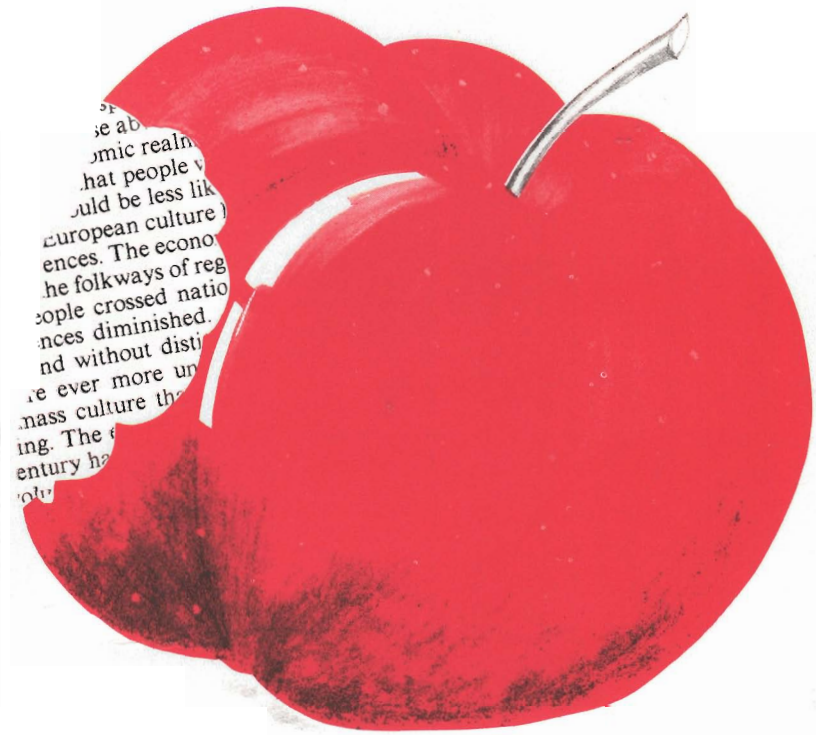
ICAME NEWS

Newsletter of the International Computer Archive of Modern English (ICAME)



NAVF

No. 9

May 1985

CONTENTS

A Survey of Computer-Based English Language Research
S.A. Blackwell 3

Revising CLAWS
B.M. Booth 29

G.E.R.A.S.
Jean-Marie Baïssus 36

Word Frequencies in Indian English: A Preliminary Report
S.V. Shastri 38

Material Available from Bergen 45

The Tagged LOB Corpus 47

Conditions on the Use of ICAME Corpus Material 48

The Humanities Research Center, Brigham Young University 49

Editor: Dr. Stig Johansson, Department of English, University of Oslo, Norway

A SURVEY OF COMPUTER-BASED ENGLISH LANGUAGE RESEARCH

Compiled by S.A. Blackwell
University of Lancaster

1 INTRODUCTION 5
2 RESEARCH AT BROWN UNIVERSITY 5
2.1 Henry Kučera: An Electronic Thesaurus as a Computational Linguistic Problem 5
3 RESEARCH AT LANCASTER UNIVERSITY 6
3.1 Roger Garside: On the Syntactic Analysis of the LOB Corpus .. 6
3.2 Sue Blackwell: Syntax Versus Orthography - Are Ditto-Tags the Solution? 6
3.3 Geoffrey Sampson: Developing the Parsing Scheme 7
3.4 Fanny Leech: Probabilistic Parsing of the LOB Corpus 8
3.5 Eric Atwell: Developing a Text Error-Detection System 9
4 RESEARCH IN OSLO AND BERGEN ON THE LOB CORPUS . 9
4.1 Stig Johansson: How to Tag: That Is the Question Some Problems in Post-Editing the Tagged LOB Corpus 9
5 RESEARCH AT LUND UNIVERSITY 10
5.1 Jan Svartvik: A New Research Project: Text Segmentation for Speech 10
5.2 Mats Eeg-Olofsson: Implementation of the System on a Microcomputer 10
5.3 Anna-Brita Stenström: Discourse Items and Pauses 11
6 RESEARCH AT BIRMINGHAM UNIVERSITY 11
6.1 John Sinclair: Lexicographic Evidence 11
6.2 Antoinette Renouf: A Corpus of Spoken Interaction 12
7 RESEARCH AT HELSINKI UNIVERSITY 13
7.1 Matti Rissanen: A Corpus of English Texts from Old English to Modern English 13
7.2 Ossi Ihalainen: A Corpus of Dialectal British English 14
8 RESEARCH AT NIJMEGEN UNIVERSITY 14
8.1 Jan Aarts: Six Years of Corpus-Related Research in Retrospect 14
8.2 Theo Van Den Heuvel and Nelleke Oostdijk: The LDB Project: An Interim Report 15
9 NEW AREAS OF CORPUS RESEARCH 16
9.1 Mahavir P. Jain (Indian Institute of Technology, New Delhi): Tagging a Corpus of Indian English 16
9.2 Peter Littlechild (Camerino University, Italy): Text Compression for the Microcomputer 16
10 LEXICOLOGY 17
10.1 Jacques Noël and Archie Michiels (Liège University, Belgium): Studying the Definition Language of Two English Dictionaries with the STAIRS Package 17

10.2 Willem Meijs (Amsterdam University, Netherlands): Lexical Organisation from Three Different Angles 18
10.3 Willem Meijs (Amsterdam University): An ASCOT Progress Report 19
11 SEMANTICALLY-RELATED RESEARCH 20
11.1 Nina Devons (Hebrew University of Jerusalem) : FREQSUCON 20
11.2 Göran Kjellmer (University of Gothenburg, Sweden): Some Phraseological Characteristics of English 22
11.3 Chris Paice (Lancaster University): Automatic Abstracting, with Particular Reference to Problems of Endophora 22
12 COMPUTER CORPORA IN EDUCATION AND E.S.T. 23
12.1 Dirk Geens (Free University of Brussels, Belgium): Applications of Language Corpora in Language Education 23
12.2 Archie Michiels And Jacques Noël (University of Liège, Belgium): CALL Software for the Newbrain 24
12.3 Yang Hui-zhong (Jiao Tong University, Shanghai): JDEST Computer Corpus of Texts in English for Science and Technology 24
12.4 Rodolfo Delmonte (University of Trieste, Italy): Complex Noun-Phrases in Scientific English 25
13 SPOKEN VS. WRITTEN ENGLISH 26
13.1 Gunnel Tottie (University of Uppsala, Sweden): Negation in Spoken English 26
13.2 Bengt Altenberg (University of Lund, Sweden): Contrastive Linking in Spoken and Written English 27
13.3 John M. Kirk (Queen's University, Belfast): Scottish Dramascripts and Syntactic Variation 27

1 INTRODUCTION

This paper presents a report on the Conference on Computers in English Language Research, held on 20th-23rd May 1984 at the Low Wood Hotel, Windermere, Cumbria, England.

