to it. Since millions of people read the newspapers and listen to the radio, it should be easy to acquire large quantities of this type of data, and this will be assumed. Local radio reaches fewer, but still hundreds of thousands, and so do magazines and best-selling paperback books. Speeches to large rallies reach thousands of people, as do leaflets and notices.

The figures are merely typical. There are some leaflets printed in millions by governments and advertisers, and some very modest magazine circulations and audiences to radio. But note that if a speech at a rally is repeated on television, or a magazine article is reprinted in the newspapers, its intended audience is still the original one — what happens to it afterwards does not affect its linguistic constitution.

It could be argued, however, that its very linguistic constitution is the factor that caused it to be transferred to another medium; or that the speaker/writer was attempting to achieve the wider publicity and contrived his/her language accordingly, and the text is thus more appropriately classified at its final destination. In practice, it frequently happens that a text is transferred from one medium to another and such a text may well be in a corpus twice.

Lower circulation material, involving hundreds of people, is to be found in workplace and institutional documentation, and in the spoken medium in lectures, talks and some sermons. When the audience can be numbered in tens we are down to documents with a circulation list and seminars. Individuals can be identified and in some cases the role of speaker/writer can change. At these levels, written material is not easy to identify, but there is of course private correspondence, with a readership of one or two only. Here there is a lot of variety in the spoken medium, with all kinds of discussions, interviews, meetings and conversations involving very small numbers of participants. E-mail tends to be used in fairly small groups but there is more and more circular material coming out.

Two points emerge from this discussion. One is that forms of the spoken language are composed, given to mass audiences, and made available in electronic form. This contrasts with the view, often expressed in corpus circles, that spoken language is difficult and expensive to obtain. This point will be elaborated later.

The other point is that a classification of texts by approximate audience size is worth further consideration as a way of quantifying the size default. It is open to various criticisms, particularly that it is just making a virtue out of necessity. Other, more suitable but equally realistic measures are solicited so that there may be general guidance to the very first question asked by a novice in corpus linguistics — How big a corpus do I need?

3.5.2 Quality

The default value for **Quality** is **authentic**. All the material is gathered from the genuine communications of people going about their normal business. Anything which involves the linguist beyond the minimum disruption required to acquire the data is reason for declaring a *special corpus*. Such a declaration protects the interest of those who wish to make statements about the way the language is used in ordinary communication, and who might be misled into including data which had arisen in experimental conditions, or in artificial circumstances of various kinds.

It is difficult to draw the line. For example, some television shows deliberately put participants into artificial and indeed bizarre conditions and induce extremely odd responses. Casual conversation is expected to be impromptu but it can be rehearsed by one or more parties.

However difficult round the margins, it is important that serious intervention by the linguist, or the creation of special scenarios, is recorded in the name of the corpus. **Experimental corpus** may be suggested as a general category.

One well-recognised type of experimental corpus is the **speech corpus**, which is assembled for the study of fine details of the spoken language. Such a corpus may be very small and be the product of asking subjects to read out strange messages in anechoic chambers. The classification of speech corpora is not the concern of this document. (For more on speech corpora, refer to the reports of the EAGLES Spoken Language Working Group.)

A special category is the **literary corpus**, of which there are many kinds. Biblical and literary scholarship began the discipline of corpus linguistics long ago, and there is a lot of expertise available in literary circles on such things as establishing a canon of an author's works. Classification criteria include, as well as the **author**, the **genre** (odes, short stories, etc.), the **period**, the **group** (Augustan poets, campus novelists, etc.), the **theme** (revolutionary writings, etc.). **Drama** is usually kept separate from prose and poetry.

Corpora of the language of children, geriatrics, non-native speakers, users of extreme dialects and very specialised areas of communication (like the heraldic blazon or the knitting pattern, or the auctioneer's patter) should also be designated special corpora because of the unrepresentative nature of the language involved.

Note that the special corpus is different in principle from a corpus that features one or other variety of normal, authentic language. A corpus of conversations is not a special corpus, nor is a corpus of newspaper text, or even one particular newspaper. There is a distinction made here between variety within the limits of reasonable expectation of the kind of language in daily use by substantial numbers of native speakers, and varieties which for one reason or another deviate from the central core. The special corpora are those which do not contribute to a description of the ordinary language, either because they contain a high proportion of unusual features, or their origins are not reliable as records of people behaving normally.

Each component, then, illustrates a particular kind of language, and for each component there should be a descriptive label that indentifies the homogeneity of the material inside. The particularity of the language may be retained at corpus or subcorpus level without transferring the corpus into the 'special' category.

