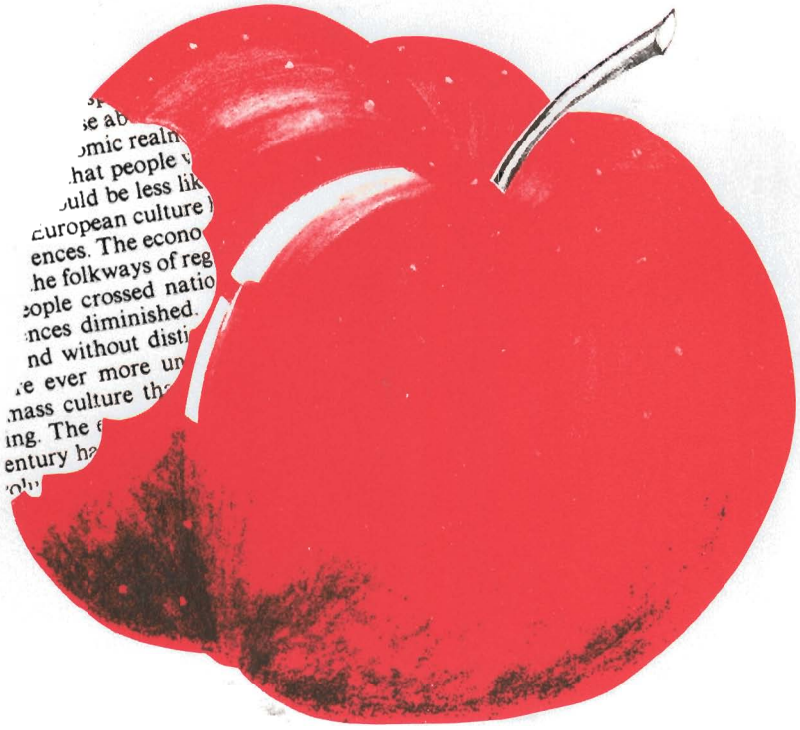
ICAME Journal

International Computer Archive of Modern English

No. 12
April 1988



NORWEGIAN COMPUTING CENTRE
FOR THE HUMANITIES

CONTENTS

The *ICAME Journal* is the continuation of *ICAME News*.
Editor: Stig Johansson, Department of English, University of Oslo

A corpus for studying linguistic variation

Nelleke Oostdijk
University of Nijmegen

1. Abstract

Corpus-based studies of linguistic variation have so far failed to make a substantial contribution to the development of a descriptive theory of linguistic variation. This failure must be attributed in part to the lack of proper data and in part to the lack of a proper methodology. Recent developments in the field of corpus linguistics, however, have provided the means to get access to data that before could only be obtained on a very small scale or not at all, while at the same time various techniques for the manipulation and interpretation of data have been discovered and/or further developed (in this light the work of Biber & Finegan (1986) must be considered valuable). It is therefore assumed that future corpus-based research potentially does have a substantial contribution to make with respect to the development of a variety theory. Crucially important is the way in which future research is set up and carried out, starting with the collection of proper initial data, including a computer-readable corpus for studying linguistic variation. In the present paper[1] the design of such a corpus is discussed.

2. Introduction

Linguists have long been aware of the fact that a language is not a homogeneous phenomenon but rather a complex of many different varieties. Over the years, however, linguistic variation has remained a problem area in linguistics and linguists in their descriptions have tended to abstract away from this variability. Studies of linguistic variation have been restricted to the setting up of small-scale hypothetical models which have contributed fairly few insights into the phenomenon of language variation.

In part, the apparent failure to cope with the complexity of such a phenomenon as language variation can be attributed to the fact that linguists have not been very well equipped to carry out large-scale formal empirical analyses which would enable them to systematically vary extra-linguistic factors and examine the accompanying linguistic variation. In the past numerous variety studies have been undertaken based, of necessity, on very small subsets of the language since they involved a manual processing of the data. With the advancement of computer technology and the development of corpus linguistics it would appear that the study

of vast amounts of data had become manageable. So far, however, corpus linguistics – while obviously well-equipped for such an undertaking – has failed to play any substantial role in the study of linguistic variation. This failure is to a large extent due to the following factors:

1. the corpora that were used were not suited for studying linguistic variation;
2. computer methods had not yet reached a stage that made it possible to retrieve sufficient information relevant to the study of linguistic variation.

The major computer-readable corpora had been compiled with the intention of representing a cross-section of either American or British English (the Brown Corpus and the LOB Corpus, respectively). Random samples of 2,000 words each were selected from a wide range of texts. In selecting a great many different samples it was attempted, on an intuitive basis, to ensure representativity and neutralize any variety-specificity. Other, 'specialized', corpora in contrast tend to be restricted to very small subsets of the language, so that they can only be used in a variety study when used in combination, each corpus exemplifying a particular subset.

So far corpus-based variety studies could only make use of quantitative data available from a word-based analysis of the corpus, such as the frequency of occurrence of words or the mean sentence-length. The study of linguistic variation can, however, only be seriously undertaken if ample quantitative data are also available about the frequency of occurrence of syntactic structures. A further handicap earlier studies were confronted with was that the statistical techniques were by no means as sophisticated as need be for such complex data. It is only through recent developments in the field of corpus linguistics, such as the implementation of systems for the automatic syntactic analysis of text corpora and the development of sophisticated quantitative techniques, that the study of linguistic variation can now be considered a feasible proposition.

3. Designing a variety corpus

3.1 Objectives

In our view a descriptive theory of linguistic variation should provide answers to at least the following three questions:

1. Under what extra-linguistic conditions is a particular variety used?
2. By what features is it characterized linguistically?
3. How do we describe the correspondences between linguistic and extra-linguistic categories?

Since it is, as Ellis & Ure (1969) point out, typical features rather than unique ones that are the subject of a variety study, a certain amount of prejudging is inevitable as part of the preliminaries to any variety study, in that texts will usually need to be grouped together (using an intuitive judgement of linguistic/extra-linguistic correspondences) to obtain a large enough initial corpus for study. Previous research may be surveyed to facilitate this prejudging. It is clear, however, that although it may be wise to survey previous research in order to establish what linguistic and extra-linguistic features should be included in the next study, it also entails the danger not only of incorporating inadequately defined text categories (while others may be overlooked), but also of restricting the scope of the research to those linguistic features that have previously been identified as possibly important without exploring the newly accessible data for other such features.

At this point it is evident that care must be taken both in the collection of data and the selection of features. The first presupposes the careful description of text categories in terms of their extra-linguistic variables (as Gregory, 1967, suggested), and the compilation of a corpus for the specific purpose of variety study. Such a corpus should meet at least the following criteria:

- the corpus should comprise a variety of samples which differ with respect to a number of extra-linguistic variables;
- the individual samples of the corpus should be large enough to be representative of a particular variety; for the study of linguistic variation rather long samples are required. Experiences with a small subset of English have led us to believe that samples of 20,000 words each are sufficiently large in order to yield reliable information about the frequency of occurrence of most syntactic structures.
- the corpus should allow for the systematic varying of extra-linguistic variables;
- the complete corpus must be large enough to make it possible to obtain ample quantitative data about the frequency of occurrence of various linguistic features as they occur in the different samples and allow for comparisons between samples.

Given such a corpus it will be possible to contrast texts or groups of texts and allocate those linguistic features that are characteristic of the texts that are being investigated to the corresponding extra-linguistic determinants. Co-occurrence patterns among linguistic features may be identified using sophisticated quantitative techniques such as the Multi-Feature/Multi-Dimensional (MF/MD) approach that Biber & Finegan (1986) propose. This approach will also make it possible to establish textual dimensions and to identify text types.

Little can be said about the total size of the corpus that will be needed for the study of linguistic variation, if one is to obtain statistically significant results. Much

will depend on the number of variables involved which relates directly to the number of samples minimally required. Given the complexity of the matter and the fact that any research is bound by limited resources we would do well to proceed by studies on well-defined subsets of the language in individual research projects. One such project is currently being carried out at the University of Nijmegen, where we are investigating a corpus of present-day British English.

3.2 Selectional criteria

At Nijmegen University we took it as our objective to compile a variety corpus that would be representative of a well-defined subset of the English language. It soon became apparent that a 'well-defined subset' can only be established when reference is made to particularly clear external criteria. Thus we found that many criteria that had been provided in the literature and also criteria that were handed to us by experiences in other research projects,[2] could not be employed since they lacked a proper definition, were too coarse or too refined. It was decided to restrict the corpus to a subset that could be described as 'intended written to be read', printed, educated, contemporary British English prose. Texts should be original British English publications, ie translations were not to be included. Likewise, texts that were of American English, Australian English, or Indian English, origin were to be left out. By restricting the subset to prose we intended to exclude poetry. Similarly, the restriction to texts that were 'intended written to be read' led to the exclusion of plays, speeches, songs, etc, in other words, texts that are primarily meant to be spoken, recited, or sung rather than read. Since we intended to exclude private material such as personal letters, and also material that had only a limited distribution such as memos, we introduced the label 'printed' which would ensure that the texts were both intended for and available to a wider public. Only texts in educated British English were to be selected, thus excluding non-educated and/or substandard English. Finally, restricting the subset to contemporary British English, although apparently meant to refrain from including archaic material, contemporary was to some extent rather arbitrarily taken to be post 1975.[3] This was done for various reasons. Given the fact that the existing corpora were not suited for our objectives, we were forced to undertake the compilation of a new corpus. Since the research we have embarked upon will proceed for several years, we wanted our material to be as recent as possible. This, as we found, proved to be a more general desire. The main British English corpus, the LOB Corpus, takes 1961 as its sampling year and is becoming rather dated. Another, minor[4] reason was that we wanted to avoid as much as possible any unpredictable influences of private time (the age of the author) on public time (the year of publication). Allowing for a briefer time-span would probably neutralize any differences in language use that might occur between the language use of, say, a 70-year-old in 1950 and that of a 20-year-old in 1987.

Once the subset had thus been established a number of initial text categories were defined. As we aimed at an acceptable minimum[5] of samples to represent each text category we could only have a limited number of categories. This resulted in some instances in rather coarse categories such as 'religion and mythology' (which makes up one category),[6] whereas other text categories that might well have been included were left out. Text categories that were included are:

Non-fiction

I Arts

NAUT	autobiography/biography
NEDU	education
NHIS	history
NLIN	language and linguistics
NLIT	literary criticism
NPHI	philosophy
NPSY	psychology and psychiatry
NSOC	sociology and anthropology
NWOM	women's studies

II Sciences

NBIO	biology
NCHE	chemistry
NECO	economics
NGEO	geography
NMED	health and medicine
NPHY	physics

III Miscellaneous

NGEN	non-fiction, general
NLAW	law and government
NMYS	mysticism and the occult
NPOL	politics
NREL	religion and mythology
NTRA	travel

Fiction

FCRI	crime and mystery
FHOR	horror
FHUM	humour
FNOV	general fiction, novel
FPSY	psychological novel
FROM	love and romance
FSFF	science fiction and fantasy
FSTO	general fiction, short story
FTHR	thriller and adventure

3.3 Sampling principles

Having established the text categories that were to be included in the corpus we then had to decide on the procedure that was to be followed in the sampling of the material. The following questions were raised:

1. What is to be the total size of the corpus?
2. What size should the individual samples be?
3. How many samples are required to represent a particular text category?
4. What samples should be selected within a certain text category?
5. What sampling procedure is to be followed within a selected source text?

As far as the total size of the corpus was concerned it was decided that the corpus should comprise at least one million words. A much larger corpus would hardly be feasible because of the time and money that would be required, while a smaller corpus was undesirable since it would force the subset to be further narrowed down if we were to maintain a certain minimum representation of samples per text category. Moreover, experiences with the 130,000 word Nijmegen Corpus had already demonstrated that such a corpus was far too small. With one million words the corpus would be comparable in size to the other major corpora, the American Brown Corpus and the British LOB Corpus.

In our view the size of the individual samples in the corpus had to be fairly large since the corpus was to be employed in the study of linguistic variation. To have samples of 2,000 or 5,000 words, as in other major corpora, would hardly yield representative samples of linguistic variation; rather, it would be more likely to reveal chance differences while neutralizing other, more significant differences. From experiences with the 130,000 word Nijmegen Corpus we knew that samples of 20,000 words each are sufficiently large in order to yield reliable information

about the frequency of occurrence of most syntactic structures. A sample size of 20,000 words would yield samples that are large enough to be representative of a particular variety.[7]

As with the total size of the corpus and the number of text categories included, the number of samples per text category should ideally be as large as possible. However, simple arithmetic shows that a corpus of one million words with samples of 20,000 words each allows for 50 samples to be selected. Given the text categories we wanted to be represented in the corpus the representation per text category was going to be rather low which led to the decision to fix the number of words for the total corpus at one and a half million. The distribution of the 75 samples over the text categories selected was mainly based on intuitive judgement. Generally we would aim at a representation of a particular text category by at least two samples. For some categories, however, we chose to select a large number of samples since these were categories that we felt were rather broad, eg 'general fiction', or 'biology'. A few of the miscellaneous categories – because they merely had an exemplifying function – were only represented by a single sample (eg 'mysticism and the occult', 'religion and mythology'). The distribution of the samples over the text categories then looks as follows:

Non-fiction (45)

I Arts (20)

NAUT	autobiography/biography	4
NEDU	education	2
NHIS	history	2
NLIN	language and linguistics	2
NLIT	literary criticism	2
NPHI	philosophy	2
NPSY	psychology and psychiatry	2
NSOC	sociology and anthropology	2
NWOM	women's studies	2

II Sciences (16)

NBIO	biology	4
NCHE	chemistry	2
NECO	economics	2

NGEO	geography	2
NMED	health and medicine	3
NPHY	physics	3

III Miscellaneous (9)

NGEN	non-fiction, general	1
NLAW	law and government	2
NMYS	mysticism and the occult	1
NPOL	politics	2
NREL	religion and mythology	1
NTRA	travel	2

Fiction (30)

FCRI	crime and mystery	4
FHOR	horror	2
FHUM	humour	3
FNOV	general fiction, novel	7
FPSY	psychological novel	2
FROM	love and romance	3
FSFF	science fiction and fantasy	3
FSTO	general fiction, short story	4
FTHR	thriller and adventure	2

The question what samples to select within a certain text category was rather central in the sampling procedure. As we observed above (section **3.1**), the corpus should make it possible to contrast texts or groups of texts and allocate those linguistic features that are characteristic of the texts that are being investigated to the corresponding extra-linguistic determinants. Therefore, samples could not be selected randomly. The selection of samples was obviously to be determined by a closed set of extra-linguistic variables. A number of the extra-linguistic variables that appeared in earlier studies had been accounted for in the definition of the subset. For example, taking the model suggested by Gregory (1967) such categories as temporal dialect, geographical dialect, social dialect, and mode of discourse were reflected in the labels 'contemporary', 'British English', 'educated', and 'written to be read', respectively. Another extra-linguistic variable, the field of discourse, could be found in the distinction of text categories. One major extra-linguistic variable,

however, had so far not been included: idiolect. Within this variable we distinguished between the author's identity, sex, age, and origin. The author's origin was restricted to Britain in order to help to determine whether the language in a sample text could be looked upon as British (our first indicator of Britishness was original publication in Britain). The author's sex and age, although recorded in the documentation, were considered minor variables and therefore not used as sampling criteria.[8] The author's identity, on the other hand, was considered of major importance, as it was expected that this might account for a linguistic variablility that could not be attributed to any other extra-linguistic variables. In the selection of samples the author's identity was generally taken to be a free variable, ie it was not used to restrict the selection of samples. It was used, however, in some instances, to make it possible to investigate such issues as to what extent the personal style of an author affects the typicalness of a certain text category, or the (in)variability of personal style. Other extra-linguistic variables were text length and distribution, both of which were determined by a practical motivation: since we wanted to be able to keep a record of and control the extra-linguistic variables involved as much as possible, we restricted the selection to samples from books that were written by a single author.[9] Newspapers, magazines, essays and articles were thus excluded. The distribution was to be international.

Having decided how many samples to select and from what texts one question remained: what sampling procedure should be followed with a selected source text? There was no scientific method that we knew of by which to select 20,000 words of running text. Rather than sampling from several instances in one particular source text we opted for a more or less random selection of the 20,000 words by taking them from roughly the middle of the book.[10] Thus a sample was begun

- if a text had chapters or similar divisions, with the heading;
- if a text was lacking chapter or similar divisions, with the first utterance following a blank line or similar spacing (as eg found after a diagram);
- if none of the above could be applied to a text, with the first utterance starting a new paragraph on the page selected for sampling.

A sample ends with the utterance containing the 20,000th word. Headings are included in the text. Extra-textual material in the source (such as diagrams, maps, lists, bibliographies) is excluded, as are footnotes and references. Similarly long foreign quotations are excluded as well as all poetry.[11]

3.4 The processing of the material

The samples of text that were selected were keyed onto tape. In order to be able to retrace what the original text had looked like it was decided to have some form of coding of the text. For this purpose the coding that had been used in the processing of the LOB Corpus was adapted.[12] For the typing student assistants were employed.

These would type the text together with the codes required. Afterwards the text would be proofread and corrected, and a count would be made in order to yield a 20,000 word sample.[13] When this had been completed each sample of text was prefixed with a tag with the format

(XXXX TEXT nn)

where XXXX stands for a text classification code (eg NAUT) and nn for the rank number of the text in a particular text category. The XXXX code as well as the rank number of the text form the first part of the location code which precedes each line of the corpus text. A full location code occupies 13 positions: apart from the 6 positions taken up by the text classification and the rank number, positions 7-10 are used to indicate the number of the page the text was on, and finally, positions 11-13 indicate the line number. Each sample of text is ended with an end-of-subcorpus tag *#.