The conference was organised by the University of Lancaster under the auspices of the British Council and the International Computer Archive of Modern English (ICAME). The Conveners were Professor John Sinclair of Birmingham University and Professor Geoffrey Leech of Lancaster University.

Financial support for the Conference was received from the British Council, the British Academy, IBM (UK) Limited, ICL and USIS.

There were over fifty participants, representing fourteen different countries – perhaps an indication that the importance of computer corpora in linguistic research is becoming more widely recognised. The following is a summary by Sue Blackwell (Lancaster), based on abstracts and her own notes.

2 RESEARCH AT BROWN UNIVERSITY

2.1 Henry Kučera: An Electronic Thesaurus as a Computational Linguistic Problem.

Statistics from a corpus can improve the accuracy of dictionary or thesaurus entries; e.g. most dictionaries describe *spring* first as a verb, although it is five times more likely to be used as a noun.

Lemmatization is crucial to any work on a thesaurus, but introduces some complications: lemmata are often more ambiguous than their inflected forms, and in any case it is often the inflected form that the user wants.

An automatic thesaurus would have to recognise the lemma from the inflected form and generate a correct inflected form from the same or a synonymous lemma. This would require the use of semantic categories: count nouns and mass nouns are being introduced for this purpose. One application is for a spelling corrector - much less trivial than a mere error-finder, since it has to suggest improvements to the user. Such a system would be even more complex, since it would have to be able to accept mis-spellings from the user without generating any mis-spellings of its own.

3 RESEARCH AT LANCASTER UNIVERSITY

3.1 Roger Garside: On the Syntactic Analysis of the LOB Corpus

This presentation described one piece of work being carried out by the UCREL team (Unit for Computer Research on the English Language) at the University of Lancaster, under SERC Research Grant GR/C/47700. This is the syntactic analysis or parsing of the LOB (Lancaster-Oslo/Bergen) Corpus of written British English, using probabilistic techniques similar to those the team had previously used to assign grammatical tags to the words of the LOB Corpus.

RG dealt mainly with the first stage of analysis, by which successive pairs of grammatical tags (for example «noun followed by verb») are looked up in a dictionary which gives information about what grammatical constituents are likely to begin, continue or terminate at such a position; and he discussed how this dictionary is, in fact, structured as a hierarchy of successively more detailed dictionaries containing exceptions to earlier dictionaries, and how the dictionaries are being constructed by the automatic analysis of a bank of manually parsed sentences from the LOB Corpus.

Since this look-up technique will often not be able unambiguously to distinguish the point at which a constituent should be closed, the main task of the second stage of the analysis is to insert constituent closures using a probabilistic technique. (This stage was described in detail in the paper given by Fanny Leech.)

RG concluded by indicating areas where the basic technique of dictionary look-up followed by probabilistic constituent closure needs to be elaborated by procedures to deal with special features of grammar, particularly co-ordination.

3.2 Sue Blackwell: Syntax Versus Orthography - Are Ditto-Tags the Solution

The TAGGIT program devised for the Brown Corpus consisted of two parts: tag assignment and tag selection. This distinction was realised in the LOB suite by two separate programs - WORDTAG and CHAINPROBS, respectively. CHAINPROBS selected tags by a statistical method based on a probability matrix of tag co-occurrence.

To deal with special sequences, the UCREL team inserted a new program, IDIOMTAG, between WORDTAG and CHAINPROBS. IDIOMTAG searches the text for sequences of words whose syntactic role in combination differs from the role of the same words in other contexts: such sequences are termed «idioms».

For example, the phrase *to and fro* would be ambiguously tagged by WORDTAG as follows:

to	TO IN
and	CC
fro	RB

WORDTAG has allowed for *to* being either part of an infinitive or a preposition; however, in this context it is an adverb. IDIOMTAG changes it accordingly:

to	RB
and	CC
fro	RB

A Ditto-Tag marks a special king of «idiom»: it indicates that in this context a *group* of words functions as a single lexical item. Thus the phrase *so as to* forms a single infinitival *to*, represented in the LOB Corpus as:

so	TO
as	TO"
to	TO"

and *according to* functions as a single preposition (IN IN").

SB voiced some misgivings about this strategy: it was difficult to know where to draw the line in defining what constituted an idiom, and some such decisions seemed to have been influenced by semantic factors. Nonetheless, IDIOMTAG had played a significant part in increasing the accuracy of the Tagging Suite to an overall figure of over 96%.

3.3 Geoffrey Sampson: Developing the Parsing Scheme

An automatic parsing system requires a manually defined scheme of parsing to serve as the target or standard against which automatic parsing is assessed. Since early 1983 we have been evolving a parsing scheme for written English and applying it to samples from the LOB Corpus.

Several considerations force us to create our own scheme rather than adopt existing linguistic parsing practices:

1. The scheme deals with written language, whereas traditional linguists' parse-trees do not incorporate, for example, punctuation marks.

2. The scheme has to be robust enough to offer an analysis for everything in real-life written text, even *sentences* such as *Chapter 1: Introduction.*

3. The scheme should ideally be rigorous enough to specify a single *correct answer*, even to parsing issues that traditional linguists might see as trivial: e.g. in *Bloomfield (1935) claims ...*, is *(1935)* part of the noun-phrase which begins *Bloomfield*?

Our scheme is intended to fall within the range of alternative views on grammatical structure held by various linguists, and within that range to adopt alternatives that avoid creating unnecessary difficulties for automatic parsing. The result is a scheme which (1) uses relatively *flat* trees; (2) represents surface

rather than deep structure; and (3) tends to analyse in terms of form rather than function when the two principles conflict.

The process of evolving the scheme has offered striking refutations of the generative view according to which language has a discrete, determinate grammatical structure. The process has rather been akin to the evolution of case law, whereby principles are constantly extended by analogy to unexpected new examples, and frequently-used analogies are in turn elevated into new principles; and the need to impose categorial classification on intrinsically continuous grammatical gradients is highly reminiscent of the work of biological taxonomists, with some researchers proving to be instinctive *splitters* while others are «lumpers».

3.4 Fanny Leech: Probabilistic Parsing of the LOB Corpus

This paper described the methodology of the final stage of the automatic syntactic analysis system summarised in the paper by RG above. The first stage - dictionary look-up plus special heuristic procedures - will pass across a 'partial parse' in the form of a string of constituent markers plus opening and closing brackets. Some closing brackets will still be missing and it is the task of the final stage to close the remaining open constituents and to output the completed parse-tree.

The task is performed by working backwards through the partial parse-tree, selecting each unclosed constituent in turn, building all possible immediate constituent structures for it, and assigning each a probability. The probabilities are obtained from a table of mother-daughter statistics giving the relative likelihood of a particular daughter sequence for a given mother. Structures with probabilities below a certain threshold will be eliminated. Each of the possible structures with its associated probability can then be incorporated at the next level, and the process repeated up to the topmost S level where many structures will be ruled out because no constituent may be left attached as sister to the S. An overall probability for each possible parse can be obtained by multiplying together the probabilities of the particular structures used to complete that parse. It is then possible to select the parse with the highest probability.