Hence it is proposed that special corpora are designated as follows, e.g.

Special corpus: Poetry of Aphasics

A corpus which illustrates a particular variety of normal language is designated, e.g.

Corpus of The Times newspaper, 1991

3.5.3 Simplicity

The default value of **Simplicity** is **plain text**. This means that the user can expect an unbroken string of ASCII characters, with any mark-up clearly identified, and separable from the text. Nowadays it is likely that many texts will be in SGML format, and in the future perhaps TEI. These mark-ups have been carefully designed and do not impose any additional linguistic information on the text. Largely, their role in relation to text representation is to preserve in linear coding some features which would otherwise be lost. They are perceived as helpful but their presence must be recorded, and the original text must be easily retrievable.

The same conventions for mark-up are extendable to various annotations that add information provided by expert linguistic analysis. Such information is the organisation and interpretation of textual features, and it varies from analyst to analyst and from purpose to purpose. Other sections of this report deal with types of analytic annotation. A plain text policy is not opposed to such annotation, nor to the use of the same mark-up conventions. For clarity in the future it would be helpful to distinguish between added codes which encode only surface features of texts that would otherwise be lost in transfer to a machine, and added codes which encode analyses and interpretations. This distinction would have to be made carefully for spoken transcription, because it can be claimed that the orthographic transcriber adds analytic notation of a kind, but one which is so conventional and familiar that people can treat it as a sophisticated mark-up, and quite distinct from intonation annotations or grammatical tags (see 4 below).

More difficult is the question of *annotated corpora*. It is proposed that this term is used for any corpus which includes codes that record extra information — provenance, analytical marks, etc. Again the annotations should be separable from the plain text in a simple and agreed fashion. A set of conventions for removing, restoring and manipulating annotations is necessary, especially as the next few years will see a large growth in the provision of annotated corpora. It is naive to expect that big corpora will remain easy to manage if they are full of various annotations; retrieval times are already critical.

3.5.4 Documented

The default value is **documented**. This means that, as proposed in NERC (1994), full details about the constituents of a component are kept separately from the component itself. The model for this is the DTD or *header* of SGML, and, following that, TEI. In contrast to the recommendations of those bodies, corpus users seem to prefer to keep the documentation of texts in a separate place from the texts themselves, and to include only a minimal header that contains a reference to the documentation. For the management of corpora this practice allows the effective separation of plain text from annotation with only a small amount of programming effort; since DTDs can be extremely verbose, the efficiency of real time search procedures is hampered if they are not detachable.

4 Spoken corpus

There is considerable confusion over the use of this term, and it would be helpful to achieve a consensus. First, it is to be distinguished from *speech corpus*, a term for a special corpus described above (3.5.2). Then there is a choice. In the usage of some authorities (e.g. Chafe (1995)), it means a corpus of informal, impromptu conversation, with no media involvement. On the other hand, it is used by some to mean any language whose original presentation is in oral form — that is, the speakers involved behave in oral mode. If such a text is later presented in written form, without change except for the transcription, it should be classified as spoken — a BBC Reith Lecture, for example. If, in time, a spoken corpus can be stored as sound wave as well as transcript, then such a text may exist in two versions and a special kind of parallel corpus can be introduced.

Similarly, any text composed to be presented in written form can be read out, but its expression need only change in ways required by the change of medium. It is, therefore, primarily a written text.

Our preference is for the latter interpretation of 'spoken corpus'. There are doubtful areas whichever meaning is chosen — how impromptu is impromptu, how informal is informal, etc. How does one know whether a composer intends a text to be written or spoken, or both? But to reserve the term for only one small class of spoken language texts seems to distort the meaning of the words involved.

The problem is that informal, impromptu speech is regarded by many scholars as the most important variety of all, closest to the core of language, revealing the characteristic patterns of a language in a way that no other variety does. It is also the most difficult and expensive to acquire, difficult to classify and manage. The crudities of transcription make a spoken corpus unsatisfactory as held in most centres, and there is no consensus as yet about the conventions of transcription. The nearest we have is the recommendations of NERC (NERC, 1994).

5 Samples

It has become traditional to use a sampling technique in the assembly of corpora — typically following the Brown model summarised above. Samples are small, in relation to texts such as newspapers, books and radio programmmes, and of a constant size, hence not qualifying as texts, as has been pointed out above. A distinction is proposed here between a *text corpus* or a *whole text corpus* and a *samples corpus*. We would like to see 'whole text' as a default condition, thus classifying samples corpora as one of the categories of special corpora, but there are still many corpora in use made up of small samples. It should, however, be realised that this feature is just a remnant of the early restraints on corpus building and it confers no beneift on the corpus. The use of samples of a constant size gains only a spurious air of scientific method.