All information about a sample and its source text is contained in the manual that accompanies the corpus, including information on corrections or deviations from the original text.

Notes

1. This paper is the result of research that was supported by the Foundation for Linguistic Research, which is funded by the Netherlands foundation for the advancement of pure research, ZWO.
2. See also Oostdijk (to appear – 1).
3. Actually the average year of publication of the texts that were selected turned out to be 1982.
4. We call this criterion 'minor' because so far there is no evidence that private time is a relevant variable in a study of linguistic variation.
5. See below, section **3.3**. The number of text categories was restricted for practical reasons. From a methodological point of view, however, it would be desirable to have not only a stronger representation of samples per text category, but also to distinguish between text categories that are more refined.
6. The same goes for categories like 'psychology and psychiatry', 'sociology and anthropology', 'law and government', but also for categories such as 'biology', 'chemistry', 'health and medicine', and 'physics', where the labels define seemingly rather precise fields of discourse that cover, however, a large range of topics.
7. See, for example, de Haan (1984) and de Haan & v. Hout (1986).

8. However, only adult authors were selected.
9. In order to investigate what possible effect text-length could have on linguistic variation we included a few samples that only deviated with respect to this variable but otherwise conformed to the set of criteria.
10. The decision to select the samples from the middle of the book rather than the beginning or the end was to some extent rather arbitrary. However, it was decided to select the samples from the middle of the book rather than the beginning because it was felt that in this way any differences in the time writers take to introduce the various characters and so on would be neutralized. Then, for more practical reasons, we chose the middle section rather than the end because it would be difficult to determine at what place to start in order to get a 20,000 word sample running up to the very end of the book.
11. Material that was not included was generally replaced by tags. A full account of these tags and other coding symbols that were used in the processing of the text can be found in Oostdijk (to appear – 2).
12. For the coding key, see Oostdijk (to appear – 2).
13. 20,000 words would be the minimum since a sample would end with the utterance that contained the 20,000th word.

References

Aarts, J. & W. Meijs (eds.). 1984. *Corpus linguistics. Recent developments in the use of computer corpora in English language research.* Amsterdam: Rodopi.

Aarts, J. & W. Meijs (eds.). 1986. *Corpus linguistics II. New studies in the analysis and exploitation of computer corpora.* Amsterdam: Rodopi.

Biber, D. & E. Finegan. 1986. An initial typology of English text types. In: Aarts & Meijs (eds.) (1986), 19-46.

Ellis, J. & J.N. Ure. 1969. Language varieties: Register. In Meetham & Hudson (eds.) (1969), 251-259.

Gregory, M. 1967. Aspects of varieties differentiation. In *Journal of Linguistics* 3: 177-198.

Haan, P. de. 1984. Relative clauses compared. In *ICAME News* 8: 47-59.

Haan, P. de & R. v Hout 1986. Statistics and corpus analysis. In: Aarts & Meijs (eds.) (1986), 79-98.

Johansson, S. with G. Leech and H. Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English.* Oslo: English Department. University of Oslo.

Meetham, A.R. & R.A. Hudson (eds.). 1969. *Encyclopaedia of linguistics, information and control*. Oxford: Pergamon Press.

Oostdijk, N. (to appear – 1). A corpus linguistic approach to linguistic variation. To appear in *Literary and Linguistic Computing*, Vol. 3 No. 1.

Oostdijk, N. (to appear – 2): *TOSCA Corpus – Manual*.

Renouf, A. 1984. Corpus development at Birmingham University. In Aarts & Meijs (eds.) (1984). 3-39.

The Kolhapur Corpus of Indian English and work done on its basis so far

S. V. Shastri
Shivaji University
Kolhapur, India

1. Introduction

The Kolhapur Corpus of Indian English has now been completed and copies are available from the Department of English, Shivaji University, Kolhapur-416 004, to researchers in India and from the Norwegian Computing Centre for the Humanities to researchers outside India (see the announcement on p 85).[1]

This is a million-word corpus of Indian English comparable to the existing Brown and LOB corpora. The genres of writing, types of text, the weighting of these, etc. are all kept as close to the other two corpora as possible. However, there are two differences: (1) while the Brown and LOB corpora draw their samples from material published in 1961, the Indian corpus draws its samples from those published in 1978; and (2) the section on Imaginative prose comprising six categories of Imaginative prose departs from the other two. Both the weighting given to the different categories of material and the ratio between full-length novels and short stories differ to some extent. This is because of the inherent difference in the Indian situation. The actual composition of the Indian corpus as compared to those of the other two is shown in Table 1.

Table 1. The basic composition of the American, British and Indian English corpora

	Text categories	No. of texts in each category American corpus	British corpus	Indian corpus
A	Press: reportage	44	44	44
B	Press: editorial	27	27	27
C	Press: reviews	17	17	17
D	Religion	17	17	17
E	Skills, trades and hobbies	36	38	38
F	Popular lore	48	44	44
G	Belles lettres	75	77	70
H	Miscellaneous (Govt. documents, foundation reports, industry reports, college catalogue, industry house organ)	30	30	37
J	Learned and scientific writings	80	80	80
K	General fiction	29	29	58
L	Mystery and detective fiction	24	24	24
M	Science fiction	6	6	2
N	Adventure and Western fiction	29	29	15
P	Romance and love story	29	29	18
R	Humour	9	9	9
	TOTAL	500	500	500

2. Work done so far on the basis of the Kolhapur Corpus

Almost simultaneously with the inception of the project for building the Indian English corpus several researchers got interested in probing it with a view to discovering features of Indian English. The present author is about to complete a comparative study of word frequencies in Indian, British and American English (Shastri 1985). A somewhat theoretical work on 'Code-mixing in the process of Indianization of English' is due to appear in *Indian Linguistics*. Researchers in the Department of English, Shivaji University, Kolhapur, have explored several areas of Indian English. The following are summaries of the studies completed so far.

2.1 The meanings of the modals in Indian English

Smt. Pratibha Katikar

This is the first large-scale study of an important aspect of IndE based on the Kolhapur Corpus of Indian English. It takes a cue from similar studies of the meanings of the modals based on the Brown Corpus of American English (Hermerén 1978) and the LOB and Survey of English Usage corpora of British English (Coates 1983). The scope of the study is restricted to the nine modals: *shall, should, will, would, can, could, may, might* and *must*, as they occur in 35 out of the 500 texts of the corpus. As the main aim of the work is to compare the use of these modals in IndE with that in the native varieties, AmE and BrE, the methodology is kept as close to the earlier studies as possible.

By and large the modal usage in IndE conforms to the modal usage in the native varieties. This may be due to the fact that modals in English are polysemous and so there is little chance of a modal being used wide of mark. Hence, the monolythic character of the modal auxiliaries in all the varieties of English may be attributed to the polysemous nature of modals rather than to any special mastery over this area by the users of a particular variety.

Regarding differences, according to the British and American scholars, the contracted forms like *'ll* and *'d* for *will* and *would*, are not used for conveying 'determination'. But in IndE these forms are not infrequent as exponents of this modality. It may be that the Indian users of English do not differentiate between the contracted and full forms.

Secondly, there are a few cases of confusion in the use of modal forms in conditional constructions and indirect speech. The most common slip seems to be a failure to conform to the rules of the sequence of tenses. There are instances of the

use of a present tense form of the modal where a past form is required and *vice versa*. Speculatively, the use of the past tense form instead of the present, may be attributed to the over-generalization of the politeness factor of the meaning. The formwise frequencies of the modals in the data studied also show an overall predominance of the past forms of the modals in IndE as compared to native English. This may marginally be due to a failure to master their use in a few contexts such as conditional and indirect form of speech but the bulk of the data seems to indicate that the Indian user of English prefers a past form where a present tense form is common.

Another modal which has a higher frequency in IndE is *shall*. This may be due to the predominance of written language over spoken in the Indian pedagogical context. Also, English in India, taught as a second language, tends to retain some of the older usages which might have lost currency in the first language situation.

A comparison of the frequency figures of the modalities in IndE and native English shows that the modality of 'futurity' and 'hypothesis' have a low frequency in IndE whereas the modality of 'certainty' has a higher frequency. One wonders if these are reflections of the peculiar Indian mode of thought. Maybe the Indian mind is not given to thinking much in terms of the future and if the western culture shies away from categorical or strong views, the expression of certainty, the Indian mind does not. In the case of the modality of 'hypothesis', the Indian user of English seems to find it difficult to cope with the complicated syntactic structure. This is suggested by the considerable number of instances of idiosyncratic use of modals in such constructions.

Within its limited framework, the study indicates that IndE differs very marginally in respect of one very important area of the structure of English, namely, modality and its expression.

2.2 A linguistic study of Indian English newspaper headlines with special reference to speech reporting

Smt. Sumati Salunkhe

This study is based on the press materials in the Kolhapur Corpus of Indian English, comparable to the Brown and LOB corpora. All the major headlines occurring in the three categories of press materials (cf. Table 1), were extracted manually and analysed. The main object of the analysis was to discover how one makes sense of newspaper headlines which obviously have a reduced kind of structure. It seems that several levels of language structure are involved in the process of

understanding newspaper headlines – syntactic, semantic and discourse. A good deal of work had already been done on the subject by scholars like Straumann (1935), Halliday and Hasan (1976), Kress *et al* (1979), Leech (1963, 1966), Leech and Short (1981). A working model based on the works of these scholars was evolved and used in the analysis of headlines. A detailed analysis of a typical headline took the following shape:

Headline:		*Crushing*	*Delay*	*UP*	*Cane*	*Farmers*	*Sore*
		Present Participle	N(deverbal)	N	N	N	Adj
Syntactic labels	Formal	NP			NP		Adjectival
	Functional	Z			S		Cs
Semantic labels	Implication	process			attribuend		attribute
	Function	ideational (event)					
Discourse function	Internal	cause			effect		
	External	summary					
Meaning		transparent					

When the verbal group is absent, the subject and the complement get neutralized. Leech has called this neutralized element 'Z'.

In all 334 major headlines occur in the press materials of the IndE corpus. They represent a wide variety of syntactic structures within the framework of block language as well as some full major clauses.

It was discovered that the frequency of occurrence of different headline structures varied from category to category. Single-item group structures were found to be the most frequent in the 'B' category (editorials). More complex group structures were the most frequent in the 'A' category (reportage) closely followed by the 'B' category. Clause structures including full clauses were the most frequent again in the 'A' category and the least frequent in the 'C' category (reviews). On the whole, however, clause structures predominated in the 'B' and 'C' categories.

The point of view of the reporter was discovered to be reflected quite clearly in a good number of headlines in the 'A' category. A pair of headlines reporting a similar event might clarify the point:

Three Die in Akali-Police Clash in Delhi, Curfew Imposed
Nine Killed, Thirty Hurt in Hyderabad Police Firing

The first headline reports the event in a neutral way (*die* as a process not indicating the agent) whereas the second headline reports the event as a process (*kill*) involving the agent (*police*). The phenomenon has been taken note of because it happens to be an important aspect of the language of press reporting. It determines both the dissemination and control of information. Obviously there is a need for research on this aspect of newspaper reporting.

The declared objective of the investigation was to focus on speech reporting. But, since very few such instances turned up in the materials used, the focus shifted away from speech reporting. However, the few instances encountered in the materials were analysed according to the framework of Short (1984). It was discovered that four types of speech reporting occurred in the headlines: DS (Direct Speech), IS (Indirect Speech), SS (Speech Summary), NRSA (Narrative Report of Speech Act). In most cases the veracity of speech reporting turned out to be sustainable. Only in one case was it found to be questionable. Almost all the reporting verbs used in speech reporting were discovered to be highly 'loaded' whereas the actual report did not warrant those. In some cases they were overstated and in others understated.

2.3 *If*-constructions in Indian English

Mr. B. N. Patil

The investigation of the nature of *if*-constructions in IndE was suggested as a topic by Katikar (1984) in her thesis on 'The Meanings of the Modals in Indian English'. One of the findings of her work was that the use of modals and tense forms of verbs was found to be deviant in Indian English.

The present study of *if*-constructions in IndE is based on all the 500 texts in the corpus. In all 1,655 instances of *if*-constructions occur in the entire corpus; out of these 137 sentences appear to be deviant (i.e. 8.27%) from standard English usage. No attempt was made to compare the usage with that in the British or the American corpora. Standard English usage was determined on the basis of the works of well-known grammarians and used as a basis for comparison.

Following Declerck (1984) the deviant instances were classified into standard conditionals and non-standard conditionals. The standard conditionals were further classified into open, hypothetical and counter-factual types, and the non-standard conditionals were further classified into 8 types proposed by him. The distribution of the deviant instances of *if*-constructions in IndE is shown in Table 2.

From the table it is clear that the bulk of deviant instances is concentrated on standard conditionals. The exact nature of deviation again pertains to the use of the past forms of modals instead of the present forms mostly in the head clause. The motivation for this seems to be an anxiety to be 'polite'. Where past forms occur in both clauses the motivation seems to be semantic, i.e. expression of tentativeness. There are some instances where the tense form shifts from the present to the future (*will*+infinitive), maybe because the present tense form in English also has future time reference.

Thus the study supports the hypothesis that the past forms of modals are predominant in *if*-constructions and the sequence of tenses is often distorted.

Table 2. Deviant *if*-constructions

		Description	No. of instances
1	2	3	4
Standard conditionals	Type 1 Open	A verb in *if*-clause deviant	7
		B verb in head clause deviant	21
		C tense forms of verbs in both clauses deviant	3
		D past forms of modals CAN, WILL and MAY used instead of present tense forms	32
		E inappropriate or superfluous use of modals	8
		F a double modal used	2
	Type 2 Hypothetical	A verb in *if*-clause deviant	2
		B verb in head clause deviant	13
	Type 3 Counterfactual	A verb in *if*-clause deviant	10
		B verb in head clause deviant	7
		C tense forms of the verbs deviant	2
		D doubtful simple past for past perfect in the *if*-clause (typically with BE and HAVE)	7
		E verb form in head clause deviant	1
		F verb form in both clauses deviant	1
			Total 116
Non-standard conditionals	Type 1	closed condition	9
	2	utterance condition	1
	3	q-primary	1
	4	p-primary	2
	5	q-primary with appended free clause	2
	6	expression of strong wish	1
	7	deductive open, close to standard conditional	2
	8	open condition complex time reference	3
			Total 21
			Grand total 137

2.4 A study of *some* and *any* in Indian English

Smt. Sheela Ramtirthakar

This is a study of the syntactic and semantic behaviour of *some* and *any* and some of their compounds in Indian English. The forms of *some* and *any* chosen for the purposes of this study are *some*+N, *someone, somebody* and *something*, and *any*+N, *anyone, anybody* and *anything*.

The occurrences of these strings in the corpus were manually extracted, analysed and compared with Standard English usage as described in the literature on the grammar of *some* and *any* (for example, Quirk *et al* 1972, Klima 1964, Jackendoff 1972, Lakoff 1969, Hogg 1977).

In standard English the occurrence of *some* and *any* is typical of assertive, and non-assertive and conditional constructions, respectively. However, although *any* is the province of Negative, Interrogative and Conditional (NIC) constructions, *some* can also occur with a difference in 'meaning'. The study is, therefore, further restricted to the analysis of *some* and *any* in NIC's.

The absolute frequencies of occurrence of *some* and *any* in our entire data are 1489 and 1377, respectively. The NIC constructions account for 96 occurrences with *some* and 730 with *any*. On the face of it the general rule that *any* is the province on NIC's is largely observed in Indian English. However, it was necessary to subject the data to further investigation to ascertain whether the choice of *some* and *any* in the syntax types (NIC's) was semantically in order, i.e. *some* where the presupposition/implication is positive and *any* where the presupposition is neutral or negative. Some 44 sentences required inspection of a 'larger context' to ascertain this and the others were self-explanatory. It was discovered that the choice of *some/any* was in order in all but 37 instances (6 with *some* and 31 with *any*), i.e. 4.4% of all the NIC's in the data. Of these 37 instances, again, it was discovered that 28 were syntactically unacceptable and 9 semantically deviant. The syntactic unacceptability of the 28 instances is the result of violating a variety of rules not necessarily directly connected with the choice of *some/any*; they are: (1) lack of inversion in a negative construction with the negative element in initial position (but *seldom* anyone expressed.....), (2) insertion of dummy auxiliary where not required (....plot which *does* never build up to any effective climax), (3) confusion of negation of NP and negation of VP (*any* employee shall not be required to...), and so on.

But the 9 instances of semantically deviant sentences contain *some/any* in NIC's not matched by the required presupposition/implication.

A broad conclusion that can be drawn from the study is that there is hardly any difference between standard English and Indian English in the use of *some/any*.

2.5 Verb-particle constructions with *up* and *down* in Indian English

Mr. S. T. Shingate

The object of this study was to explore the hypothesis that particles in verb-particle constructions (VPC's) in Indian English are often superfluous/redundant. All the five hundred texts of the Kolhapur Corpus were used as source material for the study. The particles *up* and *down* occur 2,310 times as part of VPC's in the corpus. These instances cover 376 different VPC's of which 242 VPC's contain *up* and 134 *down*. Out of these, 59 VPC's with *up* and 45 with *down* were discovered to display several semantic features peculiar to Indian English usage. In terms of percentages 27.65% of VPC's in the data show peculiar features. However, it must be mentioned that not all the instances of these VPC's show peculiar features. In other words the peculiar features may be said to be tendencies in Indian English. The peculiar semantic features noted in IndE usage are the following:

1) VPC's in which a different verb or particle synonymous with the standard English verb or particle occur, eg: *billow* up = send up, draw *up* a chair = draw out a chair.
2) Transitive VPC's used intransitively, eg: Vinayak Shastri looked *up* at him.
3) Peculiar collocation with the subject, object and complement NPs, eg: Organization ... make up its mind.
4) Non-causative VPC's used causatively, eg: The tube will be *sprung up*.
5) VPC's in which the particle is superfluous, eg: *Dig up* wells.
6) VPC's in which the particle is redundant, eg: *Rise up, stoop down*.
7) VPC's used in the passive or attributively, eg: He was *settled down* in New Zealand.
8) Idiosyncratic, eg: If we *cut up* your goat ...
9) VPC's used with a creative meaning, eg: Willie's eyebrows *arched up* a good quarter inch.
10) VPC's which seem to reflect some kind of Indian social or cultural reality, eg: If the previous evening his favourite foot-ball team has lost the match, he *beats up* his children for no reason.