The mother-daughter statistics used were derived from a sample of tagged corpus sentences which were parsed manually using the scheme described above by Geoffrey Sampson. Three different methods of statistics-gathering had been tried. The first method was simply to extract a list of daughter sequences for every mother, with an associated relative likelihood figure. The second method was a 'kernelising' technique in which elements not thought to be crucial to the internal structure of a constituent were stripped out of the daughter sequences. Using the third method, the daughter sequences were split into consecutive pairs, statistics for the relative likelihood of each daughter pair for every mother were collected, and the parser calculated the probability of an immediate constituent structure by multiplying together the probabilities of consecutive

pairs of daughters. Because the third method provides more robust statistics for handling new sentences, and because it is more simple to automate than the 'kernelising' technique, it is this method that is being used for the present, though it is recognised it will need further refinement.

FL finished by outlining work that remains to be done, modifying the parser to cope with options within the partial parse, refining the normalisation techniques to take account of tree-depth and constituent-length, and optimising the threshold point for rejection of constituent structures.

3.5 Eric Atwell: Developing a Text Error-Detection System

This project, funded by ICL, is currently in progress at Lancaster. EA talked about the problems of adapting the Lancaster tagging programs to detect errors of typing, spelling and grammar in typed English texts. Often an error results in a non-English character-string: for such cases a large, comprehensive wordlist is needed to check the validity of each text-word. This wordlist could be based on currently available wordlists and machine-readable dictionaries; every root-form will have to be expanded, so that the final wordlist includes all valid inflected and derived forms. Sometimes an error-form coincides with some other well-formed English word: this kind of error can only be detected by measuring the acceptability of the word in its grammatical context, using the contextual-likelihood measure used in the Lancaster tagging programs.

4 RESEARCH IN OSLO AND BERGEN ON THE LOB CORPUS

4.1 Stig Johansson: How to Tag: That Is the Question. Some Problems in Post-Editing the Tagged LOB Corpus.

Three types of indeterminacy confound the task of the post-editor:

1. *Ambiguity*. Sometimes this is deliberate, as in «Melodrama in *A Flat*». Ideally both tags should be retained but the present post-editor was forced to choose between them.

2. *Merger*, as in «He didn't like *her wearing* jeans», where *her* could legitimately be tagged either PP3O (objective personal pronoun) or PP (possessive pronoun). In such cases it is best to give the word its most «normal» tag.

3. *Gradience*. For example, *like* in «feeling very *like* a child» shares characteristics both with adjectives and with prepositions. The corpus linguist must formulate rules or guidelines to divide the gradient into chunks: such rules inevitably involve a degree of arbitrariness.

SJ gave many detailed examples to illustrate the dilemmas he had encountered while post-editing the LOB Corpus.

5 RESEARCH AT LUND UNIVERSITY

5.1 Jan Svartvik: A New Research Project: Text Segmentation for Speech

In 1983 a new project began at Lund entitled «Text Segmentation for Speech» (TESS), with the aim of «describing the rules that govern the natural prosodic segmentation of continuous discourse». The London-Lund Corpus of Spoken English (LLC) is prosodically analyzed in terms of *tone units*, which are defined as «intonationally coherent sequences of stressed and unstressed syllables with at least one peak of prominence called the nucleus».

TESS has two specific goals:

1. To derive information from the LLC concerning the grammatical, pragmatic and prosodic properties of spoken English in general, and of tone units in particular.

2. To apply this information to methods of speech synthesis.

Work being undertaken on the LLC includes grammatical tagging at word, phrase, clause and discourse levels. However, it is not only spoken syllables which are thought worthy of analysis by the TESS team: up to 50% of utterance time consists of pauses, and this is considered significant. One theory is that pauses precede a sudden increase in the flow of information. This and other ideas are currently under investigation.

5.2 Mats Eeg-Olofsson: Implementation of the System on a Microcomputer

The TESS project required a system which was independent, yet had access to the UNIVAC for transfer of LLC text files; which had sufficient storage space for the LLC; and which could be directly linked to a speech synthesizer at a future date. The team therefore opted for three micro-computers plus a 10Mb Winchester disk.

Semi-automatic (i.e. interactive) tagging of the Corpus is currently being carried out on the tone units, using a set of parsing programs. At *word level*, there are about 100 tags based on surface syntactic function, which are assigned to words by means of a high-frequency lexicon and a list of suffixes. In general a tag is assigned to each word, but in some cases two or more words share a single tag (cf. the «ditto-tags» used in tagging the LOB corpus).

At *phrase level*, five sets of cyclical rules operate on the word-tags in each tone unit. Next, *clause level* rules (numbering five at present) operate on the phrases and assign clause tags. However, those items which have been assigned

«discourse tags» at word level are not analysed in terms of phrases or clauses, but are dealt with at their own *discourse level.*

5.3 Anna-Brita Stenström: Discourse Items and Pauses

To date no serious research has been conducted on truly spontaneous speech. As a contribution towards remedying this situation, ABS has been studying nine dialogues and one monologue from the LLC (the former being rather more «spontaneous» than the latter). ABS's present work concentrates on the use of pauses and discourse items in spontaneous speech. Pauses can be silent or voiced (*mmm...*) they vary considerably in duration they may occur in «initial response» or «clause boundary» position. Discourse items (*well, sort of, you know*) can carry varying degrees of stress, and they may be preceded or followed by a pause, depending partly on whether the speaker wishes to invite a response.

Further study of the correlations between factors of this kind can contribute both to our understanding of spoken English and to the quality of future synthesized speech.

6 RESEARCH AT BIRMINGHAM UNIVERSITY

6.1 John Sinclair: Lexicographic Evidence

There are three main ways of gathering descriptive linguistic information:

1. *Received and documented descriptions of language.* Such works are revised infrequently, making it difficult to know when a particular usage has lapsed. Moreover, structural errors are hard to detect and it is difficult to decide that a meaning, usage or form has lapsed. It may be concluded that existing lexicography is only useful at a late stage of research, for checking data which has been assembled independently.

2. *Introspection.* Formal experiments on informants are useful but time-consuming, and should therefore be used only in crucial areas of linguistic research. A far more common practice among linguists, long established by tradition, is that of being one's own informant! However, people are generally very bad reporters of their own usage, and this practice in lexicography gives rise to misleading entries and unnatural examples.

3. *Observation of language in use* is altogether more reliable, and computer corpora have rendered this a viable proposition. Previous observational methods (including OED's) have had an initial stage of selection on intuitive grounds, which the availability of a complete corpus avoids. The picture

presented is very different. Much firmer decisions can be taken about what is and is not extant, and which semantic distinctions are clear cut, and which fuzzy. For example, Collins English Dictionary gives *declinate, declinational* and *declensionally* as valid forms of *decline*, yet the Birmingham Corpus (BC) can provide no instances of these. However, the BC also lacks any example of *declined* as a passive, which we know from other evidence to be extant in English usage. Corpus linguists must beware of mistaking «absence of knowledge» for «knowledge of absence». Therefore, although the corpus work will change and improve lexicography, the evidence must be interpreted in the light of other things known or felt about the language.