6 Sublanguages

Sublanguage is an important concept in natural language processing. It is assumed that by narrowing the subvariety, usually in a specialised communicative context, the actual structure of the language will simplify, and thus become more amenable to automatic processing. The vocabulary, too, is restricted and often specialised. There are corresponding constraints at semantic, conceptual and pragmatic levels (Harris, 1988).

A sublanguage is thus defined simultaneously by internal and external criteria, but the internal criteria are crucial. It remains to be seen whether the external and internal criteria actually match in practice. The study of genre, and LSP (languages for special purposes) shows that writers conform to quite elaborate prescriptions when composing in a technical or professional context, so it is not surprising to find many similarities.

A sublanguage is at the other end of the linguistic spectrum from a reference corpus (see 7). The homogeneity of its structure and specialised lexicon makes it useful for NLP purposes and allows the quantity of data to be kept small, i.e. it demonstrates, typically, good closure properties. The concept of sublanguage is to be distinguished from those of artificial language or reduced language. The latter are deliberately created, whereas sublanguages evolve naturally (although at the level of terminology there may be deliberate acts of creation).

Increasingly, the natural language processing community is finding that it needs access to corpora containing sublanguage material, in order to build systems capable of handling specialised texts (McNaught, 1993). Under our previous definition, corpora consisting of sublanguage material are *special corpora.*

7 Reference corpora

A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials. The model for selection usually defines a number of parameters that provide for the inclusion of as many sociolinguistic variables as possible and prescribes the proportions of each text type that are selected. A large reference corpus may have a hierarchically ordered structure of components and subcorpora.

Questions of balance and representativeness recur in the discussion of reference corpora. They are extremely difficult to define, and yet fairly easy to work with. While it is not normally claimed that there is a core variety of a language, there appear to be a large number of heavily overlapping varieties, sharing the bulk of their vocabularies and almost all the syntactic rules. Marginal vocabulary items differentiate them and slight individuality of phraseology. Some general features, associated with such things as formality, speech, preparedness and broad subject-matter, group them in people's minds, and a rough kind of representativeness is achieved by ensuring that a large quantity of text exemplifies each of these parameters.

Special corpora are made up of texts that do not overlap as much with the large central pool. To be clearly 'in a language' they must show quite a number of the grammatical and lexical features of that language, but the markedness of the patterns unique to them serves to differentiate them clearly from the general varieties of the language. In due course, and with the growing influence of internal criteria, reference corpora will be used in order to measure the deviance of special corpora.

Reference corpora are at the heart of the future development of corpus-based work in Europe and elsewhere. Reference corpora in several languages, constructed on similar principles, form a group of *comparable corpora* (see 10 below).

Example: **The Bank of English**

The Bank of English is a reference corpus. From the total holdings in Birmingham, a corpus is identified from time to time and made available world-wide via the Internet, with appropriate software. At present the corpus contains around 167 million words, soon to top 200 million. This is divided into several subcorpora:

Newspapers	43 million words
Books	37
Magazines	38
Radio	39
Ephemera	1.5
Informal Spoken	8.5

The first four subcorpora are samples from fairly plentiful material (though magazines have only recently become available). They are kept roughly comparable in size, centring on 40 million words.

There is a lot more in the vaults, so to speak — perhaps 150 million words of newspaper text alone — but only a proportion is on-line.

The Ephemera subcorpus has to be rekeyed from a wide variety of pamphlets and brochures, and is laborious and expensive to acquire. The Informal Spoken corpus, though dwarfed in size by the others, is possibly the largest available of its kind.

In turn these are broken down into further subcorpora, e.g. for Newspapers, UK is a subcorpus; and then into components such as The Times (10m), The Guardian (12m), etc. The Radio subcorpus is divided into BBC (18m), and American NPR (21m), and each of these is broken down into components by date. The Magazines subcorpus picks out The Economist (7m) and the New Scientist (3m), and puts the rest together as a general subcorpus.

The retrieval software allows the user to consult the whole corpus or any grouping of its components, either temporarily or as the default for that user. There is no restriction on the combinations of characters and codes that can be selected, with gaps, wildcards and varying sequence. Co-occurrences are particularly easy to examine, either within a putative sublanguage or in the full range of the corpus, for building translation tools or lexicons.