Of these the most frequent are superfluous/redundant use of particles occurring 59 times representing 6 different VPC's. The next is that of peculiar collocations with a frequency of 58, also limited to 6 VPC's.

According to many linguists, VPC's in English are productive or creative, and syntactic and semantic rules governing the well-formedness of VPC's have been proposed. Our observations regarding the peculiarities of IndE usage may well be within the creative mechanism of the language. It is necessary, therefore, to verify how far our observations are tenable through native speaker attestation tests. Even before that it would be useful to investigate VPC's in the LOB and BROWN corpora and see how they behave in actual use.

3. Prospects

Other topics being investigated currently are: *-s* genitives in Indian English, the use of articles in Indian English, the behaviour of a selected set of verbs in Indian English and the collocability of certain verbs in Indian English.

The present author is planning to prepare a KWIC concordance of the Kolhapur Corpus and undertake a project on grammatical tagging of the Corpus. Long-term plans include the compilation of a dictionary of Indian English based on the Corpus.

Note

1. This project was supported by U.G.C. with substantial financial assistance.

References

Coates, J. 1983. *The semantics of the modal auxiliaries.* London: Croom Helm.

Declerck, R. 1984. 'Pure future' *will* in *if*-clauses. *Lingua* 63: 279-312.

Fodor, J. A. and J. J. Katz (eds.) 1964. *The structure of language.* Englewood Cliffs: Prentice-Hall.

Halliday, M. A. K. and Ruqaiya Hasan. 1976. *Cohesion in English.* London: Longman.

Hermerén, L. 1978. *On modality in English: A study of the semantics of the modals.* Lund: GWK Gleerup.

Jackendoff, R. S. 1972. *Semantic interpretation in generative grammar.* The MIT Press.

Katikar, P. B. 1984. The meanings of the modals in Indian English. Unpublished Ph. D. dissertation, Kolhapur, Shivaji University.

Klima, E. S. 1964. *Negation in English.* In Fodor and Katz (1964).

Kress, Gunther (ed.) 1979. *Language as ideology.* London: Routledge and Kegan Paul.

Lakoff, R. T. 1969. Some reasons why there can't be a *some-any* rule. *Language* 45: 608-615.

Leech, G. N. 1963. Disjunctive grammar in British television advertising. *Studia Neophilologica* 35: 356-64.

Leech, G. N. 1966. *English in advertising: A linguistic study of advertising in Great Britain.* London: Longman.

Leech, G. N. and M. Short. 1981. *Style in fiction.* London: Longman.

Quirk, R., S. Greenbaum, G. N. Leech and J. Svartvik. 1972. *A grammar of contemporary English.* London: Longman.

Shastri, S. V. 1985. Word frequencies in Indian English: A preliminary study. *ICAME News* 9: 38-44.

Shastri, S. V., C. T. Patilkulkarni and Geeta Shastri, 1986. *Manual to accompany the Kolhapur Corpus of Indian English for use on digital computers.* Kolhapur: Shivaji University, Department of English.

Shastri, S. V., (forthcoming). Code-mixing in the process of Indianization of English: A corpus-based study. To appear in *Indian Linguistics.*

Short, M. H. 1984. Speech presentation, the novel, and the press. Paper presented at the 7th AILA World Congress, Brussels, Belgium, August 1984. (unpublished)

Straumann, H. 1935. *Newspaper headlines. A study in linguistic method.* London: George Allen Unwin Ltd.

The subjunctive in British and American English

Stig Johansson
Else Helene Norheim
University of Oslo

1. Aim

The subjunctive is one of the few areas of grammar where there are differences between standard British and American English. Attention is often drawn to the more frequent use of the subjunctive in American English (e.g. in Galinsky 1952:207ff, Harsh 1968, and Quirk *et al* 1985:156ff). In our paper we examine the evidence provided on this point by two comparable million-word corpora, the American Brown Corpus (cf Francis and Kučera 1979) and the British LOB Corpus (cf Johansson *et al* 1978). A fuller account is given in Norheim (1985).

2. Definition

The subjunctive is a fairly marginal category in present-day English. It has been claimed that it is dying, 'except in a few easily specified uses' (Fowler and Gowers 1965:595). Palmer (1974:48) even states that it 'has no place in English grammar'. There is no doubt, however, that English verbs have distinctive forms under certain circumstances which differ from the normal indicative forms and convey the meaning of 'non-fact', which is characteristic of the subjunctive in other languages. In our paper we follow Quirk *et al* (1985:156ff), who distinguish between three main uses of the English subjunctive: the mandative subjunctive, the formulaic subjunctive, and the *were*-subjunctive.

3. The mandative subjunctive

The mandative subjunctive is used in *that*-clauses after expressions of demand, order, wish, etc:

1. He insists that it (should) be placed in a special museum.
2. He insisted that she (should) not come in late.
3. They expressed the wish that the visit (should) be postponed.
4. It is important that he (should) join us.

The subjunctive is identical to the base form of the verb. There is no concord with the subject, no backshifting of tense depending upon the superordinate verb (cf examples 1 and 2), and no *do*-periphrasis in negative constructions (example 2).

The modal auxiliary *should* can normally be inserted, with no appreciable difference in meaning. The generally accepted view is that the mandative subjunctive is more common in American English, while British English prefers the *should* construction and uses the subjunctive only in formal and legalistic style.

A study of evidence from the two corpora confirms the general view on the mandative subjunctive. The choice of verb form was examined in *that*-clauses after a selection of 'suasive' verbs and adjectives (see Table 1).[1] Under some of the verbs we also include corresponding nouns which show similar patterns of complementation: *demand, desire, proposal, recommendation, request, requirement, suggestion, wish*. Among the verb forms in the *that*-clauses we include, apart from distinctive subjunctive forms and *should* constructions, 'non-distinctive' forms, as in:

5. We insist that you go.
6. I suggested that we leave at once.

If we insert a third-person singular subject here, we could easily get a distinctive subjunctive form:

7. We insist that he go.
8. I suggested that she leave at once.

In our material we also found examples with other modal auxiliaries than *should* and a single example (11) a distinctive indicative form:[2]

9. The Charter does stipulate that "due regard" shall be paid to the importance of recruiting the staff on... BROWN B07:0150.
10. But it coupled with this a requirement that Indians must bring their pelts to Mobile and thus... BROWN F45:1050.
11. Feeling it would not be wise to rush matters so soon he finished his drink and suggested they returned to the dance room. LOB P07:87.

Such examples are too few to be included in our comparison.

Table 1. The mandative subjunctive vs *should* after selected verbs (including corresponding nouns) and adjectives. Non-dist = non-distinctive forms.

Governing	Brown Corpus			LOB Corpus		
word	*should*	subj	non-dist	*should*	subj	non-dist
advise	1	2	-	3	-	-
ask	-	5	4	2	1	-
beg	-	1	-	-	-	-
demand	-	19	1	3	2	1
desire	1	1	-	1	-	-
direct	-	2	-	1	-	1
insist	2	9	4	8	-	1
move	-	1	-	-	1	-
order	1	2	-	-	1	-
propose	1	9	3	5	-	1
recommend	1	10	3	13	1	-
request	-	6	1	-	2	-
require	-	14	2	6	1	1
stipulate	-	2	-	1	-	-
suggest	7	12	7	34	2	6
urge	-	6	1	2	-	-
wish	-	3	-	2	1	-
anxious	-	1	-	2	-	-
essential	1	2	-	7	1	-
important	3	4	3	-	-	-
necessary	1	5	1	5	-	-
sufficient	-	-	-	2	1	-
Total	19	116	30	97	14	11

Table 1 shows clearly that preferences are quite different in British and American English. While the subjunctive is the normal choice in the Brown Corpus, the number of subjunctives and non-distinctive forms is very low in the LOB Corpus. Of the 14 distinctive forms subjunctive forms in the LOB Corpus 11 contain a passive verb construction (while the proportion of passives among the *should* constructions is only 35%).[3] This agrees with the previous observations that the verb *be* is a 'stronghold of the subjunctive' (Turner 1980:276). It is the only verb which has distinctive subjunctive forms with all types of subject. The cooccurrence with the passive also illustrates the formal nature of the subjunctive in British English; as is well-known, the passive is characteristic of formal, impersonal prose. All the examples except one occur in the categories of informative prose of the LOB Corpus. A further indication of formality is that only one of the LOB subjunctives is found in a clause not introduced by the conjunction *that*.[4]

Our corpus material also confirms previous observations on the negative subjunctive, which has been called a 'typical American construction' (Kirchner 1954:123). The material contains very few *that*-clauses with negative constructions (7 in each of the corpora), presumably because it is more natural to command what is to be done than what is not to take place. The 7 negative examples in the LOB Corpus all contain *should*. Of the 7 examples in the Brown Corpus only one has a *should* construction; the other examples are:

12. The council advised the governor that 'large supermarkets, factory outlets and department stores not be allowed to do business' on Sunday. BROWN A05:0880
13. This dissatisfaction led to Howsam's request that the video not be terminated before the end of the game. BROWN A13:1620
14. ... it is also essential that small shopping areas 'not be overlooked if our small merchants are to survive'. BROWN A19:1050.
15. In establishing conditions of self-help, it is important that we not expect countries to remake themselves in our image. BROWN H02:0630
16. I urge once again that the Congress not reenact this rider. BROWN H21:1120
17. On the one hand do we argue the Supreme Court decision required only that a child not be denied admission to a school on account of his race? BROWN J48:1330

The effect of the negative subjunctive is to underline the negation, as *not* is in a position where it attracts stress and cannot be reduced. All the examples of the negative subjunctive occur in formal contexts; note that four of the examples above

contain passive constructions. The positive examples as well typically occur in fairly formal contexts, but there is also sprinkling of examples in less formal texts. In other words, the mandative subjunctive is not only more frequent in American English than in British English but also has a wider range of application.

4. The formulaic subjunctive

Like the mandative subjunctive, this use is expressed by the base form of the verb. It is limited to set expressions of the type *God save the Queen, Be that as it may.* Both corpora contain a sprinkling of examples, e.g.

18. Be it enacted by the Senate and the House of Representatives... BROWN H09:0010, 1340, 1660
19. So be it... BROWN B20:0130
20. Be that as it may, the trade is of very great importance... LOB H21:5
21. ... if that was the white man's custom, he had said, so be it. LOB K29:81

The formulaic subjunctive is rare in both corpora and has similar stylistic implications. In the words of Quirk *et al* (1985:158), it 'tends to be formal and old-fashioned in style'.

5. Other uses of the base-form subjunctive

The base-form subjunctive is occasionally found in adverbial clauses, chiefly clauses of condition (see Table 2). Examples:

22. 'Now there's a man I'd tie to, if he ever give me the chance,' the constable told himself... LOB N03:27
23. His objection to it is that if mind be the product of the brain, it would be subject like the brain to the law of atomic change. LOB D14:79
24. If there be a disinterested patriot in America, 'tis General Washington, and his bravery, none can question.' BROWN G58:1270
25. ... if such a demonstration be made, it will not find support or countenance from any of the men whose names are recognized as having a right to speak for Providence.' BROWN J58:1180.
26. Here, too, a change of wording is imperative; unless, indeed, question 53 be deleted altogether, which we ourselves would favour... LOB D10:54
27. Nothing in English has been ridiculed as much as the ambiguous use of words, unless it be the ambiguous use of sentences. BROWN R05:0020

Such examples are infrequent in both corpora; most of them are found in fairly formal texts and contain the verb *be*. Both corpora provide a few instances of the subjunctive where *whether* introduces a clause listing alternatives:[5]

28. ... a relentless rooting-out of all inefficiency, restrictiveness and waste, whether it be of capital resources or of labour. LOB A21:208
29. The same delivery vehicles – whether they be airplanes, submarines or guided missiles – should be usable. BROWN J08:1760

The limited evidence does not suggest that there are any major differences between British and American English in the use of base-form subjunctives in adverbial clauses. The only exception is clauses introduced by the formal and archaic-sounding *lest*, which is generally recognised as being more typical of American English:

30. When she appeared at the store to help out for a few hours even my looking at her was surreptitious lest my Uncle notice it. BROWN N18:0630

There are 17 sentences in the Brown Corpus where *lest* occurs; 8 of these have a subjunctive form. There is no case where a distinctive indicative form is used where the subjunctive might have been chosen; the subjunctive is used wherever it is possible. There are only three clauses with *lest* in the LOB Corpus, and none of them have a subjunctive form.

6. The *were*-subjunctive

The verb *to be* allows a contrast in the past tense between indicative *was* and subjunctive *were* with 1st and 3rd pers sing subjects. The *were*-subjunctive is 'hypothetical and unreal in meaning, being used in adverbial clauses introduced by such conjunctions as *if, as if, as though, though*, and in nominal clauses after verbs like *wish* and *suppose*' (Quirk *et al* 1985:158). Examples from the two corpora:

31. It felt as if she were alone in the world. LOB P16:79
32. ... look at the tortured flesh as though it were a bone dug up from London Clay. LOB G25:94
33. Most men would unhesitatingly use a machine if it were available. LOB E02:122
34. Almost as if I were talking about something quite unreal. BROWN J62:0370
35. But there was a look about her mouth as though she were tasting lemons. BROWN K23:0080
36. If there were only darkness, all would be clear. BROWN G12:1050

Table 2. The base-form subjunctive (except the mandative use) in subordinate clauses

Type of clause	Brown Corpus	LOB Corpus
if	9	14
on condition	1	0
provided	1	1
unless[6]	5	2
even though	1	0
though	2	1
however+ adj	1	0
whatever (=*no matter what*)	0	1
wherever	1	0
whether	4	8
for fear that	0	1
lest	8	0
so that	2	0
before	0	2

Table 3. Indicative *was* vs subjunctive *were* after 1st and 3rd pers sing subjects

Type of clause	Brown Corpus		LOB Corpus	
	was	*were*	*was*	*were*
as if	8	35	15	33
as though	1	19	9	22
even if	4	3	10	7
even though[8]	4	0	2	0
if (hyp cond)[9]	28	56	38	64
if (=*whether*)	(not counted)	1	(not counted)	2
whether	(not counted)	2	(not counted)	2
unless	(not counted)	0	(not counted)	6
nominal clause after *wish/suppose*	(not counted)	5	(not counted)	5

According to Quirk *et al (ibid)*, the *were*-subjunctive is 'nowadays a less usual alternative to the hypothetical past indicative'. However, the *were*-subjunctive is clearly the dominant choice in hypothetical-conditional clauses and in clauses introduced by *as if* and *as though* (see Table 3).[7] A study of the distribution in text categories reveals that the subjunctive percentage is higher in the categories of informative prose (A-J), approximately 70-80%, but the subjunctive is also common in fiction (K-R), where it occurs in about half of the possible cases. Both corpora show the same tendencies.

In other types of subordinate clauses the *were*-subjunctive is used much more rarely. The limited evidence does not suggest that there are any differences between British and American English. A detail worth pointing out is, however, that the phrase *as it were* (not included in Table 3) occurs 4 times only in the Brown Corpus, as against 17 in the LOB Corpus.

7. Conclusion

A study of the two corpora reveals two main areas of subjunctive use: the mandative subjunctive and the *were*-subjunctive.

A. The mandative subjunctive is characteristic of the American material, while a construction with *should* is the overwhelmingly preferred alternative in the British corpus. This agrees with many previous observations on American vs British usage.

B. The *were*-subjunctive is used to much the same extent in the two corpora. In both corpora the *were*-subjunctive is distinctly preferred to indicative *was* in hypothetical-conditional clauses and clauses introduced by *as if* and *as though*. The result is hard to reconcile with Quirk *et al*'s (1985:158) observation that the *were*-subjunctive 'is nowadays a less usual alternative to the hypothetical past indicative'.

It is possible that usage has changed since the publication of the texts of the two corpora (1961), which may partly explain the disagreement between our findings and the observations on the *were*-subjunctive in Quirk *et al* (1985). If the *were*-subjunctives may be getting less frequent, results from elicitation tests (Johansson 1979, Turner 1980) suggest that the mandative subjunctive may be on the increase in British English.[10] To study such changes, we need two new comparable British and American corpora. The field is open for new initiatives in corpus building.