6.2 Antoinette Renouf: A Corpus of Spoken Interaction

The Birmingham Collection of English Text now contains over 18 million words of general English, 7.5 million words of which have been concordanced for the purposes of analysis and are called «the Birmingham Corpus» (BC). There are additional 1-million-word subcorpora of TEFL material and of scientific texts in applied science, biology and economics, which are also serving different research needs.

The BC includes 1.5 million words of spoken English, but these do not represent «spontaneous» speech. In March 1984, therefore, an attempt was made at finding an efficient way of eliciting and recording such data, when a pilot study of guided interaction was set up. Four undergraduates of each sex took part in the experiment; the same students were used to transcribe the data, but it was checked twice in case the students had «corrected» the text. The experiment totalled 8 hours of recording time spread over 3 days, and yielded about 70,000 words.

The students were paired off and given instructions on cards which they were not to show their partners. This was an attempt to introduce an «embarrassment factor» into the conversations. For example, one card read: «Your partner is going to ask you questions about money which you have or spend. Be evasive whenever you don't feel like giving the information asked for.»

Some of the findings were predictable: the relatively frequent occurrence of discourse markers like *right* and *Okay* was to be expected, in view of the heavy structuring of the dialogue. There were also many pauses and false starts. However, some unexpected complications arose. In cases where the students were motivated and able to discuss the topic, for example when recalling their first days at university and discussing the advice they would give to new students, they did so. When they were asked to project beyond their own experience, for example in discussing their future career plans, the typical response was to offer the nearest alternative, in this case by recounting what they had done successfully or happily in past years. Being students, they were predictably relaxed in their attitude to traditionally taboo subjects, and it

proved difficult to introduce an element of embarrassment into their discussions of personal matters such as finance.

It was clear that the nature and use of instructions was also a crucial factor in the recordings. Unclarity of wording led the students to spend time discussing what was expected of them. Long and complex instructions were only partially carried out. This factor, combined with a certain degree of self-consciousness about the experiment, undoubtedly produced skewed frequency statistics, since words and phrases like *write down, pen, card, tape*, etc., recurred repeatedly.

One possible solution to this problem is to record subjects without their knowledge, but this raises serious ethical questions. Some members of the audience were perceptibly shocked when AR reported that data had been obtained by tapping Professor Sinclair's telephone!

There is much work still to be done on these recordings, in terms of identification and description, and in developing criteria for evaluation. Questions as to how this method of elicitation can be improved are already framing themselves: Should the tasks be rehearsed? Should we use «good communicators»? What constitutes a good communicator?

7 RESEARCH AT HELSINKI UNIVERSITY

7.1 Matti Rissanen: A Corpus of English Texts from Old English to Modern English

In the autumn of 1983, a project was launched at the English Department of the University of Helsinki to compile an extensive corpus of English texts for the purpose of diachronic studies, primarily on the syntactic and lexical levels. In March 1984, the Finnish Cultural Foundation awarded a grant to employ a project secretary during 1984.

The Corpus will consist of extracts representing a variety of texts covering the period from the earliest Old English to the early 18th century. At the first stage, it will be divided into a Basic Corpus of approximately one million words and a Supplementary Corpus of unlimited size. The Supplementary Corpus will include materials collected for individual research by scholars affiliated with the project and, initially, its use will be restricted. Later on, parts of the Supplementary Corpus may be added to the Basic Corpus.

The Basic Corpus will be divided into century-long sections to ensure a fairly even chronological distribution of the extracts. Each period (Old, Middle, and early Modern English) will be covered by approximately 300,000 words of prose extracts plus some 50,000 words of verse. It is hoped that permission can be acquired to use existing computer tapes of texts dating from the periods in question.

The most important parameters used in the selection of the extracts will be the date, dialect, possible relationship to Latin or French originals, subject-

matter, style (mainly on the formal/informal basis), and relationship to spoken language. The number and applicability of the parameters varies, of course, from one period to another.

7.2 Ossi Ihalainen: A Corpus of Dialectal British English

In the 1970s a number of graduate students from the English Department of the University of Helsinki carried out interviews with elderly natives of rural Devon, Somerset, East Anglia and Yorkshire. The purpose was not to gain information about any specific linguistic point, but rather to elicit as much natural speech as possible. The interviews are now being transcribed and stored in computer memory.

The machine-readable texts are then converted into the tagged Brown format by using a small program called MakeBrown. Since the researchers are working on a number of limited problems such as verb forms, pronominal case, ellipsis, dislocation, etc., the texts will at this stage be only partially tagged, each text naturally reflecting the research interest of its maker.

Although the final form of the tagging system remains to be worked out, it is clear that, in addition to word class tags of the Brown type, there will be tags to indicate ellipsis, zero relatives, phrase boundaries and dislocated elements, to mention just a few features needed in the study of sentence structure.

The Brown format was chosen because we already have a program (called BrownScan) that searches the Brown corpus for morphological, lexical and syntactic patterns. BrownScan also produces basic statistics on the texts studied, a useful feature for anyone interested in variation and language change.

The information retrieved by BrownScan has been used as input (after a number of transformations) to programs like CLUSTAN to generate hypotheses about natural groups in the data.

8 RESEARCH AT NIJMEGEN UNIVERSITY

8.1 Jan Aarts: Six Years of Corpus-Related Research in Retrospect

Two main factors influence the progress of corpus linguists: *external* developments (the reception given to them by the world at large) and *internal* developments (the results of their own work).

Throughout the 1960s, largely due to Chomsky's influence, linguists held a negative attitude towards corpus analysis. Emphasis was shifted away from the purely descriptive towards psychological applications. In any case, it was generally felt that native speakers were an adequate source of data for descriptive purposes.

Against this hostile background the Computer Corpus Pilot Project (CCPP)

emerged. Students were first asked to tag texts manually for both word-class and constituent boundaries. While the manual tagging was very laborious, an attempt at automatic tagging left numerous mistakes. This dilemma gave rise to TOSCA (TOols for Syntactic Corpus Analysis), an interactive tagging system which could run automatically if the user chose not to intervene.

The original LISP program was eventually abandoned in favour of a Context-Free Grammar (CFG). However, the resulting failure rate was unacceptably high: 10% tagging errors and 50% parsing errors. This was partly due to the drama component of the corpus, which contained some highly unorthodox sentences; but it was also felt that the CFG was inadequate because it was based solely on textbooks and intuition rather than on data obtained from a corpus. A further drawback was its unwieldy size.

The next development was therefore an Extended Affix Grammar (EAG), which is smaller than the CFG. However, two problems remain:

1. How is deep structure to be represented?

2. At present each sentence is treated as a self-contained unit rather than as part of a text. How can the analysis take account of this wider context?