8 Monitor corpora

The understanding of large corpora has increased with experience of handling them and the development of electronic applications to publishing has made data available in very large quantities. It became clear some years ago that the assumption of a finite limit on a corpus for any length of time was an unnecessary restriction. For some uses, it is essential to achieve a steady corpus size and constitution, but this is easy to devise within a large and constantly moving collection. The question that arose was how to manage the large quantities of data that were foreseen for what is known as a *monitor corpus*.

The first model was of a corpus of a constant size, so that the software of the day could cope with it, which would be constantly refreshed with new material, while equivalent quantities of old material would be removed to archival storage. The constitution of the corpus would also remain parallel to its previous states.

This model gave rise to the idea of *rate of flow* as the best way of managing the corpus. Instead of setting, say, 10 million words as the proper proportion of that genre, the setting could just as easily be 10 million words a year. Or a month, or a week. The language would flow through the machine, so that at any one time there would be a good sample available, comparable to its previous and future states.

Such a model opened up new prospects for those interested in natural language processing, and it added another dimension to contemporary corpora — the diachronic. New words could be identified, and movements in usage could be tracked, perhaps leading to changes in meaning. Long term norms of frequency distribution could be established, and a wide range of other types of information could be derived from such a corpus.

Some scholars were less than happy about disposing of the older texts as new ones came in. That problem, however, was neutralised by the fast-expanding power and memories of the machines. There is no need to move any text out of modern systems. However, to manage a monitor corpus to the best advantage, it is convenient to divide it into batches of a similar size and constitution.

Over time the balance of components of a monitor corpus will change. New sources of data will become available and new procedures will enable scarce material to become plentiful. The rate of flow will be adjusted from time to time.

9 Parallel corpora

A *parallel corpus* is a collection of texts, each of which is translated into one or more other languages than the original. The simplest case is where two languages only are involved: one of the corpora is an exact translation of the other. Some parallel corpora, however, exist in several languages. Also, the direction of the translation need not be constant, so that some texts in a parallel corpus may have been translated from language A to language B and others the other way around. The direction of the translation may not even be known.

Parallel corpora are objects of interest at present because of the opportunity offered to align original and translation and gain insights into the nature of translation. From this work it is hoped that tools to aid translation will be devised. Probabilistic machine translation systems can moreover be trained on such corpora. Parallel corpora are made in the business of communication in multilingual societies, such as the United Nations, Nato, the EU and officially bilingual countries such as Canada.

10 Comparable corpora

A comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora. One of the clearest is ICE — the International Corpus of English (Greenbaum, 1991). Corpora of around one million words in each of many varieties of English around the world are being assembled following the same model, which prescribes genres and the target quantity of words to be gathered in each. Originally, the corpora were all to be gathered in the same year.

The possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus.

Note: **'Multilingual corpora'**
At present there are no multilingual corpora apart from parallel and comparable corpora; there are plenty of centres that have collected text material in several languages, and some of these collections are corpora in their own right. But unless the collections share common features of selection, at least at the level of the comparable corpus, then they are just text resources in different languages. It therefore seems unhelpful to use the term 'multilingual corpus'.

References

Atkins, S., Clear, J. & Ostler, N. (1992) Corpus design criteria. *Literary and Linguistic Computing* 7:1–16.

Biber, D. (1988) Cambridge University Press. Cambridge.

Chafe, W. (1995) Conference contribution. In: Leech, G. & Myers, G. (eds.) *Spoken English on the computer*. London: Longman.

Greenbaum, S. (1991) The development of the International Corpus of English. In: Aijmer, K. & Altenberg, B. (eds.) *English corpus linguistics*. London: Longman.

Harris, Z. (1988) *Language and information*. New York: Columbia University Press.

Hofland, K. & Johansson, S. (1982) *Word frequencies in British and American English.*. London: Longman.

Jones, S. & Sinclair, J. (1974) English lexical collocations. Cahiers de Lexicologie. Institut des professeurs de francais a l'etranger.

Kucera, H. & Francis, W. (1967) *Computational analysis of present day American English*. Providence: Brown University Press.

M[c]Naught, J. (1993) User needs for textual corpora in natural language processing. *Literary and Linguistic Computing* 8(4):227–234.

NERC (1994) Feasibility of a network of European reference corpora. Report submitted to the CEC. Pisa. Forthcoming as a book.

Nakamura, J. (1993) Text types. In: Baker, M., Francis, G. & Bonelli, E. T. (eds.) *Text and Technology*. Amsterdam: John Benjamins.

Sinclair, J., ed. (1988) *Looking Up*. London: HarperCollins.