Notes

1. If there was more than one clause after a governing word, only the first was included in the material.

2. References to the corpora give the sample code (letter for text category + number of text) and the line number.
3. The proportion of active and passive clauses was more equal in the material from the Brown Corpus: 53 active constructions and 63 passive constructions.
4. Cf the observations in Elsness (1982) on the conditions affecting the use or non-use of the conjunction *that*.
5. Cf also occasional examples like: 'However, be this election year or not ....' LOB A03:141.
6. Four of the five examples from the Brown Corpus occur in quotations from the Bible ('... unless a man be born...').
7. If there was more than one verb in the clause, only the first was included in the material.
8. Note, incidentally, that there is a clear difference between the two corpora in the use of *even if* vs *even though:*

	Brown Corpus	LOB Corpus
even if	61	104
even though	76	37

   No such difference was recorded between *as if* and *as though,* though it has sometimes been pointed out that Americans tend to substitute *as though* for *as if* (McDavid's edition of Mencken 1979:250).
9. Both corpora also provided a handful of instances of hypothetical-conditional clauses with inversion, e.g.: '... were I on my death-bed...' LOB N26:23, '...were it not for my fear that...' BROWN R02:1030. The *were*-subjunctive is obligatory in this case.
10. Haegeman (1986), who examined the use of the mandative subjunctive in the Survey of English Usage corpus, found a number of instances both in formal legal writing and in informal speech. She tentatively suggests that 'it might well be ... that the spoken variety of English has tended to introduce the subjunctive more recently, perhaps because of the influence of American English, while the occurrence of the subjunctive in legalistic writing is the remainder of its original use' (p. 66). In her material she also found indicative and non-distinctive forms, but the *should* variant was the dominant one both in speech and writing.

References

Elsness, J. 1982. *That* v zero connective in English nominal clauses, *ICAME News* 6, 1-45.

Fowler, H.W. and E. Gowers. 1965. *A dictionary of modern English usage.* 2nd ed. London: Oxford University Press.

Francis, W.N. and H. Kučera. 1979. *Manual of information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers*. Rev ed. Providence, R.I.: Department of Linguistics, Brown University.

Galinsky, H. 1951-52. *Die Sprache des Amerikaners*. 2 vols. Heidelberg: Kerle.

Haegeman, L. 1986. The present subjunctive in contemporary British English, *Studia Anglica Posnaniensia* 19, 61-74.

Harsh, W. 1968. *The subjunctive in English*. University of Alabama: University of Alabama Press.

Johansson, S. 1979. American and British English grammar: An elicitation experiment, *English Studies* 60, 195-215.

Johansson, S., G.N. Leech and H. Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Oslo: Department of English, University of Oslo.

Kirchner, G. 1954. *Not* before the subjunctive, *English Studies* 35, 123-125.

Mencken, H.L. and R.I. McDavid. 1963. *The American language*. Abridged edition. London: Routledge & Kegan Paul.

Norheim, E.H. 1985. The subjunctive in present-day British and American English, Unpubl. 'hovedfag' thesis. Oslo: Department of English, University of Oslo.

Palmer, F.R. 1974. *The English verb*. 2nd ed. London: Longman.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Turner, J. F. 1980. The marked subjunctive in contemporary English, *Studia Neophilologica* 52, 271-277.

The *New Oxford English Dictionary* project: A presentation[1]

Stig Johansson
University of Oslo

Aims

The creation of the original *Oxford English Dictionary (OED)* was a monumental achievement. Such a combination of imagination, insight, and sheer hard work has rarely been matched. This classic among dictionaries is the starting-point for the *New OED* project, begun just a few years ago and already promising to become the most important event in computerised lexicography.

The aims of the *New OED* project are 'to bring the *OED* up to date, keep it up to date, use up-to-date tools to carry out these tasks, and offer users of the Dictionary up-to-date tools for using it' (Weiner 1987:6).

The printed integrated *OED*

The aim of the first major phase of the project is to produce a printed integrated edition of the original *OED*[2] and the *Supplements*.[3] To carry out this task, OUP employed a team of lexicographers and computer specialists, supervised by a full-time management team.[4] The approximately 60 million words of complicated text (lightly marked up) were keyborded by some 120 typists over an 18-month period at International Computaprint Corporation of Fort Washington, Pennsylvania.[5] More than fifty freelance proof-readers checked the proofs.

The text is broadly the same as that in the printed versions already available, though with some limited editing to handle slight inconsistencies in the dictionary and some discrepancies in style between the original *OED* and the *Supplements* (Weiner 1987:6f).

During the keyboarding it was only possible to tag a limited number of easily identifiable structural elements. A parser was developed to define the full structure of the text and convert the mark-up to a version of SGML (Standard Generalized Mark-up Language). This work was done in cooperation with researchers at the University of Waterloo, Canada.

The next step was to integrate the text of the *Supplements* into the main body of the dictionary. It has been proven that 80 per cent of the text can be integrated automatically, by the use of computer programs. The remaining 20 per cent involves more complicated adjustments and requires interactive editing. Some 5,000 words and senses, mainly neologisms from the last two decades, are added during

the editing stage. Murray's system of phonetic notation has been converted by computer program to the IPA system.

The integrated editon is scheduled to be published in about twenty volumes in early 1989 (Weiner 1987:7).

A lexical database

The second major phase of the *New OED* project is the enhancement of the electronic database, a task already begun during the integration phase and undertaken by researchers at the University of Waterloo, Canada.[6]

The database management system will provide possibilities of direct access to each distinct information category in the dictionary: pronunciation, part of speech, definitions, quotations, etc. Some suggested database applications which cannot be performed using the printed *OED* are (quoted from Weiner 1987:8, Raymond and Blake 1987):

- Which words contain the sequence of sounds (or letters) ...?
- List all words with definitions that contain a particular word or phrase.
- Most adjectives of the type ... can form adverbs/nouns ending in ... . Are there any adjectives which do not appear to have such corresponding derivatives?
- List the participial adjectives that end in *-ate* (i.e. modelled on the Latin use). Which of them are obsolete or archaic? What is their date range? Who first used each of them?
- How many, and what kind of, words were borrowed in each of the periods specified from ... (language specified)?
- Isolate all Americanisms by first citation rather than by label.
- List words in a particular subject area (e.g. Aeronautics, Agriculture, Anatomy).
- Construct a complete dictionary for the year 1840 (so that one can tell what words and senses were available to a writer at that time).
- Give all references to the concept of 'progress'.
- List all words descended from a particular Greek or Latin root.

In addition, it will of course also be possible to use the electronic database for the same purposes as the printed versions, i.e. ordinary dictionary look-up.

CD-ROM

The electronic database may be made available on line in the same way as newspaper archives, legal databases, etc. Also envisaged are versions on magnetic tape and CD-ROM (Weiner 1987:9). A CD-ROM version of the twelve-volume

OED, without the supplements, was produced in late 1987;[7] see Hodgkin (1987), Hodgkin and Benbow (1987). The CD-ROM version holds approximately 700 Mb and comes on two disks; '... the equivalent of one disk is occupied with indexes and inverted files, but only eight of the forty major fields in the dictionary are searchable and many of the searches one might wish to do within a particular field are not practicable with the technology of the CD-ROM' (Hodgkin 1987:12). Nevertheless, it will be possible, for example, 'to search the whole corpus for occurrences of any word or string, ... to generate wordlists (such as lists of words supported by quotations from Shakespeare, words whose definition(s) contain the word *crime*, words from a given register...), ... to print out lists or extracts from the corpus, and ... to view structure maps which may summarize the form of an entry and give the location of the user within it...' (Hodgkin and Benbow 1987:237). It is expected that a 'much more powerful and flexible dictionary [can be produced] in five years' time' (Hodgkin 1987:12).

Prospects

The existence of the electronic *New OED* will not make the printed versions obsolete; revised editions will continue to be printed (Weiner 1987:9). The existence of the electronic database will simplify continual updating and revision.[8] It will also make the revision more systematic, as it can be organised by category of information rather than alphabetically.

Among the long-term prospects mentioned by Weiner (*ibid*) is the possibility that the *New OED* might be supplemented by 'satellites', such as dictionaries for regional varieties, a thesaurus, a supplementary quotation file, etc. The principal new aspect of the *New OED* is not, however, the extension of the database. The electronic form makes it possible to use the material in new ways, to ask new types of questions, and to discover relationships which were hidden in traditional form of the printed dictonary. In the process, dictionaries will continue to develop, and our insight into the language will be more profound.

After the *OED* lexicography could never be the same. The *New OED* project has made advances which few would have thought possible just a few years ago. Lexicography has reached a new age.

Notes

1. The presentation draws heavily on papers written by members of the *New OED* project team, particularly on the recent article by Weiner (1987). I am grateful to Adam Hodgkin, Director of Electronic Publishing, and Sue Bennett, Assistant Manager of the *New OED* project, for their assistance in providing material for this article.

2. *The Oxford English Dictionary*. Vols 1-12. Ed by James A.H. Murray, Henry Bradley, W.A. Craigie, and C.T. Onions. Oxford 1933.
3. *A Supplement to the Oxford English Dictionary*. Vols 1-4. Ed by R.W. Burchfield. Oxford 1972-86.
4. Important contributions have also been made by IBM (United Kingdom), who have donated equipment, software, and expertise to the value of one million pounds. Assistance has also been obtained from Government sources.
5. Experiments were made with optical character recognition, but this means of data capture was rejected (Weiner 1985c:67).
6. See Gonnet and Tompa (1987), Raymond and Tompa (1987).
7. Co-published with Bowker and Tri Star.
8. See Raymond and Warburton (1987).

References

Benbow, Timothy. 1987. The *New Oxford English Dictionary* project: An introduction. Oxford University Press.

Gonnet, Gaston. H. and Frank Wm. Tompa. 1987. Mind your grammar: A new approach to modelling text. Center for the New Oxford English Dictionary, University of Waterloo, Canada.

Gray, J. C. 1986. Creating the electronic *New Oxford English Dictionary*, *Computers and the Humanities* 20, 45-49.

Hodgkin, Adam. 1987. The uses of large databases. In *Proceedings of the Third Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary*, Centre for the New Oxford English Dictionary, University of Waterloo, Canada, 9-16.

Hodgkin, Adam and Timothy Benbow. 1987. Publishing dictionaries on CD-ROM, in *OPTICA '87*, Oxford: Learned Information (Europe) Ltd, 233-239.

Raymond, Darrell R. and G. Elizabeth Blake. 1987. Solving queries in a grammar-defined *OED*. Center for the New Oxford English Dictionary, University of Waterloo, Canada.

Raymond, Darrell R. and Frank Wm. Tompa. 1987. Hypertext and the *New Oxford English Dictionary*. Center for the New Oxford English Dictionary, University of Waterloo, Canada.

Raymond, Darrell R. and Yvonne L. Warburton. 1987. Computerization of lexicographical activity on the *New Oxford English Dictionary*. Center for the New Oxford English Dictionary, University of Waterloo, Canada.

Simpson, J.A. 1987. The *New OED* project: A year's work in lexicography, *University Computing* 9, 2-7.

Weiner, Edmund. 1985a. Computerizing the *Oxford English Dictionary, Scholarly Publishing*, April 1985, 239-253.
Weiner, Edmund. 1985b. The *New Oxford English Dictionary, Journal of English Linguistics* 18, 1-13.
Weiner, Edmund. 1985c. The *New OED:* Problems in the computerization of a dictionary, *University Computing* 7, 66-71.
Weiner, Edmund. 1987. The electronic *Oxford English Dictionary, Oxford Magazine*, Second Week, Hilary Term, 6-9.

ICAME 8TH

The 8th International Conference on English Language Research on Computerized Corpora in Hanasaari/ Hanaholmen, Espoo, 21-24 May, 1987

More than seventy participants from eleven countries (Australia, Canada, England, the Federal Republic of Germany, Holland, Israel, Norway, Sweden, Switzerland, the United States and Finland) participated in ICAME 8th. Thirty papers were read (see the list below); most of them will be included in the Proceedings of the conference. The publisher will be Rodopi, Amsterdam, and the editors are Ossi Ihalainen, Merja Kytö and Matti Rissanen, who were also responsible for the organization of the conference. Abstracts of most of the papers are given below.

It was most satisfying to notice that many computer corpus projects in progress in the last few years were now either finished or nearing completion. All in all, the reports showed that the development of both corpus building and software continues to make rapid progress. Also, corpus-based studies of individual language problems clearly show the usefulness of this approach to linguistic research.

In addition to the papers, there was a workshop on information exchange through networks and on the coordination and exchange of programs. One afternoon was reserved for a visit to the Research Unit for Computational Linguistics at the University of Helsinki. The work of the Unit was introduced and demonstrated by Prof. Fred Karlsson, Dr. Kimmo Koskenniemi and Dr. Lauri Carlson of the Department for General Linguistics of the University of Helsinki. The Research Unit concentrates on developing a general computational theory which offers principles and formalisms for describing and processing as many natural languages as possible. The Unit focuses on morphology and syntax.

On the leisure program there was a visit to Hvitträsk, the home of Eliel Saarinen and two other Finnish architects of the early 20th century.

List of papers

Jan Aarts and Nelleke Oostdijk (Nijmegen) "Progress report on TOSCA II" – see abstract

Karin Aijmer (Lund) "Report on ongoing research on conversational phrases in English"

Eric Akkerman (Amsterdam) "ASCOT: A computerized lexicon with an associated scanning system" – see abstract

John Algeo (Athens, Georgia) "A computer corpus for a dictionary of Briticisms" – see abstract

Bengt Altenberg (Lund) "Making Brown talk" – see abstract

Eric Atwell (Leeds) "Transforming a parsed corpus into a corpus parser" – see abstract

Nancy Belmore (Montreal) "The use of tagged corpora in defining informationally relevant word classes" – see abstract

Douglas Biber and Edward Finegan (Los Angeles) "Drift in three English genres from the 18th to the 20th centuries: A multidimensional approach" – see abstract

Jeremy Clear (Birmingham) "Towards a process grammar of English"

Peter Collins (New South Wales) "The Australian Corpus Project" – see abstract

Pieter de Haan (Nijmegen) "Structure frequency counts of Modern English: Preliminary results" – see abstract

Mats Eeg-Olofsson (Lund) "Computer processing in the TESS Project" – see abstract

Raymond Hickey (Bonn) "LEXA – A database package for lexicological work on personal computers"

Knut Hofland (Bergen) "Report on the ICAME Program Library and demonstration of the FAFSRV Server on EARN/BITNET"

Ossi Ihalainen (Helsinki) "Working with dialectal material stored in a dBase file" – see abstract

Stig Johansson (Oslo) "Further work on the tagged LOB Corpus" – see abstract

Randall Jones (Provo, Utah) "Using WordCruncher to access the Brown Corpus"

Geoffrey Kaye (Winchester) "The design of the database for the Survey of English Usage" – see abstract

Göran Kjellmer (Gothenburg) "*Even if* and *even though* in Modern English"

Geoffrey Leech (Lancaster) "Corpus analysis for speech recognition"

Willem Meijs (Amsterdam) "*All but* and *if not* in BROWN and LOB" – see abstract

Antoinette Renouf (Birmingham) "Research at Birmingham University" – see abstract

Matti Rissanen and Merja Kytö (Helsinki) "Progress Report on the Helsinki Corpus" – see abstract

Geoffrey Sampson and Robin Haigh (Leeds) "Why long sentences are longer than short ones" – see abstract

Anne-Brita Stenström (Lund) "Adverbial commas and prosodic segmentation" – see abstract

Jan Svartvik (Lund) "TESS at the end of the road"

Lita Taylor and Gerry Knowles (Lancaster) "Progress report on the Lancaster Spoken English Corpus" – see abstract

Gunnel Tottie (Uppsala) "*No*-negation and *Not*-negation in spoken and written English" – see abstract

Wolfgang Viereck (Bamberg) "The data of the *Survey of English Dialects* computerized: Problems and applications" – see abstract

Piek Vossen, Marianne den Broeder, Willem Meijs (Amsterdam) "The LINKS Project: Building a semantic database for linguistic applications" – see abstract

Abstracts

Progress report on TOSCA II

Jan Aarts and Nelleke Oostdijk
University of Nijmegen

The TOSCA II project that is currently carried out at the University of Nijmegen aims at the automatic syntactic analysis of a one million word corpus of contemporary English. For this purpose the tools will be employed that were developed in two earlier projects and which enable linguists to analyse corpora syntactically and to exploit the analysis results (see Aarts & vd Heuvel, 1985). Activities within the project have so far included the writing of a formal grammar for English and the compilation of a corpus.

Although it was not one of the aims of the project, a new corpus of contemporary English has been compiled with a total of one and a half million words. It was not set up as a counterpart of the Brown Corpus or the LOB Corpus. No attempt has been made to arrive at a 'cross-section' of English. Rather, as we intend to use the analysed corpus – in any case – for the study of linguistic variation it consists of a number of fairly large samples (20,000 words) which all belong to the category of printed English, incl. both fiction and non-fiction. It is our experience that such samples are both manageable and sufficiently large for dealing with most questions arising in connection with the study of linguistic variation.

The formal grammar that is being written is nearly complete. It describes most syntactic structures, incl. declarative and interrogative sentences, actives and passives, cleft and extraposed sentences, coordination and subordination in terms of syntactic functions and categories. A number of constructions remain to be included although it is not altogether clear to what extent this will prove to be feasible. From testing the grammar it has become evident that certain realizations are extremely problematic because they are bound to yield a vast amount of ambiguity which on the basis of purely syntactic criteria cannot be controlled. Examples of such problematic instances are prepositional phrases occurring in final position and noun phrases in adverbial position.

Reference

Aarts, J. & Th. vd Heuvel. 1985. Computational tools for the syntactic analysis of corpora, *Linguistics* 23: 303-335.

ASCOT: A computerized lexicon with an associated scanning system

Eric Akkerman and Willem Meijs
University of Amsterdam

In this paper, the output of the ASCOT project[1] was discussed, with emphasis on the lexical products developed in the course of the project. The ASCOT software-package consists of two major components:

1. the ASCOT lexicon (Aslex)
2. the scanning program MULTI_FLEX

The ASCOT lexicon is based on the computer-tape version of the *Longman Dictionary of Contemporary English* (henceforth LDOCE), and was created in four stages:

(a) A computer program was developed to transform the original LDOCE files in such a way that all information became optimally accessible.

(b) The few imperfections which were still left (either due to the program or to the structure of LDOCE) were corrected manually. This resulted in the first ASCOT intermediary file.