Parts of the EAG are now being tested on the corpus with a view to «feeding back» the results into the system to improve its accuracy.

8.2 Theo Van Den Heuvel and Nelleke Oostdijk: The LDB Project: An Interim Report

The LDB is an offshoot of the TOSCA Project: a general-purpose system for manipulating Linguistic Data-Bases (i.e. corpora), with a variety of facilities, including:

1. Interactive data-retrieval

2. The ability to search for key words

3. Clever graphics for displaying tree structures (assuming that the text has already been parsed), with the nodes picked out in contrasting colours

4. The ability to scroll in all directions through tree structures

After TH had described the main features of the system, NO gave an audio-visual presentation showing it in operation.

9 NEW AREAS OF CORPUS RESEARCH

9.1 Mahavir P. Jain (Indian Institute of Technology, New Delhi): Tagging a Corpus of Indian English

Funds have just been granted for this new corpus. MJ justified the concept of «Indian English» on the grounds that some 22 million Indians speak fluent English and that there is a powerful English press in India. The purpose of the corpus is threefold:

1. Comparison with American and British English. For this reason the texts are all chosen from publications in 1961, and the sampling techniques of LOB and Brown are being used. However, it is proving difficult to obtain samples of humorous Indian fiction and of Indian novels analogous to western novels.

2. Extracting «Indianisms»: phrases which fulfil a cultural need in India but which have no meaning in the rest of the English-speaking world. E.g. *dung wash* (cf. *whitewash*) *school ceremony* (on starting school) *turmeric ceremony* (for brides).

3. Identifying more general linguistic features which distinguish Indian English.

MJ admitted that these aims were ambitious and problematic. How could corpus linguists discriminate between authentic «Indian» expressions and errors made by people learning English as a second language? A further problem in analysing texts culled from newspapers is editorial interference. In one paper, the English editor alters the writings of his Indian sub-editors; while in others, all the staff are Indian. Some correction needs to be made for such discrepancies.

This paper was probably the most controversial of the conference, with animated exchanges among the audience as to whether a corpus linguist's first efforts should be purely descriptive and as free as possible from value-judgements.

9.2 Peter Littlechild (Camerino University, Italy): Text Compression for the Microcomputer

PL is using an Apple III to collect a corpus of «political English». This will consist of a variety of written English texts relating to political subjects, with the exception of Hansard which, although seemingly the most obvious candidate, comprises mainly transcriptions of spoken English and is therefore ineligible.

Even with a 5Mb Winchester disk, the project presents a considerable challenge in terms of storage: PL has been forced to investigate possible ways of compressing his corpus. He realised that, while each character was represented by 8 bits (= 1 byte) and the total number of possible combinations for 8 bits was 256, most standard character-sets used about 80 characters. Thus, the 176 or so «spare» combinations could be used to represent a whole word or phrase which commonly recurred in the corpus, such as *The Prime Minister*.

The volume of text could be reduced by up to 50% by using this means of compression in conjunction with either an n-gram algorithm or an algorithm for efficient cluster generation based on frequency scores. Processing could then take place on the compressed text, with no need for repeated encoding and decoding. The codes could be looked up in a key as required.

Implications of text compression for software would be minimal, since there is as yet no standard linguistic software for micro-computers and any concordance packages, etc, for use on the corpus would have to be purpose-built!

10 LEXICOLOGY

10.1 Jacques Noël and Archie Michiels (Liège University, Belgium): Studying the Definition Language of Two English Dictionaries with the STAIRS Package.

Computations

Two dictionary subfiles - letters T from LDOCE by Longman, and from CCED by Collins - were converted into STAIRS-searchable automatic dictionaries. Searching for occurrences of conjunctions in the definitions yielded the following:

CCED: ETC, 588; OR, 2,884; AND, 879; ESP, 461; USU, 149; BUT, 16

LDOCE: ETC, 506; OR, 1,380; AND, 782; ESP, 326; USU, 89; BUT, 81

There is little overlap between the populations of headwords and of conjoined elements thus identified in CCED and in LDOCE. A KWIC program (by Chr. Delcourt) was then applied to the definitions selected by STAIRS in order to search for commas and semi-colons (possible conjunction equivalents), and also to demonstrate that automatized segmentation of conjoined strings can help considerably towards disambiguating (segmenting *and* cross-referencing) definition language.

Interpretation

1. *Linguistic Relationships.* Typically, OR, ETC, and semi-colon, with the comma(s) preceding, serve to relate various equivalent strings; ESP and USU

to relate superordinates and illustrative subordinates; AND, to state necessary conditions for the equivalence between definition and term defined. Cp. also the *strong* (AND) and the *weak assertive force* of OR (cp. ESP, USU, ETC), which may co-occur with CAN, OFTEN, POSSIBLY («hedges»).

2. *Referential*, i.e. encyclopaedic factors also account for various uses of conjunctions, as in *Put two and two together*, and with asymmetric AND, to conjoin processes, etc.

3. *Lexicographical Factors*. Conflated definitions and examples account for many uses, and some overuses of OR; cp. definitions of TYPE and TRANSFER in LDOCE and in CCED.

10.2 Willem Meijs (Amsterdam University, Netherlands): Lexical Organisation from Three Different Angles.

The question of lexical organization - the systematic storage and retrieval of lexical information - is presented from the three angles of theoretical linguistics, experimental psychology, and computer data bases. The central issue raised for each of these spheres is how morphologically complex items - both existing and novel ones – can best be accounted for. WM suggests that his findings in the three spheres can be explained by reference to one and the same model of the lexicon.

The theoretical model advanced (based on Meijs 1975, 1981) is a «full-entry model» (as proposed by Jackendoff 1975), with all existing complex lexical items (CLIs) listed in full in the «Item-Familiar Lexicon», and all potential novel CLIs derivable, via word-formation rules, in the «Type-Familiar Lexicon».

The experimental evidence derives from a number of experiments (reported on in Meijs 1984 and Long-Cornelissen 1984) involving Lexical Decision Tasks (LDTs). It turns out that existing CLIs are retrieved just as fast as existing simplex words (around 700 msecs), while potential novel CLIs like *knittable, longness, unsweet, haired, payer* take an average of 150 msecs longer to process. At the same time the relatively high YES-answer rates for such items (17, 35, and 43 per cent for specified subgroups) shows that these are indeed felt to be «possible words» quite different from such «standard non-words» as *charp, greep, wug* etc.

The issue is then brought to bear on the principles for the construction of a lexical data-base as envisaged in the ASCOT project (cf. progress report in this issue). WM suggests that the ASCOT-lexicon should be constructed along the lines of the theoretical and mental lexicon models, with full entries for all existing items, both simplex and complex, and a morphological component which can relate unlisted complex forms to their (listed) bases and assign (sub)category codes in the process. This goes for flexion-forms like *adds,*

switched, meandering etc., but also for possible regular derivatives like *non-smilers, non-recodable, pseudo-pervert* etc. Interestingly, the «L-tree» storage-system, which is to be adopted for the ASCOT lexicon, shows some remarkable parallels in terms of relative retrieval-time for various kinds of words with the experimental findings: thus frequent words are normally accessed faster than less frequent ones, standard non-words are identified fairly quickly, etc. It is to be expected that, as in the experimental case, access to regular novel CLIs will take somewhat longer, due to the extra processing time.