(c) On the basis of a detailed analysis of the LDOCE coding system, a number of changes were made affecting the contents of the lexicon. This resulted in the second intermediary file.

(d) Finally, the ASCII file resulting from the work done in the first three stages was used as input for a program which created the actual ASCOT lexicon (Aslex).

In Aslex, the entries, together with their wordclass code and information about inflection and spelling are stored in an L-tree format, which combines efficient use of memory space with very fast access to each lemma. A code-file associated with the lexicon contains the grammatical information pertaining to each lemma.

The scanning system MULTI_FLEX analyses words from (pre-edited) input. It contains an inflection component which relates inflected word forms (including multi-words) to their stem, and a derivation component which establishes the wordclass of derived words which are not in the lexicon, but which consist of a base form and one of the affixes which are incorporated in LDOCE.

When a word is fully coded, it will have codes for wordclass, inflection and all its grammatical characteristics. While the ASCOT software-package can be used interactively, it is basically intended for batch-processing.

For a complete description of the ASCOT project, the reader is referred to:

Akkerman, E., P.C. Masereeuw, W.J. Meijs. 1985. *Designing a computerized lexicon for linguistic purposes*, ASCOT Report No 1, Amsterdam: Rodopi.

Akkerman, E., W.J. Meijs, H.J. Voogt-van Zutphen, forthcoming, *Context free tagging on the basis of a computerized lexicon*, ASCOT Report No 2, Amsterdam: Rodopi.

Note

1. ASCOT, which stands for 'Automatic Scanning system for Corpus Oriented Tasks', was a research project funded by the Dutch Organization for Pure Academic Research (Z.W.O.), under project number 300-169-004. The project started on 1 March 1984 at the English Department of the University of Amsterdam and it was finished on 1 March 1987. Additional funding was supplied by the Arts Faculty of the University of Amsterdam for the period of 1 March 1987 – 1 July 1987, to round off computational work.

The main goals of the ASCOT project were the development of a lexical database system and an associated scanning system, to be employed in (semi-)automatic syntactic analysis.

A computer corpus for a dictionary of Briticisms

John Algeo
University of Georgia

British English is the national variety of the language that has been least satisfactorily described with respect to other national varieties, largely because it has generally been taken as the norm for descriptions of the others. The most convenient way to describe British is to compare it with American.

In the 1930s, Allen Walker Read began work on a Dictionary of Briticisms that would meet the need for a lexical description of the national variety of the United Kingdom seen through American eyes. Since then he has compiled a file of some 100,000 citations as evidence for the dictionary. His citations are drawn chiefly from works of the 19th and early 20th centuries. Read's file is to be combined with the smaller file (at present about 10,000 citations) collected by J. Algeo chiefly

from contemporary sources. The combined files, supplemented to fill gaps and totaling perhaps 5,000,000 words, will be computerized and become the corpus for editing the dictionary.

The Dictionary of Briticisms will aim to include all and only the characteristic lexical features of British English. For each entry it will specify the features that distinguish the word as British, as distinct from American, and will provide illustrative citations as evidence.

The corpus that will serve as the basis of the dictionary will be expanded before and during the editorial process as resources and time permit. When the dictionary has been completed, the corpus will be available for other uses.

Making Brown talk

Bengt Altenberg
Lund University

An automatic text-to-speech conversion system that aims at producing natural-sounding speech must contain, as one of its components, a set of rules that 'chunk' the written input text into appropriate information units or tone units (TUs). A central aim of the research project 'Text Segmentation for Speech' (TESS) at Lund University has been to establish such a rule system on the basis of natural spoken data and to apply the rules to written text (see Svartvik 1984). The present report focuses on the latter of these tasks: the testing and evaluation of a system of segmentation rules applied to a text (newspaper editorial) from the Brown Corpus. The rules are largely based on correlations between TU boundaries and grammatical boundaries in a popular lecture from the London-Lund Corpus (described in Altenberg 1987a and 1987b:46-124), but some adjustments have been made to accommodate the system to the requirements of a written text and to take advantage of punctuation.

The system presupposes a machine-readable input text tagged for word classes and parsed up to phrase level (see Svartvik 1987 and Eeg-Olofsson 1987). The program approaches the text in caterpillar fashion, successively sizing up working strings of a certain maximal length and scanning each string from left to right for potential 'breaking-points' in its grammatical structure. If the structural pattern matches the conditions of any of the segmentation rules, and if no constraints apply, a TU boundary is assigned at the relevant point in the string and the program advances, repeating the pattern-matching process until no further rules apply. Overlong strings without a matching pattern are divided (somewhat arbitrarily) by a default rule.

The current system contains thirteen segmentation rules, roughly ordered according to syntactic type and frequency of application. Each rule is designed to handle a specific type of grammatical boundary where a prosodic break is likely to be made in speech, eg between certain conjoins, before certain subordinate clauses and postverbal phrases, and after subordinate clauses, complex subjects and certain clause-initial adverbials. Some rules are provided with context-sensitive constraints which block the application of the rule under certain conditions; in addition there are general constraints to prevent the division of simple phrases and specific word-class sequences. The punctuation of the text is used as an additional cue in the segmentation.

Roughly estimated, the success rate of the current system is about 87% (disregarding some doubtful cases but including errors due to incorrect tagging and parsing). Further refinements of the system, including a cyclic rearrangement of the rules, a more delicate specification of the contextual constraints, and elimination of the blind default rule, can be expected to raise this figure considerably.

References

Altenberg, B. 1987a. Predicting text segmentation into tone units. In Meijs (1987), 49-60.

Altenberg, B. 1987b. *Prosodic patterns in spoken English. Studies in the correlation between prosody and grammar for text-to-speech conversion.* Lund Studies in English 76. Lund: Lund University Press.

Eeg-Olofsson, M. 1987. Assigning new tags to old texts. An experiment in automatic word class tagging. In Meijs (1987), 45-47.

Meijs, W. (ed.). 1987. *Corpus linguistics and beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora.* Amsterdam: Radopi.

Svartvik, J. 1984. 'Text Segmentation for Speech' (TESS): Presentation of a project. Survey of Spoken English, Lund University.

Svartvik, J. 1987. Taking a new look at word class tags. In Meijs (1987), 33-43.

Transforming a parsed corpus into a corpus parser

Eric Steven Atwell
Leeds University

General probabilistic search techniques have been proposed as methods for parsing the unrestricted English text found in corpora such as the LOB Corpus, but these are computationally much more expensive than standard rewrite-rule-based parsers.

The extra cost can be justified if standard techniques cannot cope with very large rule-based grammars. This paper describes an experiment to extract a very large phrase-structure grammar from a treebank of parsed sentences from the LOB Corpus, and feed it through a parser-generator. The resultant parser was far too large for our Prolog interpreter, as expected. Furthermore, the grammar does not cover all possible syntactic constructs, yet is larger than other known rule-based systems.

The use of tagged corpora in defining informationally relevant word classes

Nancy Belmore
Concordia University

A potential area for fruitful cooperation between specialists in corpus linguistics and specialists in information processing is in the specification of word classes and sub-classes for natural language understanding systems. In such systems, the word classes must be of sufficient delicacy to facilitate an informational analysis of the sentence. The purpose is to be able to translate this information into a standard format so that it can be used for fact retrieval.

This paper reports on some preliminary studies in the use of tagged computer corpora as an aid in arriving at the informationally relevant word classifications which fact retrieval systems necessitate. They are based on a pilot study in which the QUERY program from the University of Amsterdam was used to extract from the Brown Corpus patterns in which words ending in *-ed* and tagged adjective (JJ) or verb (VBN) had occurred as an element in a predicate (*Use mugs to keep beer chilled*) or as an element in a structure of modification (*chilled beer*). From the hundreds of sentences extracted, three small subsets were created, each of which exemplified a problem in classification which the pilot study had brought to light: 1. The problem of specifying the grammatical relations among the constituents of *-ed* modification and predication structures when such relations are obscured by orthographic conventions or require some morphological analysis for their identification, 2. The problem of arriving at an accurate description of the grammatical structures which signal contrasts in meaning between *-ed* words which have one meaning when they function as adjectives and another when they function as verbs, and 3. The problem of specifying informationally relevant sub-classes of the major word classes.

To study the first problem, all the sentences in the relevant subset were tagged by the LOB tagging suite. A new data set was then created which included each *-ed* item, its immediate left constituent, whether another word, a hyphenated word or a prefix, the context of the *-ed* item, and the associated Brown and LOB tags as in the following examples:

	Brown	LOB
was *probably* **prepared**	BEDZ+RB+**VBN**	BEDZ+RB+**VBN**
have been equally *un*prepared	HV+BEN+RN+**JJ**	HV+BEN+RB+**JJ**
the *well*-**prepared** program	AT+**JJ**+NN	ATI+**JJ**+NN

This made it possible to look at the same stem in different grammatical contexts as a means of determining when grammatical analysis below the word level is required.

To study grammatical signals of contrasts in lexical meaning, the data set exemplifying this problem was first tagged by the LOB tagging suite. Then a new data set was created which included the *-ed* word, the contexts in which it had occurred and the Brown and LOB tags, as in the following examples:

	Brown	LOB
the divinely **appointed** means of rescue	AT+RB+**VBN**+NN+IN+NN	ATI+RB+**VBN**+NNS+IN+NN
the small, perfectly **appointed** table	AT+JJ+RB+**VBN**+NN	ATI+JJ+RB+**VBN**+NN

The use of tagged corpora in specifying sub-classes of the major word classes was examined by extracting all instances from the QUERY outputs of words tagged adjective (JJ) which had occurred in Brown in positions characteristic of passive verbs, as in the following examples:

He was universally **beloved** by his neighbors	PPS+BEDZ+RB+**JJ**+IN+PP$+NNS
be well-**ruled** by either militant minority	BE+**JJ**+IN+DTX+JJ+NN

For systems requiring informationally relevant word classifications, such words should probably be treated as a special sub-class of verbs, viz. verbs which occur only in the passive voice.

The three data sets, each of which could be automatically generated from a QUERY-type output, all indicate that tagged computer corpora can be used not only for the special purpose of defining informationally relevant word classes for natural language understanding systems but also to help achieve better general linguistic descriptions.

Drift in three English genres from the 18th to the 20th centuries: A multidimensional approach

Douglas Biber and Edward Finegan
University of Southern California

This study shows how a multidimensional analysis can be used to trace the historical evolution of genres in English. We briefly review five major dimensions of variation identified in previous work. We then show how narratives, essays, and letters have evolved over the last three centuries with respect to three of these dimensions: 'Involved versus Informational Production', 'Elaborated versus Situation-Dependent Reference', and 'Abstract versus Non-Abstract Style'. Our findings show that texts vary diachronically along each dimension. In particular, we show that all three genres are undergoing similar patterns of drift to less elaborated styles of reference and less abstract styles of expression. Narratives shifted earlier and to a greater extent than essays, but both genres are drifting markedly in the same direction.

The Australian Corpus Project

Peter Collins
University of New South Wales, Australia

Three Australian linguists (Pam Peters and David Blair from Macquarie University, and Peter Collins from the University of New South Wales) initiated a project in 1985 to compile a computer corpus of Australian English. The project is planned in two stages, though these are not envisaged as necessarily consecutive.

1. Preparation of a million-word sample corpus parallel to Brown and LOB, to serve as a basis for interdialectal studies. 1986 has been selected as the sampling year (in the interests of contemporary relevance, and in the expectation that recently published material will be readily available in computer-typeset form). Some reweighting of text categories has been found to be required by local differences. For example, with the Press categories (A, B and C) a daily-weekly distinction has been adopted as the primary form of classification (as in Brown, but not LOB). The sampling of newspapers was based on circulation figures, but "intervention" was necessary to ensure that – in the interests of demographic uniformity – some lower-circulation papers from the less-populous states were represented.

2. Compilation of a larger "monitor" corpus, comprising whole texts (in the first instance, those from which the 2,000 word samples for the million-word corpus are extracted). This will incorporate media and genres not represented in Brown and LOB (spoken, non-printed written, dramatic fictional, etc.) This corpus will enable lower-frequency lexical, grammatical and discourse items to be investigated and permit comprehensive register study.

The sample corpus is being compiled in a "modular" fashion, with modules of categories being selected for exploitation in computational studies conducted by the principal investigators. At this stage Peters and Collins are examining informal and interactional language features (grammatical and orthographic) in newspapers and government documents, while Blair is investigating punctuation practices in newspapers, government documents and fictional texts.

Some preliminary comparisons of selected words in Brown, LOB and Australian newspaper material were presented. Findings suggest that while Australian English is clearly moving away from the norms of British English towards those of American English, the process is an intermittent one, with American variants being selectively assimilated into the dialect.

Structure frequency counts of modern English: Preliminary results

Pieter de Haan
University of Nijmegen

In this paper a report is presented of an investigation into the behaviour of so-called prepositional verbs. Two aspects are looked into:

1. The occurrence of prepositional verbs in passive verb phrases, as opposed to the occurrence of adverbial insertion between the verb and the preposition.
2. The position of prepositions in finite relative clauses.

It is argued that the fact that combinations of verbs and prepositions occur in passive verb phrases is the only justification for their classification as prepositional verbs. Nevertheless, the question remains whether within the group of prepositional verbs a distinction can be made between combinations that are more close and those that are less close. This distinction might be visible in their behaviour in the two syntactic environments mentioned above.

The investigation of a small amount of corpus data, currently resident in the *Linguistic Database*, shows us that there are certain tendencies in this respect. However, it appears that a larger amount of data would have to be examined before we can draw any definite conclusions.

Computer processing in the TESS project

Mats Eeg-Olofsson
Lund University

The most important computer processing in the TESS project is performed by the program for automatic segmentation of written text into tone units. The Segmentation program presupposes that the input text has been previously tagged by other programs on the word class and phrase levels.

There are two versions of the Segmentation program at present. A prototype version was written in Prolog and run on a mainframe (VAX) computer. Almost in parallel, a Turbo Pascal version for IBM Personal Computers was developed as a programming project for a group of students of computer science.

The prototype version has proved to be a flexible tool for experimentation with various rules and strategies for segmentation. New rules must be manually translated into the appropriate Prolog code, however, and the execution speed of the program is too low for interactive use.

The alternative version is much faster, being specially designed for graceful interaction with the end user (linguist), emphasizing the use of the terminal screen for menus, status fields, etc. A disadvantage is that its segmentation rules are fixed and cannot be altered by the user. An optimal design would include a rule compiler, embodying the knowledge about suitable general rule formats and strategies gained by experimentation with the prototype version.

The input of the Segmentation program is provided by the Phrase Level Tagging program, written in Prolog and using the Definite Clause Grammar formalism. It analyses each sentence as a sequence of phrases of five kinds: adverb phrase, verb phrase (which can be quite complex, including various auxiliaries and linking verbs), adjective phrase, noun phrase, and prepositional phrase, The program can produce several alternative analyses by backtracking, but some care is taken to present the most likely analysis first. Only heuristic solutions are presented to the classical problems of conjunction reduction and prepositional phrase attachment.

The input of the Phrase Level Tagging program is produced by the Word Class Tagging program, which is written in Pascal (see Eeg-Olofsson, 1987).

In addition, several minor search programs have been written to extract files of instances of various prosodic and grammatical phenomena, for inclusion and further processing in the relational database handling system dBASE III on IBM PC.

Reference

Eeg-Olofsson, M. 1987. Assigning new tags to old texts: An experiment in word class tagging. In *Corpus linguistics and beyond* ed. by W. Meijs. Amsterdam: Rodopi. 45-47.

Working with dialectal material stored in a dBase file

Ossi Ihalainen
University of Helsinki

This short paper shows how from a running text a dBASE file can be automatically created that can be used in actual research and subsequently enriched by further findings.

The language material stored in the demonstration dBASE file is dialectal South-Western British English. The linguistic point discussed is the WAS/WERE variation (e.g. THAT WAS alternating with THAT WERE in the language of a single speaker).

First, the speakers showing this variation are identified. On the basis of the distributional pattern that emerges from their language, it is hypothesized that the variation is not random, but is linguistically conditioned.

Once the primary linguistic data is stored in a dBASE file, findings that are made during the investigation can be inserted in separate fields to accompany the linguistic material. For example, in the case of the WAS/WERE alternation, one would comment on the syntactic position of the verbs and the stress patterns involved. Thus a very powerful database incorporating the analyst's observations of the data is gradually built up that can be used for further computations.

Further studies of the tagged LOB Corpus

Stig Johansson
University of Oslo

The tagged LOB Corpus is being analysed on four main levels: tag frequencies, word frequencies, tag combinations, and word combinations. The results will be presented in a forthcoming book co-authored with Knut Hofland, Bergen (to be published shortly by Oxford University Press). The report focused on problems in analysing word combinations. A study which only includes recurrent word

combinations in a million-word corpus will neither be exhaustive nor sufficiently revealing. The tagged LOB Corpus makes it possible to produce lists of, for example, all the nouns that follow a particular adjective or all the adjectives that precede a particular noun (including both combinations which recur and those which are instanced once only). Each individual combination may not mean very much, but together they may build up a good picture of the combinatory possibilities of words.

The design of the database for the Survey of English Usage

Geoffrey Kaye
IBM UK Scientific Centre, Winchester

The Survey of English Usage was started in 1959. It has as its objective the collection of 500,000 words of written English, and the same amounts of spoken English (transcribed). This is made up of 100 spoken and 100 written texts each of 5,000 words giving 7-10 million characters in all. The intention has been to capture all significant grammatical constructs, and occurrences of closed word categories in these texts. The task is nearing completion.

The primary form of the coded data is in hard copy, filed in about 100 filing cabinets. Significant progress has been made by the sister project at Lund University to convert, in a restricted form, a large amount of the spoken data to machine-readable form. This paper describes the initial stages of, and the long-term objectives for, the conversion of the written texts and their associated grammatical codings into a computer based system.

The main reasons for moving the survey to a computer based system are:

- Easier access to the data. Hard copy in filing cabinets is slow to search.
- Greater usability. The choice of categories under which the coded data is stored makes some investigations extremely time-consuming, and makes others nearly impossible.
- Vulnerability of the data. A single hard copy as voluminous as the Survey is impossible to 'back-up', which means that the whole corpus could be lost in a fire.
- Exploit the power of computer techniques of data storage and interrogation, laying a foundation for the next 25 years of the Survey.

The design of a relational database which is to be used to store the grammatically coded data is described. The emphasis has been on expandibility, and ease of access to the data by linguists with no computational skills. Also described is a real-time interactive concordance browser that runs on an IBM PC and is to be used as a research tool in its own right and as a mechanical aid to the collection of grammatical data. The written texts of the Survey are now in machine-readable form, they have been concordanced and are available at the Survey for use by researchers.

All but and *if not* in BROWN and LOB

Willem Meijs
University of Amsterdam

1. *All but*

All but is difficult for foreign learners because it is ambiguous: it can mean "almost", "nearly" (meaning A), as in (1), or "all, everything except" (meaning B), as in (2):

(1) The large medieval cathedrals look all but empty.

(2) In 1952 Mr. Eisenhower won all but Missouri.

Learners also tend to confuse it with *anything but* meaning "not at all", "far from". Corpus-distribution of *all but* is fairly even: 16 and 8 cases of meaning A versus 14 and 9 cases of meaning B in BROWN and LOB, respectively.

Examination of the cases found in BROWN and LOB suggests that in the A interpretation the **whole** combination should be treated as a (complex) degree adverb, while the **B** meaning should be brought out by part-of-speech labels for the individual **components**, *all* functioning as a pronoun and *but* as a preposition. If *all but* (A) is a degree adverb we will expect it in syntactic contexts in which (degree) adverbs typically occur, i.e. (a) as an intensifier modifying an adjective in adjective phrases (GCE 5.51) or (less commonly), (b) modifying an adverbial prepositional phrase or (c) at sentence level in medial 2 position (GCE 8.7) immediately before the lexical verb or (d) following *be*, before the complement. Similarly, if *all but* (B) consists of a pronoun plus a preposition, we expect a noun phrase to follow, as signalled by: (e) an article, (f) a cardinal numeral, (g) a common or proper noun, or possibly (h) an adjective. The corpus-data on the whole confirm these expectations, as illustrated in the article.

The grammatical tagging of *all but* in BROWN turns out to be rather unsystematic and often incorrect. The LOB tagging, by contrast, is excellent. The 8 A cases are tagged *all RB but RB"*, while the 9 B cases are all tagged *all ABN but IN*. This tallies with our view that in the A meaning the whole combination should be treated as a complex adverb, which is exactly what the "ditto tag" *RB RB"* does, while the B meaning calls for individual tags for the component parts.

2. *If not*

The problem with *if not* is its ambiguity, an excellent account of which is presented in Kjellmer (1975). The ambiguity can be brought out by contrasting the two paraphrases "(possibly) even" and "though not" that could be used to replace *if not* in sentence (3) from LOB, as in (4) and (5):

(3) Christians on both sides of the Curtain face similar, if not identical problems...

(4) Christians on both sides of the Curtain face similar, possibly even identical problems...

(5) Christians on both sides of the Curtain face similar though not identical problems...

The "A-interpretation" exemplified in (4) could be described as "intensifying" or "culminative", the "B-interpretation" going with (5) as "downtoning" or "concessive" – Kjellmer (1975) calls them "inclusive" and "exclusive", respectively.

We found 42 cases of *if not* in BROWN and 44 in LOB. Examination of these on the whole confirmed Kjellmer's observations on aspects which may help to disambiguate it. Thus when there is inversion (i.e. fronting of *if not*), as in "... a way of life which, if not identical, is congenial with ours", only a B reading is possible. We found one further disambiguating feature: certain adverbs also force a B reading, as in "... remarkably similar in atmosphere if not *entirely* in content" and "... *at least* an optical, if not, *properly speaking*, a pictorial illusion". However, in other, unmarked cases the inherent ambiguity of *if not* may well be intentionally exploited, as in "Ah well, our love affair was short if not sweet".

References

Kjellmer, Göran. 1975. "'The weather was fine, if not glorious': On the ambiguity of concessive *if not*". *English Studies* 56/2:140-146. Reprinted in *Eichosha gogaku nenkan*, Tokyo: Eichosha 1977:142-154, 240-243.

Meijs, Willem. 1984. "Data and theory in computer corpus research", in J. Lachlan Mackenzie & Herman Wekker (eds.), *English language research: The Dutch contribution I*. Amsterdam: Free University Press, 85-99.

Research at Birmingham University

Antoinette Renouf
Birmingham University

In December 1986, with the completion of the dictionary, the first stage of the COBUILD lexicographic project came to an end. In January 1987, a re-grouping of project members took place and two new teams were formed. One of these is 'COBUILD Ltd', a company set up by Collins Publishers and the University to produce further reference works based on Birmingham corpus data; the other is 'The English Language Research and Development Unit', consisting of Professor Sinclair, Jeremy Clear and myself, a self-financing unit with the objective of pursuing advanced research in computational linguistics.

A major interest of the Development Unit is collocation, and one particular focus of our study has been linguistic prediction for the purposes of speech recognition. Jeremy Clear's paper at the conference introduced our work on a 'process grammar', a device which makes a structural evaluation of a text fragment and calculates, on the basis of linguistic information derived from a large text corpus, the most likely structural pattern of which it is part. Other of our research interests have been furthered by the work of two M Phil students, attached to the Development Unit under the sponsorship of Collins, in the areas of 'naturalness' and the parsing of nominal groups.

The Unit also continues to be active in corpus development and exploitation. We have recently acquired a one-million-word collection of spoken and written English from the SHAPE Language Centre; and we are about to assemble a corpus of examination papers, published by the University of Cambridge Local Examinations Syndicate, for the purposes of critical examination. The latter enterprise will involve the coding of sub-texts within the data, a matter which I examined briefly at the conference. We are also shortly to begin on the creation of a system that will monitor the linguistic content of on-line text such as newspaper output on a regular basis. The rationale for such a 'monitor corpus' was outlined by Jeremy Clear at the 1986 ICAME conference in Amsterdam.

Progress report on the Helsinki Corpus

Matti Rissanen and Merja Kytö
University of Helsinki

The Helsinki Corpus of English Texts is a computerized corpus divided into two parts: diachronic and dialectal. The dialect part consists of computer transcripts of tape recordings of present-day British English dialects; the diachronic section comprises texts and text extracts dating from the 8th to the 18th century. The main purpose of the Helsinki Corpus is to offer a textual basis for morphological, syntactic and lexical studies.

Most of the texts to be included in the historical part of the Helsinki Corpus have been keyed in and are now being edited. The Basic Corpus so far totals 1.7 million words (Old English 540,000 words, Middle English 582,000 words and Early Modern British English 545,000 words). The dialectal part consists of some 253,000 words transcribed and keyed in from the tapes recorded in various parts of England.

The texts of the diachronic part will be coded according to date, dialect, author, type of text, and a number of sociolinguistic parameters. It was finally decided to omit the style parameter from the coding system. Based on extra-linguistic factors, this parameter appeared redundant and unnecessarily "pre-determining". An attempt to label stylistic levels of texts by linguistic criteria might result in circular statements on the co-occurrence of linguistic and stylistic characteristics of the texts in the later use of the corpus.

A parameter defining the formality level of the text will be retained. Further, the possibility of including a code which would define "diachronic textual prototypes" has been discussed. The following possible categories have been considered: law, secular instruction, religious instruction, non-imaginative narration, imaginative narration, drama, law-court trial, private correspondence, official correspondence. This proposed categorization will not, however, be included in the first version of the corpus because its usefulness can best be estimated through pilot studies based on the corpus.

The LINUS program pack, originally devised by Dr Kimmo Koskenniemi of the University of Helsinki in Extended Algol is now being converted into C-language versions for both main frame and micro-computer use. The preliminary version of the program is scheduled to be available at the same time as the Helsinki Corpus.

The details of the distribution format of the material are still open. It is hoped that the Corpus will be available on magnetic tapes for main frame use and on diskettes and streamer backup tapes for micro-computer use. Both coded (OCP conventions) and uncoded versions are under compilation.

Why long sentences are longer than short ones

Geoffrey Sampson and Robin Haigh
University of Leeds

Technical prose is strikingly different from fiction with respect to average sentence length. There must therefore be statistical differences between the productions used in expanding individual non-terminal symbols in the two genres. A database of parsed sentences is used in order to locate these differences. Certain constructions prove to be significantly commoner in one or the other genre, but this contributes little to the overall difference in sentence lengths; and many constructions are notably stable in frequency across genres. The sentence-length contrast appears to derive almost wholly from quite small differences in mean values for numerical properties of constructions that are common in all genres.

Adverbial commas and prosodic segmentation

Anna-Brita Stenström
Lund University

To facilitate the procedure of working out automatic segmentation rules for text-to-speech conversion, which is one of the aims of the TESS project (Text Segmentation for Speech), we have been looking for places where the segmentation principles of written and spoken texts might correlate. This paper reports on the relation between adverbial commas in the Lancaster-Oslo/Bergen Corpus (LOB) and adverbial tone unit (TU) boundaries in the London-Lund Corpus (LLC).

Although a fairly high degree of correlation might have been expected, the study showed that it was not high enough to serve as a starting-point for suggesting automatic segmentation rules.

The correlation between a comma in LOB and a TU boundary in LLC was fairly strong at the beginning of a sentence but decreased with the distance from initial position. The lack of agreement manifested itself in three ways:

1) the comma in LOB was not matched by a TU boundary in LLC
2) the adverbial was marked off by a TU boundary in LLC, but there was no matching comma in LOB
3) mid-sentence adverbials with a comma on both sides in LOB were either only preceded or only followed by a TU boundary in LLC

One reason for the disagreement is that punctuation in general is fairly conventionalized, whereas speech segmentation is highly dependent on individual speakers and particular speech situations. Another reason is that the same adverbial item can serve more than one function, also in the same position. The segmentation principles often differ accordingly. Sentence-initial *however* can for instance serve as a transitional conjunct ('*However*, he couldn't make it'), as a process adjunct ('*However* the tour was organized, it became a success'), and as an intensifier subjunct ('*However* much he tried, he couldn't get there'). In this case, a comma/TU boundary distinguished the conjunct from the adjunct and the subjunct, but there was not always complete agreement between the written text and speech.

The choice of texts for investigation is important; the correlation between the spoken and written segmentation would for instance most certainly have been higher if a less 'personal' kind of talk had been chosen, such as news reports, where idiosyncracies are avoided.

Rule calculation on the basis of findings in the LOB and LLC corpora alone was abandoned in favour of recommendations based not only on the data but also on previous research and, to a certain extent, on linguistic intuition and common sense.

Progress report on the Lancaster Spoken English Corpus

Lita Taylor & Gerry Knowles
University of Lancaster

The Spoken English Corpus Project was started in September 1984 at the University of Lancaster, and is a three-year project funded by IBM UK Scientific Centre. This report describes the corpus in its final state.

The corpus is composed of 52,837 words of contemporary spoken British English divided into 11 categories containing 53 texts in total:

A	Commentary	G	Fiction
B	News broadcast	H	Poetry
C	Lecture – type I	J	Dialogue
D	Lecture – type II	K	Propaganda
E	Religious broadcast	M	Miscellaneous
F	Magazine-style reporting		

Five versions of the corpus material were identified as being necessary to the work on automatic intonation assignment:

a: spoken recording

b: unpunctuated transcription

c: orthographic transcription

d: prosodic transcription

e: grammatically tagged version

The recorded, unpunctuated, orthographic, and tagged versions are now complete, and the prosodic versions should be complete by the end of summer. – The corpus material is now available to other interested research institutions (see p 85).

No-negation and *not*-negation in spoken and written English

Gunnel Tottie
Uppsala University

This is a report on ongoing research on the use of two types of negation, *no*-negation and *not*-negation, in spoken and written contemporary English. The two types of negation may be exemplified by sentences (1) – (4):

(1) He did *not* see anything

(2) He saw *nothing*

(3) He did *not* know any Scandinavians

(4) He knew *no* Scandinavians

Variation between the two constructions can only occur when an indefinite expression follows the finite verb.

A comparison of two samples of spoken and written English taken from LLC and LOB shows that *no*-negation is greatly preferred in written language, where it occurs in 63% of all possible environments, in contrast to spoken language, where it occurs only in 29% of all possible cases.

A detailed investigation of the factors conditioning the use of either type of negation in spoken English was carried out by a two-step process, viz interactive computer tagging of the material and subsequent variable rule analysis by means of the VARBRUL 2S program, which eliminates the effects of poor data distribution or correlated factors, establishing the effect of each factor taken individually.

This analysis showed that the most important factors favouring *no*-negation are occurrence of the verb BE in existential constructions, of HAVE meaning 'possess', and parallel constructions of the type *He had friends but no money*. The factors most strongly favouring *not*-negation are contrastive constructions like *it is **not** a symptom of lung cancer # it is a symptom of lung disease*, sentences where the indefinite element occurs in a prepositional phrase, and sentences with complex verb phrases.

Pending a final analysis of the written sample, some preliminary data concerning written English are also given.

The data of the *Survey of English Dialects* computerized: – Problems and applications

Wolfgang Viereck
Bamberg University

The objective of the *Survey of English Dialects* [SED] was to record the oldest kind of vernacular speech as this would best illustrate the historical development of English. Consequently, its questionnaire was mainly addressed to the older generation of farm workers in rural areas. Altogether it comprises 1,326 questions, divided into nine "books". The first eight are thematic in subject matter, concentrating on the farm, farming, animals, nature, the house and housekeeping etc., while the ninth book is concerned with morphology and some syntax. Fieldwork was conducted mainly in the 1950s in a network of 313 localities throughout England, i.e. mainly in villages with a rather stable population of about 500 people. In each locality, the questionnaire was answered once only, usually by more than one informant. The informants' responses were published in narrow phonetic transcription in four regional volumes, each in three parts, between 1962 and 1971.

In the past, the computerization of the published data in the SED was attempted repeatedly, but unfortunately these plans were never carried out. At the University of Bamberg, the computerization of the SED data is now nearing completion. Key and conventions followed in the project have appeared in print (Viereck 1985). The basic decision taken was not to deal with phonetics. Since a quantification of the data and a dictionary were envisaged, phonetic transcriptions had to be transformed into normal orthography. For a number of reasons, the transformation of pronunciations into spellings proved very difficult. With our present state of

knowledge, there are bound to be differences of opinion in this area. Furthermore, quite a number of inconsistencies found in the various SED volumes increased the problems. However, with the computer these inconsistencies could be systematically detected and put right. Since November 1986, the project has been receiving financial support from the German Research Foundation and will be carried out in close cooperation with the "Abteilung für Linguistische Informatik" of the "Forschungsinstitut für deutsche Sprache – Deutscher Sprachatlas" at the University of Marburg/Lahn.

Apart from derivative studies such as the contribution of loanwords, the data bank will be used for two basic projects, one being a linguistic atlas. This will go beyond what is already available both in the presentation of single item maps and, above all, in the interpretation of the data. Here a quantitative analysis of the complete material will be provided, separately for each linguistic level, to show the dialect areas of England. Certain methods to measure linguistic similarity and distance will be used, such as the so-called participation map technique, cluster analysis and dialectometry. The second basic project will be a dictionary. This will be in two parts. Part 1 will be restricted to dialectal English in the narrow sense of the term and Part 2 will consist of an index, listing the complete SED material according to certain domains. An important feature of the dictionary will be the inclusion of distributional maps, especially of those items that occur in several completely different areas. Thus, their distribution will be imprinted on the reader much better than a mere verbal listing.

Reference

Viereck, Wolfgang. 1985. "The data of the *Survey of English Dialects* computerized: Key and conventions". In Wolfgang Viereck (ed.), *Focus on England and Wales*. Amsterdam: John Benjamins, pp. 235-246.

The LINKS Project: Building a semantic database for linguistic applications

Piek Vossen, Marianne den Broeder, Willem Meijs
University of Amsterdam

Natural language users are capable of inferring all kinds of semantic information from words. This information is stored as lexical knowledge in Long Term Memory. The aim of the LINKS Project is to build a database of interrelated meaning characterizations such that a machine can generate comparable inferences.

Meaning characterizations can be regarded as language dependent expressions of relations between concepts. Since we cannot easily get at those concepts and relations directly, or by deduction, it is worthwile trying to collect them empirically. A dictionary on computer tape in effect constitutes a carefully built corpus of meaning descriptions which could be seen as a machine-readable example of such a collection. By exploiting this information we can build the database in a systematic pragmatic way. In a dictionary all concepts and relations are expressed by meaning descriptions in natural language. In order to make this information systematically accessible for a machine, therefore, we have to analyse those descriptions in terms of the words and structures they contain. The Longman Dictionary of Contemporary English is very suitable in this respect because of:

a. the fact that all types of information are kept separate and are therefore easily retrievable,
b. the use of a restricted set of words (a Controlled Vocabulary of 2200 words),
c. the use of a global system of semantic labels.