The full text of the talk will appear in JALLC, 1985, first issue.

References

Jackendoff, R. (1975), «Semantic and Morphological Regularities in the Lexicon», *Language* 51:639-671.

Long-Cornelissen, M. (1984), *Possible Words and Lexical Access*, Graduate thesis, English Department, Amsterdam University.

Meijs, W. (1975), *Compound Adjectives in English and the Ideal Speaker-Listener*, North-Holland, Amsterdam/Oxford/New York.

Meijs, W. (1981), «Synthetische Composita: Voer voor Morfologen», *Spektator* 10:250-291.

Meijs, W. (1984), «Morphological Meaning and the Structure of the Mental Lexicon», to appear in the Proceedings of the Colloquium on Meaning and the Lexicon held in Cleves, 1983 (probably Foris, Dordrecht).

10.3 Willem Meijs (Amsterdam University): An ASCOT Progress Report.

General Outline

The ASCOT project aims at the development of a lexical database and analysing system, which together can provide the coding of words in uncoded corpora. The information contained in the ASCOT lexicon will be extracted from the computer-tape version of an existing English dictionary (either the LDOCE or the OALD). ASCOT will restrict itself to word-level analysis, without looking at the context of the words in question. We intend to give a full analysis of any possible word, presenting all information which is incorporated in, and can be formally extracted from, the OALD or LDOCE. Possible ambiguities must be dealt with by the syntactic analysing system that will use the output of the ASCOT analysing system. As a part of the word level analysis, ASCOT will provide a morphological component, which consists of two parts:

1. A component which can recognize inflected forms and relate them to lemmata incorporated in the lexicon.

2. A derivational component, which can provide correct tags for derived words which are not in the lexicon, but which have been created by productive morphological rules.

Eventually, a syntactic analysing system like e.g. TOSCA (which is being developed at the University of Nijmegen, Holland), in combination with the ASCOT system, should be able to provide a fully automatic analysis of English sentences, resulting in the assignment of both word class tags and grammatical (i.e. functional) tags to words in uncoded corpora.

Progress

In the six months the project has now been under way, a lot of preparatory and exploratory work has been done. Major aspects of linguistic research have been a detailed comparison of the two dictionaries, and the development of an outline of the form and contents of the information contained in the entries of the ASCOT lexicon. Also, a beginning has been made with the development of the component dealing with inflection (which is intended to take the form of a grammar). As far as the computer-side is concerned, it has been explored how information on computer tape can be disambiguated and extracted, and how the L-tree system (developed at the Amsterdam Arts Faculty Computer Department) can be implemented, so that we can have on-line access to the ASCOT lexicon with a minimum of disk-accesses.

11 SEMANTICALLY-RELATED RESEARCH

11.1 Nina Devons (Hebrew University Of Jerusalem) : FREQSUCON

Background

Two areas of lexical statistics have hitherto been paid scant attention: (a) where the unit is word meaning rather than graphic word form, and (b) where the count relates to occurrences in particular types of text, subcategories of the language rather than in a conglomerate cross-section of all types. The FREQ-SUCON represents a TEFL-oriented attempt to redress the balance.

This project, currently in progress at the Hebrew University of Jerusalem, aims to build a FREQuency, Sense-discrimination, Usage-indication, CONtex-tual dictionary of 300 common multi-meaning English words. The research is based on the Brown Corpus and the Word Frequency List derived from it.

The Lexical Unit

Each of the lexical entries analyzed in FREQSUCON consists of the headword, occurring variant forms, and transparent elements in hyphenated words and compounds.

Examples are:

case	cent	class
cases	cents	classes
case-history	five-cent	classmate
suitcase	percent	middle-class

Sense Discrimination

The semantic breakdown of each lexical entry is based on contextual disambiguation of the occurrences in the Corpus, supported by evidence from one or more of a number of listed dictionaries.

Descriptors - not intended as definitions - are used to identify the different senses distinguished. Meanings are, however, only distinguished when they relate to different tags. Thus, no attempt is made to discriminate between *appear* in the sense of «come» and *appear* in the sense of «seem», since both would be tagged VB.

Usage Indication

This refers primarily to an indication of the type of language text in which the word occurs. Subfrequencies of the lexical unit as a whole and of each of the senses discriminated are shown over five major groupings of text, «Domains of Language Use», which are derived from the 15 genres of the Corpus, as follows:

1. Press (categories A, B and C)

2. Literary non-fiction, including religious writings (D and G)

3. Popular non-fiction, including humour (E,F and R)

4. Learned non-fiction (H and J)

5. Fiction (K,L,M,N and P)

Lexical Entries Chosen

Because of the risk of skewed statistics, simple frequency counts were considered unreliable. Accordingly, provisional ad hoc criteria for the interpretation of «common words», avoiding frequency as the determining factor, were decided upon. Candidates for inclusion are words which occur in all five of the Domains of Language Use, and in at least 70/69 of the 500 sample passages. However, to reduce the list to manageable proportions, some eligible words have been excluded. These are high frequency function words (e.g. *to, for, that*) and polysemous (non-homographic) verbs. Hence *appear* and *get* are excluded, while *set* (v. and n.) is included. This leaves an estimated 300 «common» words.

Layout

Each lexical entry consists of two sections:

1. *Statistical Analysis*. The first part of this lists the headword with its variant and compound forms, with frequency and distribution data for each form, plus a calculated composite rating. The second part comprises a Statistical Table giving a breakdown of sense variety and distribution as described above.

2. *Contextual Listing*. This contains lists of KWIC citations from the Corpus, one for each of the senses discriminated.

Applications

In addition to the original purpose (TEFL), the findings of the project might have useful application in lexical contrastive analysis studies, and also in the compilation of bilingual dictionaries.

11.2 Göran Kjellmer (University of Gothenburg, Sweden): Some Phraseological Characteristics of English

Collocations are defined as «recurring sequences of words which are grammatically well-formed». Collocations have as much claim as lexical words to be included in a dictionary; systematic presentation of collocations can teach us much about the language.

For example, the «constructional tendency» rating is very low for adverbs but high for nouns. This suggests that native speakers store adverbs in isolation but nouns in construction.

A study of *semantic* trends in collocation can also prove illuminating. GK analysed the occurrences in the Brown Corpus of adjectives denoting nationality: he found that nouns associated with *American* or *English* were primarily cultural in reference, while those following *Soviet* were mainly political or military. Collocation analysis may well yield interesting results in areas other than linguistics!