Method

The meaning of an expression can be described as a function of the structure and the meaning of the parts it contains. Therefore we are developing a typology of frequently used structures and words. This involves:

1. Division of the words of the expression into a syntactic kernel and a remainder part. The structure of the meaning description depends on the syntactic category and the semantic label of the lexical entry.
2. Evaluation of the kernels and structures with regard to the semantic relations that are expressed. Our research thus far has yielded a distinction between two major types of definition:
   a. The syntactic kernel is a hyperonym of the entry: we call such a kernel a

'link'. The entries grouped together by the kernel form a coherent set that fits within the hierarchy of hyponymous relations. A **'link'** implicitly constitutes a relation between the entry and the concept that the **'link'** itself designates.
b. The syntactic kernel is **not** a hyperonym of the entry: we call such a kernel a **'linker'**. A **'linker'** constitutes an explicit relation between the entry and the concept designated by the complement of the kernel.

3. Analysis of the remainder of the expression within the specific fields constituted by the links and the complements of the linkers. The specificity will help resolve ambiguity.

Review

John Sinclair, Patrick Hanks, Gwyneth Fox, Rosamund Moon, Penny Stock (eds). *Collins COBUILD English Language Dictionary*. London & Glasgow: Collins. 1987. Reviewed by **Kay Wikberg,** University of Oslo.

The new *Collins COBUILD English Language Dictionary* is a long-awaited contribution to the monolingual English dictionaries that are particularly suited for learners and teachers of English. A research team headed by Professor John Sinclair at Birmingham University has worked on this dictionary for seven years.

COBUILD is really new in several respects:

1. The vocabulary is based on a large data base of contemporary English for linguistic research. This means that COBUILD illustrates the uses of words with examples from actual texts.
2. COBUILD's style of presentation is simpler than that of other dictionaries. Grammatical information has been moved from the actual entries to a separate column, which also contains valuable information on synonyms, antonyms, and superordinates. Special entries (e.g. ADJ AFTER N, PHRASAL V, V+O+A) explain grammatical terminology.
3. Definitions are given in complete sentences.
4. The common everyday vocabulary is given more attention than in any other dictionary.

What, then, are the advantages and possible disadvantages of the above features?

The use of a corpus ensures that the examples are genuine and that they represent typical uses. They also provide independent information, which may have to do with collocations, phrases or compounds. Only items with highly specific senses (e.g. *anesthesiologist, killer whale*), or words whose senses are clear from the definitions anyway (*doubter, latchkey*) are not supported by examples.

A typical entry in COBUILD contains information on spelling, pronunciation, inflection (all inflected forms are listed), senses, style, and language variety. In addition, there may be notes on related uses of an item belonging to a different word class, such as *background* (N) used adjectivally (*background information*), or a count noun appearing as an uncountable.

Derivations are generally presented in the same paragraph as their base word (e.g. *divine – divinely*), except when a derived word is very common (*bad – badly*) or has a different meaning from its base. Such words are listed separately.

All instances of a given item are presented in the same entry, irrespective of whether they are homonyms or represent different word classes. Thus *bank* (N) is first listed in its institutional sense, then in its related sense as a V, as in

2. If you *bank* with a particular bank, you have an account with that bank.
3. If you *bank* money, you pay it into a bank.

Bank, as in *river bank*, appears in paragraph 6, followed by related verbal uses. Finally, there is a list of phrasal verbs (*bank on, bank up*). Towards the end of entries you also find what the editors call a 'convention', i.e. "an expression which has an established form and meaning, and which can be used by itself as a single utterance" (p. 310). Some examples under *life*:

This is the life, How is life?, Not on your life, What a life.

Although the word class indication in the extra column helps the user to find the explanation, this new feature may be confusing, particularly in long entries. On the other hand, the elaborate grammatical coding which clutters the pages of OALD (1980) and LDCE (1978) has been abandoned, and what remains is simple and should not be beyond the post-intermediate learner's capacity. A new grammar note is V-ERG (ergative verbs), which is a symbol that marks "verbs which are both transitive (V+O) and intransitive (V) in the same meaning" (p. 1620). The explanation under V-ERG shows that the intransitive use also covers 'pseudo-intransitives', as illustrated by *Peas freeze well.*

There are cross-references of several kinds, such as under *horse*,

> The word *horse* is used in the following expressions, which are explained at other places in this dictionary. • *to put the cart before the horse*: see *cart*....

Sometimes you look for a cross-reference in vain, such as under *multiple sclerosis*, where there is no mention of *MS*, found as a separate entry. The same thing is true of *let up* (V), which has no cross-reference to *let-up* (N).

The most interesting feature of COBUILD is probably the definitions or 'explanations', as the editors prefer to call them. The most traditional explanations are those of nouns, e.g.

> *Spillage* or *a spillage* is the spilling of oil into the sea from a ship.

But even this explanation tells you that *spillage* is both COUNT and UNCOUNT.

By contrast, explanations of verbs are often presented as situations:

1. When you *squeeze* something, you press it firmly from all sides, usually with your hands, often so that its shape changes or it becomes smaller.
4. When you *squeeze* through a small space, you manage to get through it, often with great effort.

Thus you can generally infer from the explanations what sort of subject and complementation the verb takes.

An example of an adjective:

> Something that is *ingenious* is very clever, involving new ideas, methods, or equipment.

Unlike OALD and LDCE (1987), this explanation excludes the use of the word applied to human beings.

Compared with other dictionaries, COBUILD's explanations are generally superior and will be of great value for both learners and teachers.

As regards the organization of the entries, I would, however, like to make a critical remark. The editor claims that "Wherever possible the first sense is a common one, and a fairly easy one – usually the sense that most people would expect" (p. xix). Although this practice is probably followed in most cases, there are instances where you start wondering about economy or the ordering of paragraphs. For example, the transitive use of *cough* and the synonymous *cough up* a few lines further down the page both have full and identical explanations. Moreover, *cough up* (= 'pay up') is given as sense nr. 1, where the more concrete sense would have appeared more natural. *Cough* and *cough up* can be compared with the treatment of *brighten* and *brighten up*, which are collapsed as *brighten (up)* wherever appropriate.

The semantic information in the extra column often serves as a shortcut if you are not looking for the complete explanation. For instance, we are told that *obstreperous = stroppy*, and vice versa. However, there is no indication that *stroppy* is more informal. Examples of superordinate words and related hyponyms are *mixer – dry ginger*, and *cause – determinant.*

One consequence of the full explanations and COBUILD's concern with the common words is that there is less room for rare words than one might expect in a dictionary of this size. Another consequence is that COBUILD is too bulky to carry every day to school or university. It will, however, be very useful to keep on your desk.

From the point of view of lexicographical method, COBUILD repesents a new development in the use of the computer. The research team has had access to a large concordance in machine-readable form and many other sources, in all about 20 million words. A complete account of the project is now available (Sinclair 1987). There is also a handy workbook by Fox & Kirby (1987) which will help teachers and learners to use the dictionary. It will be exciting to see what else will come out of the Collins Birmingham University International Language Database in the future.

References

Fox, Gwyneth & Deborah Kirby 1987. *Learning real English with Collins COBUILD English Language Dictionary.* London & Glasgow: Collins ELT.

Longman Dictionary of Contemporary English. London-Harlow: Longman 1978 (LDCE)

Longman Dictionary of Contemporary English. 2nd ed. London-Harlow: Longman 1987. (LDCE)

Oxford Advanced Learner's Dictionary of Current English. Oxford: Oxford University Press 1980. (OALD)

Sinclair, John M. (ed.) 1987. *Looking up: An account of the COBUILD Project.* London & Glasgow: Collins ELT.

Shorter notices

Margery Fee: Strathy Language Unit

In 1981, businessman J.R. Strathy left the Department of English of Queen's University a bequest to study Canadian English usage. The bequest directs the Unit to produce "an authoritative guide to correct written and oral communication in English within Canada" and to stimulate interest in the correct use of English in Canada. Well aware of the pitfalls inherent in such terms as "authoritative", "correct", and even "Canadian", the Unit's first director, W.C. Lougheed, began to establish the Unit's corpus of English-Canadian writing from magazines, books and newspapers so that the guide and the Unit's subsequent research would have a sound descriptive base. The corpus now consists of two and a half million words, and the Unit holds extensive computer files on general English usage problems. Dr. Lougheed's "Towards a National Archive on a Micro", *Literary and Linguistic Computing*, 2:4, reviews the computing facilities and programs of the Strathy Language Unit in the context of the Unit's Study of Canadian English.

The Unit has *Writings on Canadian English, 1792-1975: An Annotated Bibliography* by Walter S. Avis and A.M. Kinloch (Toronto: Fitzhenry and Whiteside, n.d.) in database form and keeps it up to date. So far we have added 160 items, and will publish an annotated checklist of these shortly as Number 2 of the Unit's Occasional Papers. We gratefully accept notices, copies and offprints of popular or learned articles dealing in whole or part with Canadian English, especially those unlikely to be listed in the major bibliographies.

The Unit would like to offer its services in keeping records on work in progress on Canadian English. Please keep us up to date on your work and that of your students. We would especially like to know of any projects that might be forwarded by the use of our corpus.

Copies of *In Search of the Standard in Canadian English*, ed. W.C. Lougheed (Kingston: Strathy Language Unit, 1986), the proceedings of a conference held at Queen's in October 1985, are still available at $10.00 from the Unit.

Margery Fee,
Director, Strathy Language Unit,
207 Stuart Street, Room 316, Rideau Building,
Queen's University,
Kingston, Ontario
K7L 3N6

[e-mail: Fee@Qucnd; tel. (613) 545-2152]

Gerhard Leitner, Freie Universität Berlin: Research with computer corpora in Germany

In April 1987, a group of German anglicists/linguists held a meeting at the British Council offices in Cologne to discuss the (deplorable) state of the art of research with computer-based corpora of English in German universities and to develop perspectives for the future. As a first practical result of this meeting an inquiry was conducted at English departments and the results published in the second issue of the newsletter, *Computer Corpora des Englischen* (Berlin). The newsletter can be obtained from the present author.

The main results of the inquiry were this:

(1) Corpus research plays a marginal role in *Anglistik* but there are signs of increasing activities, in particular amongst younger linguists, irrespective of their university status.

(2) Interests vary widely but reflect the current main lines of English linguistics in general. In other words, the main emphasis is on British English and, well behind, American English. Special registers, e.g. technical English, dialects or varieties of English outside Britain or America are even more marginal. But there are several departments, like Bamberg, Bayreuth, Berlin, where research in these areas is going on intensively and where new corpora are being set up.

(3) As far as the temporal dimension of research interests are concerned, synchronic studies on current English come first, followed by diachronic studies and, interestingly, the period between the 18th to 20th centuries.

(4) As for linguistic levels, not surprisingly, grammar and lexis outnumber research in morphology and phonology. Some interest in text linguistics, sociolinguistics, literary translation and other areas is noticeable in a number of universities, for instance, Göttingen, Konstanz, Bielefeld.

(5) Most researchers so far are relying on available corpora, such as LOB, LLC and Brown. There is a tremendous interest in the new COBUILD corpus and it would be good if it became available for research. But smaller corpora are being built up in several places. There are, for instance, corpora on educated English (Saarbrücken and Berlin), Australian and Indian English (Berlin), prepositions (Wuppertal), learner language (Marburg), (East) African English (Bayreuth), literature with a view to literary translation (Göttingen), and more will be done in the near future.

If one looks for explanations for the underdeveloped state of research, one, obviously, has to point to such facts as the non-availability of hardware and software at, or near, departments; the difficulty of getting sound advice on the possibilities that computers would open up; problems with purchasing available corpora; insufficient training and similar practical issues.

But it seems equally true to say that there is an inherent conservatism in German anglistics/linguistics that is less evident in Scandinavia, Great Britain and the Netherlands. To what extent this is related to the fact that English Studies has a longer history and tradition here than in those countries would be an interesting research question. But there is little doubt that tradition must play a role: the situation is not much better in the GDR, Austria and Switzerland, even if financial issues may play an important role, than in the FRG.

This impression is reinforced if one compares anglistics with other modern language philologies, for instance Romance Studies, Slavic Studies or Germanic Studies. It would appear that English is not (much) worse off. To give one example, there are very few corpora of German indeed. One of them is the, by now, rather dated LIMAS corpus, available from the *Institut für Phonetik und Kommunikationswissenschaft, Universität Bonn*, another one is archived at the University of Ulm *(Ulmer Datenbank)* and concerns psychotherapeutic data.

In order to get an overall picture of corpus research on English in Germany one would also have to look at departments of information science. There are, at least, two large research projects involving English in Saarbrücken (EUROTRA) and in Berlin (Technische Universität). Both aim to develop automatic grammars usable for translation and other practical purposes. But like much research on information theory and artificial intelligence, they are less corpus-based than theoretical. Linguistic theories such as GPSG play a central role.

As I have indicated above, the situation in anglistics may be improving quickly if government programs that aim to provide hardware and software for university research and teaching, the socalled CIP-program, bear fruit. An increasing number of, mainly younger, researchers are beginning to move into computer linguistics already.

The initiative taken by the group of anglicists/linguists will be pursued and it is hoped that research goals can be formulated that bring together a number of researchers in different departments of English. Co-operation with workers outside Germany will be welcomed and information can be obtained from the writer.

Project APRIL: A progress report on the Leeds Annealing Parser Project

Geoffrey Sampson, Eric Atwell, Robin Haigh
Centre for Computer Analysis of Language and Speech, University of Leeds

Project APRIL, sponsored by the Royal Signals & Radar Establishment, Malvern, has since December 1987 been developing a system which uses the stochastic optimization technique of simulated annealing (see e.g. P.J.M. van Laarhoven & E.H.L. Aarts, *Simulated Annealing: Theory and Applications*, Reidel 1987) to locate the highest-valued parse-tree for an input word-string, within the logical space of all possible labelled trees having the same number of leaf nodes as the string has words. Individual labelled trees are evaluated in terms of statistical similarity of their various local configurations to the configurations in a database of manually-parsed material, as discussed in R.G. Garside *et al.*, eds., *The Computational Analysis of English*, Longman 1987. An optimal analysis for an input is sought by trying many random local changes to an arbitrary initial analysis, and applying a growing bias in favour of positive moves so as to allow high-valued substructures to emerge gradually from the random mutation process in a manner somewhat akin to the Darwinian evolution of species.

A crude pilot version of this system was described in G.R. Sampson, "A stochastic approach to parsing", *Proceedings of COLING '86*, pp. 151-5. This has now been replaced by a relatively sophisticated system, which recognizes virtually the full range of grammatical distinctions among phrase and clause categories included in the parsing scheme evolved for corpus research at Lancaster and Leeds. APRIL now chooses between 113 alternative labels for nonterminal nodes, which means that for a 22-word (average length) input the search space contains ca $5x10^{103}$ points – hence the impossibility of exhaustive search!

Many aspects of the system still await elaboration (the project is due to run till November 1989). In particular, the tree-evaluation function will shortly become much more sophisticated than at present; and, rather than inputting raw words and extracting detailed wordclass information from a dictionary, so far we only input wordclass tags using an extremely coarse (40-member) classification. Nevertheless, results already strike us as extremely promising. We evaluate APRIL's analysis by using a measure of the similarity between an output tree and the target tree over the same string, which works roughly as follows. For each word, consider the chain of node-labels between that word and the root in each of the two trees, and compute the proportion of matching labels occurring in the same order in both chains as a

proportion of all labels in both chains. Then average over words. With respect to our approach this is a conservative assessment method, since it gives greater weight to higher constituents while APRIL is a broadly bottom-up parser. Run over 50 sentences from technical prose and fiction (mean length 22.4 words) at the beginning of February, APRIL produced analyses with a mean score of 75.3; and many of the mistakes are ones we believe we know how to cure. (NB the test sentences did not contribute to the statistics APRIL uses.) Furthermore the results appear to offer some support to our decision to use an annealing schedule which increases processing only linearly with sentence-length – a significant issue for future practical applications such as real-time speech understanding.

Two further research proposals are currently under consideration: one to use a transputer array in order to implement an algorithm we have evolved to permit parse-trees to be optimized concurrently and thus much more quickly; and the second to work on the Gothenburg Corpus (Alvar Ellegård's manually-parsed subset of the Brown Corpus of American English) in order to turn it into an adequate database of statistics permitting APRIL-like parsing techniques to extend to "deep", semantic analysis.

The Lancaster Spoken English Corpus

In September 1984, a joint research project into the automatic assignment of intonation was undertaken at the University of Lancaster in collaboration with the Speech Research Group at IBM UK Scientific Centre. The first aim of the project was to collect samples of natural spoken British English which could be used as a database for analysis and for testing the intonation assignment programs. The result is the Spoken English Corpus (SEC), a machine-readable corpus of approximately 52,000 words of contemporary spoken British English.

Unlike most other corpora currently being used in the computational linguistic field, the SEC exists in various forms. Research into speech synthesis requires some study of the relationship between the orthographic and prosodic representations of speech. The SEC material, therefore, has been transcribed orthographically and prosodically, both these versions being generated independently from an unpunctuated version. A grammatically annotated version has been produced using the CLAWS word-tagging system to allow an analysis of the influence of syntax on prosody. Recordings of speech samples were produced mainly by IBM UK Scientific Centre using high-quality equipment. The tapes are of a standard suitable for instrumental analysis (for example, the extraction of F0).

It is impossible in a corpus of this size to include samples of every style of spoken English; instead, emphasis has been placed on collecting a sizeable sample

of the type of spoken English which is suitable as a model for speech synthesis. Small samples of highly-stylized speech (for example, that used in poetry reading or a sermon) have been included, but will not be used in the initial testing of the intonation assignment programs. The majority of the texts were obtained from the BBC.

The SEC in its various versions should prove most useful to those researching in the speech synthesis or speech recognition fields. It has already proved to be a valuable tool for teaching purposes at the University of Lancaster, providing students with the opportunity for close study of the phonetics of natural spoken English.