11.3 Chris Paice (Lancaster University): Automatic Abstracting, with Particular Reference to Problems of Endophora

A simplistic approach to the production of abstracts of technical documents by computer is to select sentences from the text which are deemed to be particularly useful in indicating general subject matter (see Paice in R.N. Oddy, ed., *Research and Developement in Information Retrieval*, London: Butterworths 1981). A concatenation of separate sentences is however disjointed, and

contains many «loose ends». An aim of this research is to build self-contained passages, by identifying endophoric references between sentences. When found, such features must either be «neutralised», or the sentences linked into a longer passage.

Most endophoric references are backward («anaphoric»), though forward («cataphoric») references must also be dealt with. Most of the references link adjacent sentences, but longer range links also occur, often involving the article *the*. Some of the latter cases can only be resolved by reference to a thesaurus, namely where a «definite» concept has been introduced indirectly via an associated concept.

This research is concerned with developing a general mechanism for identifying and resolving endophoric references, and with compiling and testing a table of rules for a wide range of specific endophoric features.

12 COMPUTER CORPORA IN EDUCATION AND E.S.T.

12.1 Dirk Geens (Free University of Brussels, Belgium): Applications of Language Corpora in Language Education.

The Leuven Drama Corpus is a collection of 62 British English plays, totalling approximately 1,060,000 words. It is being used in language education as a source for:

1. *Vocabulary*. Frequency lists were used to find common English words. However, in practice all words with a frequency of more than 1 had to be used, which suggested a need for a larger corpus!

2. *Lexicography*. Contextualisation, usage.

3. *Syntax*. The texts were analysed for word class, syntactic function and surface dependency. Tagging yielded rather flat «tree» forms, from which an inventory of elementary syntactic structures could be compiled. Each structure is assigned a numerical code, from which frequency lists may be generated.

In addition, a fully automatic system for Computer Assisted Language Learning (CALL) is now in use. This incorporates exercise generators and «text banks» which provide an automatic selection of passages for comprehension.

Global variables have to be associated with each text, including text type, theme and speaker. This will necessitate deep syntactic analysis, semantic analysis and extra-linguistic information, none of which is available at present; however, the team are currently working on a processor which will provide extra-linguistic information.

12.2 Archie Michiels and Jacques Noël (University of Liège, Belgium): CALL Software for the Newbrain

This presentation had two purposes:

1. To illustrate what can be done with a cheap micro in the field of text manipulation. A suite of programs was described which produces and monitors cloze tests on the basis of word lists and lists of associated categories (in the application described, the word list conjoined the most frequent function words in English and the associated categories were grammatical).

2. To introduce the Expert System approach to CALL. At the time of the conference TAG existed only in the minds of the authors. It has now been developed and is operational. It is designed to work in two modes, production mode and test mode.

   In *Production Mode* the student types in a sentence and the system offers to build the confirmatory tag that can be appended to it. It initiates a dialogue with the user to check the assumption it makes and to elicit the structural information it needs in order to build the tag. It prints all the rules it is in the process of using so that the student can follow step by step the procedure that the sytem applies.

   In *Test Mode*, the program prints a sentence on the screen and invites the student to type in the appropriate confirmatory tag for it. The student's answer is then analyzed on the basis of sentence structure and the main parameters for tag formation (negation, nature and tense of the auxiliary, number of pronoun, etc.). Diagnostic messages are produced until the student gets the answer right.

The presenters believe that the Expert System approach is a step in the direction of more intelligent CALL software, in which the system masters a small and readily formalizable but non-trivial problem, and genuinely uses the roles it knows about for both synthesis and analysis instead of merely comparing the student's response with a range of canned answers (the latter is a self-defeating practice in that the range of wrong answers is truly indeterminate).

12.3 Yang Hui-zhong (Jiao Tong University, Shanghai): The JDEST Computer Corpus of Texts in English for Science and Technology.

It is clear that if the goal of modernizing China is to be achieved, scientists and technologists will need to have access to information and research available only in English. This places great responsibility on those drawing up syllabuses

and preparing teaching materials to ensure that the content of English for Science and Technology courses reflects accurately the language of the various disciplines involved. It was with this in mind that the JDEST Corpus was compiled.

JDEST is a bi-lingual acronym, in which *JD* stands for «Jiaotong Daxue» (the Chinese name for Jiao Tong University, Shanghai) and *EST* stands for «English for Science and Technology». JDEST comprises 2000 texts of about 500 words each – 1 million words in all. The specialised fields were not sampled on a random basis, but were dictated by need and availability. Medical texts, for example, are not represented at all. However, *within* each field the texts are sampled randomly to minimise the possible effects of bias on the part of the researchers.

The ten subjects covered by the corpus are: Computers, Metallurgy, Machine Building, Physics, Electrical Engineering, Civil Engineering, Chemical Engineering, Naval Architecture, Atomic Energy and Aircraft Manufacturing.

Each text is uniquely identified by a code which is 20 bytes in length. This indicates whether the text is written in American, British or another form of English; the date of publication; and the genre from which it is taken (periodical, abstract, textbook, etc.). Most of the texts were published after 1975, but unlike the LOB and Brown corpora they are not all taken from the same year.

The corpus was completed in June 1983, having been typed in manually and checked by three proof-readers. Research based on JDEST is now under way.

12.4 *Rodolfo Delmonte* (University of Trieste, Italy): Complex Noun-Phrases in Scientific English

Complex NPs are a major problem in teaching English as a foreign language, and science students are likely to come across more such phrases than other students.

Examples

[[[[Tom's] sister's] husband's] *mother*]

is left-branching, while its Italian equivalent is right-branching:

[La madre [del marito [della sorella [di Tommaso]]]]

This alone will cause difficulties for some students, due to the non-availability of the head NP until the end of the phrase. In Italian students an unusual temporary storage of information will have to take place, as well as an increase of the memory load for semantic processing. To make matters worse, the NPs encountered in scientific English may well have branches in both

directions! Examples are:

[a [[strongly [[wave-length] dependent]] photocromism]]

[a [[rubber industry] [world output]] survey]

NPs of this kind cannot be disentangled by syntax alone: semantics is also necessary. However, an automatic morpheme concordance based on keymorphs and keyroots could help the students to establish which nouns relate to which.

13 SPOKEN VS. WRITTEN ENGLISH

13.1 Gunnel Tottie (University of Uppsala, Sweden): Negation in Spoken English

In English, two kinds of negation are possible when there is an indefinite expression after the finite verb, *no*-negation as in (1) and *not*-negation as in (2):

(1) He saw *no* people in the street.

(2) He did *not* see any people in the street.

It has been shown previously (Tottie 1983) that in written American English, the variation between the two types of negation is conditioned by such factors as sentence type, verb type, the kind of indefinite expression after the finite verb (NP, adverb, etc.), and the syntactic function of that indefinite expression.

As it has also been claimed (cf. Jespersen 1917) that *no*-negation «yields a more elegant expression» and is therefore preferred in more formal styles, a comparative study of the two kinds of negation in spoken and written English was undertaken. The study was based on relevant negative sentences taken from two standard corpora of British English, the Lancaster-Oslo/Bergen Corpus of written English and the London-Lund Corpus of spoken English, both available on computer tape and therefore susceptible to interactive tagging at a computer terminal.