The SEC project was supported in 1984-5 by the University of Lancaster Humanities Research Fund and by IBM UK Ltd., and subsequently by IBM UK Ltd. IBM have not only given financial support, but have actively participated in the project. The project team comprised Dr. G. Knowles (University of Lancaster), Dr. P. Alderson (IBM), Dr. B. Williams (IBM) and L. Taylor (University of Lancaster). Prof. G. Leech (University of Lancaster) and Prof. G. Kaye (IBM) initiated the project and maintained an active collaborative role in it. Additional help was provided by A. Bell and N. Campbell (IBM), and S. Elliot, C. Grover and Dr. E. Briscoe (University of Lancaster).

The SEC is available through ICAME in four versions: orthographic transcription, horizontal tagged text, vertical tagged text and prosodic transcription. The audio tapes are not available from ICAME. See further p 85 and the accompanying order forms.

A Swedish TEFL corpus

A corpus of English texts for Swedish secondary schools (gymnasium) is being compiled by Magnus Ljung at the Department of English, Stockholm University. The immediate aim of the work is to evaluate the vocabulary found in a number of representative English textbooks (56 to date) by comparing the frequencies and distribution of the lexical items with the contents of an extensive corpus of modern English, i.e. the corpus compiled for the COBUILD project in Birmingham.

Once the comparative work has been finished, the texts will be made available as a computer-accessible TEFL corpus containing texts and appropriate lists and concordances. The corpus will be accompanied by a publication describing its contents. If funding can be found, plans are also being made to develop the present corpus into a continuously updated TEFL corpus for Sweden.

Texts for WordCruncher

Electronic Text Corporation, home of the text retrieval program known as WordCruncher, has recently announced a number of indexed texts that can be used with WordCruncher. These include the Riverside Shakespeare; ten volumes from the Library of America collection, viz. Franklin, Jefferson, Melville, Twain, Emerson, Thoreau, Hawthorne, Whitman, Henry James, London; a collection of approximately 50 writings on the U.S. constitution (including Federalist Papers), and the King James Bible. Texts in progress include the Oxford Shakespeare, additional volumes from the Library of America, other English bibles, the Hamburg edition of the complete works of Goethe, and the Pfeffer spoken German corpus. Because of agreements with the publishers these texts cannot be sold simply as electronic texts, but can only be used with WordCruncher. A WordCruncher version of the Brown Corpus is available through ICAME (see p 84 in this issue). Requests for additional information as well as suggestions for texts to be made available for the future can be addressed to:

Electronic Text Corporation, 5600 N University Ave Provo, Utah 84604 *or* Randall Jones, Director, Humanities Research Center, Brigham Young University, Provo, Utah 84602

HUMANIST: A presentation

HUMANIST is a Bitnet/NetNorth/EARN discussion group for people who support computing in the humanities. Those who teach, review software, answer questions, give advice, program, write documentation, or otherwise support research and teaching in this area are included. Although HUMANIST is intended to help these people exchange all kinds of information, it is primarily meant for discussion rather than publication or advertisement.

HUMANIST is an activity of the Special Interest Group for Humanities Computing Resources, which is in turn an affiliate of both the Association for Computers and the Humanities (ACH) and the Association for Literary and Linguistic Computing (ALLC). Although participants in HUMANIST are not required to be members of either organization, membership in them is highly recommended.

In general, HUMANISTs are encouraged to ask questions and offer answers, to begin and contribute to discussion, to suggest problems for research, and so forth. One of the specific motivations for establishing HUMANIST was to allow people

involved in this area to form a common idea of the nature of their work, its requirements, and its standards. Institutional recognition is not infrequently inadequate, at least partly because computing in the humanities is an emerging and highly cross-disciplinary field. Its support is significantly different from the support of other kinds of computing, with which it may be confused. It does not fit easily into the established categories of academia and is not well understood by those from whom recognition is sought.

Apart from the general discussion, HUMANIST encourages the formation of a professional identity by maintaining an informal biographical directory of its members. This directory is automatically sent to new members when they join. Supplements are issued whenever warranted by the number of new entries. Members are responsible for keeping their entries updated.

Technically speaking, HUMANIST is a list of names and addresses kept by ListServ software on the IBM 4381 known as UTORONTO. When ListServ receives an ordinary e-mail message addressed to HUMANIST@UTORONTO. BITNET by anyone on the list, it automatically sends a copy to any other person on the list. The Editor screens submissions only to prevent the inadvertent distribution of junk mail. Valid mail is usually passed on to the members within a few hours of submission. Although HUMANIST is managed by software designed for Bitnet/NetNorth/EARN, members can be on any comparable network with access to Bitnet, for example, Janet or Arpanet.

Since the start in May 1987, HUMANIST has grown rapidly and now has members in many countries. Topics taken up have included the teaching of humanities computing, availability of machine-readable texts, archiving of machine-readable texts, new storage media, electronic publishing, encoding standards for machine-readable texts, etc. The volume of HUMANIST mail is quite large, and there is a discussion at the moment (March, 1988) of possible changes in the distribution system.

New members are welcome, provided that they fit the broad guidelines described above. To subscribe, send a message to the Editor giving your name, address, telephone, and a short biographical description of what you do to support computing in the humanities. This description should cover acedemic background and research interests, both in computing and otherwise; the nature of the job you hold and, if relevant, its place in the university.

The Editor of HUMANIST is Dr. Willard McCarty, Centre for Computing in the Humanities, University of Toronto. Direct applications for membership in HUMANIST to MCCARTY@UTOREPAS.BITNET, not to HUMANIST itself.

Encoding standards for machine-readable texts

In November 1987 there was a meeting on Text Encoding Practices at Vassar College, Poughkeepsie, New York, organised by Nancy M. Ide and sponsored by the Association for Computers and the Humanities and the National Endowment for the Humanities. The participants, who represented various organisations and text archives, discussed guidelines for the encoding of texts. The following points summarize the results of the discussion:

1. The guidelines are intended to provide a standard format for data interchange in humanities research.
2. The guidelines are also intended to suggest principles for the encoding of texts in the same format.
3. The guidelines should
   a. define a recommended syntax for the format,
   b. define a metalanguage for the description of text-encoding schemes,
   c. describe the new format and representative existing schemes both in thatmetalanguage and in prose.
4. The guidelines should propose sets of coding conventions suited for various applications.
5. The guidelines should include a minimal set of conventions for encoding new texts in the format.
6. The guidelines are to be drafted by committees on
   a. text documentation
   b. text representation
   c. text interpretation and analysis
   d. metalanguage definition and description of existing and proposed schemes,

   coordinated by a steering committee of representatives of the principal sponsoring organizations.
7. Compatibility with existing standards will be maintained as far as possible.
8. A number of large text archives have agreed in principle to support the guidelines in their function as an interchange format. We encourage funding agencies to support development oi tools to facilitate this interchange.
9. Conversion of existing machine-readable texts to the new format involves the translation of their conventions into the syntax of the new format. No requirements will be made for the addition of information not already coded in the texts.

Some relevant references are:

Barnard, David T., Cheryl A. Fraser, George M. Logan. Towards a markup standard for full-text archives. Paper presented to the Fourteenth International Conference of the Association for Literary and Linguistic Computing, Gothenburg, Sweden, July, 1987.

Fraser, Cheryl A. 1986. An encoding standard for literary documents. Master of Science thesis, Queen's University, Kingston, Ontario, Canada.

Smith, Joan M. The Standard Generalized Markup Language for data base publishing. Paper presented to the 7th International Conference on Computers and the Humanities, Provo, Utah, June 1985.

Program distribution and networking within ICAME

Knut Hofland
Norwegian Computing Centre for the Humanities
Bergen

Program distribution

The Norwegian Computing Centre for the Humanities has set up a service for the collection and redistribution of programs for use in corpus linguistics and related fields. There is no restriction on type of programs, programming language or operating system. The programs will be distributed on floppy disk, tape or via network (see below). Users that have programs they are willing to share with others are requested to contact the Centre in Bergen. A list of available programs can be obtained via net or from the Centre in Bergen. Among the programs available are:

FREQ for wordlists (alphabetical and rank)
KWIC for concordances

ICAME network mailing list

An electronic mailing list of ICAME users has been set up. This list will be used to distribute information from the Centre in Bergen between the issues of the ICAME Journal. If you want to be added to this list, or have information that you want to distribute to the people on this list, please contact the coordinator at the address given below.

ICAME network server

To facilitate the distribution of information and programs, a network server has been set up at the EARN/BITNET node in Bergen. This server can be reached from any network that has a gateway to EARN/BITNET like Uninett, Janet, Arpa, Csnet etc. The server contains information about the material available, some text samples, an ICAME bibliography, programs and documentation, and network addresses. The server can be contacted in two ways:

a) via interactive messages (only EARN/BITNET)

Example from the IBM VM/CMS environment:

TELL FAFSRV AT NOBERGEN help

will give the following answer

*> NAVF/ICAME Bergen 1 Apr 1987 09:14:11
*> Available commands:
*> HELP - Send this information
*> SEND - Send list of available files
*> SEND fn ft - Send specified file
*> QUERY msg - Store msg to server operator
*> You may also send mail to server operator
*> End of NAVF/ICAME Server Help Info Bergen

TELL FAFSRV AT NOBERGEN send icame netaddr

will send you the list of names on the ICAME electronic mailing list

*> NAVF/ICAME Bergen 1 Apr 1987 09:14:53
*> The file ICAME NETADDR has been sent to you
PUN FILE 1574 FROM FAFSRV COPY 001 NOHOLD

b) via mail

The commands to the server are given in the subject line. Only one command is available in each letter at the moment.

Example:

Date: 12 Mar 87 16:45 -0100
From: Stig Johansson <h_johansson%use.uio.uninett@cernvax>
To: fafsrv@nobergen
Subject: send test boo

How to transfer MS-DOS programs

The MS-DOS programs (.COM or .EXE files) are stored as 8-bit bytes. These files can be transferred between some EARN/BITNET sites, but not all and not via gateways to other networks. To test this, request the file TEST COM. Transfer this file from your local host with the Kermit program. Make sure to set the file type to binary with the command SET FILE BINARY to the host Kermit. Run the program on your PC. You should then see all the ASCII characters displayed.

Another way to transfer binary files is to encode the file as a file of 7-bit printable characters, transfer the file, decode the file back to an 8-bit file. Files encoded in this way have the extension .BOO. To test this transfer, request the files FROMBOO PAS and TEST BOO. FROMBOO PAS is a Turbo Pascal program that decodes a .BOO file. Transfer these files to your local PC, this time as text files (default to the host Kermit). Strip off the mail headers if you have requested the files via mail. Compile the FROMBOO.PAS program and run the program. Give the name of the input file TEST. The file TEST.COM will now be generated. Run the program and you should then see all the ASCII characters displayed.

The program that decodes a file is named TOBOO.EXE or TOBOO.BOO. This can be used if you want to transfer programs via net to Bergen.

In the future other decoding and compression techniques may be used.

Server

EARN/BITNET:	FAFSRV@NOBERGEN
JANET:	FAFSRV@EARN.NOBERGEN
ARPA:	FAFSRV%NOBERGEN.BITNET@CUNYVM.CUNY.EDU

Coordinator

EARN/BITNET:	FAFKH@NOBERGEN
JANET:	FAFKH@EARN.NOBERGEN
ARPA:	FAFKH%NOBERGEN.BITNET@CUNYVM.CUNY.EDU

Material available through ICAME

The following material is currently available through the International Computer Archive of Modern English (ICAME):

Brown Corpus, untagged text format I (available on tape or diskette): A revised version of the Brown Corpus with upper- and lower-case letters and other features which reduce the need for special codes and make the material more easily readable. A number of errors found during the tagging of the corpus have been corrected. Typographical information is preserved; the same line division is used as in the original version from Brown University except that words at the end of the line are never divided.

Brown Corpus, untagged text format II (tape or diskette): This version is identical to text format I, but typographical information is reduced and the line division is new.

Brown Corpus, KWIC concordance (tape or microfiche): A complete concordance for all the words in the corpus, including word statistics showing the distribution in text samples and genre categories. The microfiche set includes the complete text of the corpus.

Brown Corpus, WordCruncher version (diskette): This is an indexed version of the Brown Corpus. It can only be used with WordCruncher. See the article by Randall Jones in the *ICAME Journal* 11, pp. 44-47.

LOB Corpus, untagged version, text (tape or diskette): The LOB Corpus is a British English counterpart of the Brown Corpus. It contains approximately a million words of printed text (500 text samples of about 2,000 words). The text of the LOB Corpus is not available on microfiche.

LOB Corpus, untagged version, KWIC concordance (tape or microfiche): A complete concordance for all the words in the corpus. It includes word statistics for both the LOB Corpus and the Brown Corpus, showing the distribution in text samples and genre categories for both corpora.

LOB Corpus, tagged version, horizontal format (tape or diskette): A running text where each word is followed immediately by a word-class tag (number of different tags: 134).

LOB Corpus, tagged version, vertical format (available on tape only): Each word is on a separate line, together with its tag, a reference number, and some additional information (indicating whether the word is part of a heading, a naming expression, a quotation, etc).

LOB Corpus, tagged version, KWIC concordance (tape or microfiche): A complete concordance for all the words in the corpus, sorted by key word and tag. At the beginning of each graphic word there is a frequency survey giving the

following information: (1) total frequency of each tag found with the word, (2) relative frequency of each tag, and (3) absolute and relative frequencies of each tag in the individual text categories.

London-Lund Corpus, text (computer tape or diskette): The London-Lund Corpus contains samples of educated spoken British English, in orthographic transcription with detailed prosodic marking. It consists of 87 'texts', each of some 5,000 running words. More texts are in preparation. The text categories represented are spontaneous conversation, spontaneous commentary, spontaneous and prepared oration, etc.

London-Lund Corpus, KWIC concordance I (computer tape): A complete concordance for the 34 texts representing spontaneous, surreptitiously recorded conversation (text categories 1-3), made available both in computerized and printed form (J. Svartvik and R. Quirk (eds.) *A Corpus of English Conversation*, Lund Studies in English 56, Lund: C.W.K. Gleerup, 1980).

London-Lund Corpus, KWIC concordance II (computer tape): A complete concordance for the remaining 53 texts of the London-Lund Corpus (text categories 4-12).

Melbourne-Surrey Corpus (tape or diskette): 100,000 words of Australian newspaper texts (see the article by Ahmad and Corbett in the *ICAME Journal* 11, pp. 39-43).

Kolhapur Corpus (tape or diskette): A million-word corpus of printed Indian English texts. See the article by S.V. Shastri on pp. 15-26 in this issue.

Lancaster Spoken English Corpus (tape or diskette): A corpus of approximately 52,000 words of contemporary spoken British English. The material is available in orthographic and prosodic transcription and in two versions with grammatical tagging (like those for the LOB Corpus). There is an accompanying manual. See also p. 76 in this issue.

Most of the material has been described in greater detail in previous issues of our journal. Prices and technical specifications are given on the order forms which accompany the journal. *Note that tagged versions of the Brown Corpus cannot be obtained through ICAME. The same applies to audio tapes for the London-Lund Corpus and the Lancaster Spoken English Corpus.*

There are available printed manuals for the LOB Corpus (the original manual and a supplementary manual for the tagged version). Printed manuals for the Brown Corpus cannot be obtained from Bergen. Some information on the London-Lund Corpus is distributed together with copies of the text and the KWIC concordance for the corpus. Users of the London-Lund material are, however, recommended to consult J. Svartvik & R. Quirk, *A Corpus of English Conversation* (see above).

A manual for the Kolhapur Corpus can be ordered from: S.V. Shastri, Department of English, Shivaji University, Vidyanagar, Kolhapur-416006, India.

The price of this manual is US $15 (including airmail charges). Payment should be sent along with the order by cheque or international postal order drawn in favour of The Registrar, Shivaji University, Kolhapur.

Information about ICAME and order forms can now also be obtained from:

Oxford Text Archive, Oxford University Computing Service, 13 Banbury Rd., Oxford OX2 6NN, England

Humanities Research Center, Brigham Young University, 3060 JKHB, Provo, Utah 84602, USA

These centres also assist in distributing material. All order forms are sent to Bergen.

Conditions on the use of ICAME corpus material

The primary purposes of the International Computer Archive of Modern English (ICAME) are:

(a) collecting and distributing information on (i) English language material available for computer processing; and (ii) linguistic research completed or in progress on this material;

(b) compiling an archive of corpora to be located at the University of Bergen, from where copies of the material can be obtained at cost.

The following conditions govern the use of corpus material distributed through ICAME:

1. No copies of corpora, or parts of corpora, are to be distributed under any circumstances without the written permission of ICAME.
2. Print-outs of corpora, or parts thereof, are to be used for bona fide research of a non-profit nature. Holders of copies of corpora may not reproduce any texts, or parts of texts, for any purpose other than scholarly research without getting the written permission of the individual copyright holders, as listed in the manual or record sheet accompanying the corpus in question. (For material where there is no known copyright holder, the person(s) who originally prepared the material in computerized form will be regarded as the copyright holder(s).)
3. Commercial publishers and other non-academic organizations wishing to make use of part or all of a corpus or a print-out thereof must obtain permission from all the individual copyright holders involved.
4. The person(s) who originally prepared the material in computerized form must be acknowledged in every subsequent use of it.

Editorial note

The Editor is grateful for any information or documentation which is relevant to the field of concern of ICAME. Write to: Stig Johansson, Department of English, University of Oslo, P.O. Box 1003, Blindern, N-0315 Oslo 3, Norway.

ICAME Journal is published by the Norwegian Computing Centre
for the Humanities (NAVFs EDB-senter for humanistisk forskning)
Address: Harald Hårfagres gate 31, P.O. Box 53, Universitetet, N-5027 Bergen, Norway.
Telephone: Nat. 05 212954, Int. + 47 5 212954

ISSN 0801-5775