Several differences in the distribution of the two kinds of negation were found, especially in connection with the use of different kinds of verb phrases, modification of postverbal NPs, and various types of idioms.

References

Jespersen, Otto. (1917), *Negation in English and Other Languages*. Historisk-filosofiske Meddelelser I,5. Copenhagen: De Kgl. Danske Videnskabernes Selskab.

Tottie, Gunnel. (1983)., *Much About Not and Nothing*: A Study of the Variation between Analytic and Synthetic Negation in Contemporary American English. Lund: LWK Gleerup.

13.2 Bengt Altenberg (University of Lund, Sweden): Contrastive Linking in Spoken and Written English

This talk presented some results of a study of contrastive linking in a sample of spontaneous conversation from the London-Lund Corpus and a sample of informative prose from the Lancaster-Oslo/Bergen Corpus. Four types of links were examined: the co-ordinator *but*, antithetic and concessive conjuncts, and concessive subordinators. The use of these links revealed a number of differences in the two samples. Stylistically, the links were distributed along a scale from those exclusively or chiefly confined to speech (*anyway,* [but] then, the conjunct *though*) to those primarily confined to writing (*while, however*). The type-token ratio was considerably higher in the written sample, reflecting a greater concern for variation in writing and a reverse spoken tendency to rely on comparatively few types, especially *but* and *anyway*. The frequent use of these links in the spoken sample was found to be closely associated with various conversational strategies, e.g. the employment of *anyway* as a topic shifter and of *but* in «countering» and «disarming» moves. Other notable differences were the higher frequency of concessive subordination in the written sample (contrasting with the spoken preference for *but* co-ordination) and the stronger right-orientation of subordinate clauses in the spoken sample.

13.3 John M. Kirk (Queen's University, Belfast): Scottish Dramascripts and Syntactic Variation

Much of the work on the syntax of Scottish English, while identifying interesting forms as well as areas of study, has remained impressionistic. With the aid of the Corpus of Scottish English now assembled on computer at the Queen's University of Belfast, the grammar of Scottish English may be described in a systematic way. As the data comprises, for the present, six dramascripts (totalling c. 115,000 words) of plausibly realistic, albeit fictional, vernacular speech, it lends itself not simply to syntactic exploitation, particularly in terms of type and token frequencies and their distributions, but also, as appropriate, to exploitation in terms of semantics, pragmatics and discourse.

In addition, the texts represent different varieties along a continuum of Scottish English, so that traditional dialect Scots options, Standard English options and Non-Standard English options, some of which latter are not confined to Scotland, co-exist among the texts in varying frequencies and densities, with clearly discernible text and, occasionally, character distributions. Thus the corpus not only provides a basis for formal comparison with Standard English (e.g. with J. Coates' description of the auxiliary modal verbs in the LOB Corpus or with the increasing number of specific studies based on the LL Corpus): it also discriminates between its own varieties and uses of Scottish English.

In a short preliminary paper it was difficult to do more than suggest some lines of enquiry which could be based on the corpus, including important wider questions such as speech realism, historicity, and geographical distribution. The largest part of the paper drew attention to points of Scottish English syntax on which it is envisaged that on-going work will be able to shed some fresh light.

REVISING CLAWS

B.M.Booth, University of Lancaster

Introduction

The major research tasks being carried out at Lancaster University by the UCREL team are the development of an automatic parsing system, the production of a parsed version of the LOB corpus using this system, and the development of a context-dependent spelling-checker. The work on the parser is funded by an SERC research grant and has been described above in summaries of the papers given at Windermere by Roger Garside, Geoffrey Sampson, and Fanny Leech. The same research grant, GR/C/47700, is also intended to fund further development of the existing automatic tagging system, CLAWS, and this paper describes the work being done concurrently on that aspect of the research.

The existing tagging system was developed for LOB corpus texts and has now been run over the entire corpus with a tagging accuracy rate of between 96% and 97%. The original project has been fully documented elsewhere (Leech et al,1983; Marshall,1983). The purpose of the current enhancement work is to develop the system so that it can be run over other corpora with equivalent or improved accuracy, to remove the need for manual pre-editing, and to modify the output from the system to be even more useful for automatic parsing and possible later semantic analysis. The work is described below in three sections which could roughly be subtitled input, output, and processing. The first section gives brief details of a new simplified input format that can be used for submitting new texts; the second describes the revision of the tag set, and explains how the new output will differ from existing output; the third outlines the changes that have to be made to the automatic tagging programs, and the dictionaries and statistics they use, both to eliminate manual pre-editing and to cater for this modified input and output.

1 Input Format

The "guiding principle" in coding the LOB Corpus was to produce "a faithful representation of the text with as little loss of information as possible" (Johansson et al,1978:7). Some of the coded information is relevant for tagging and is used by the current system; other information is simply ignored or removed before tagging takes place. The coding scheme itself is complex, and for the system to run over new texts these have to be recoded using the scheme, even though they might already be available in a machine-readable format. To avoid this time-consuming and sometimes confusing task, it was decided to

revise the tagging system to accept two different input formats, the original Corpus-coded format, and a much simplified format requiring little if any recoding of texts.

The recognised elements in this new simplified input will either be characters from a standard character set, or special codes from a limited set used to represent characters not available as standard, such as foreign orthographic symbols. The system will assume all standard characters to be used as they would be in normal English orthography and punctuation. For example, - will represent dash, hyphen, or minus, depending on context. The recognised standard characters will be those in the ISO/ASCII 95 character set, omitting _ underline, ~ overline, | vertical bar, \ backslash, ` grave, and ^ circumflex 89 characters in all. But it will be a relatively simple matter to alter the composition of this character set if required. All special codes will have the same format as 'uncoded characters' in the original LOB scheme; that is, one or two digits prefixed by *?. This allows 99 special codes in all. Certain special codes will be fixed in meaning: those appearing in the word list, for instance, or those specifically tested within the tag assignment program. But others will be left unallocated for the user to assign as required.

With the new input format, sentence divisions and abbreviations will not be specifically marked in the text. The system will be amended to allow for this. Also there will be no coding available to pick out typographical shifts, non-English expressions, headings or comments. If this information is required in the output, it would be best to use the existing LOB coding system.

2 Tagset

Because it was considered important to ensure compatibility between Brown and LOB, the tagset used in the original tagging system was based very heavily on the Brown tagset. Only small changes were made in the area of proper nouns, pronouns and adverbs. The current revision involves the relabelling of many existing tags and the abolition of a few others, the introduction of new tag distinctions, and changes in the assignment of words to tags.

Existing tag-labels have some mnemonic significance, and the purpose of relabelling has been both to improve on this mnemonic value, for human convenience, and to ensure that the labels are automatically analysable, to facilitate machine-processing. In the new system, each label can be decomposed into component characters, each signifying some sub-categorisation within the tag-class.

The first character of the tag-label stands for the major grammatical class, for example, N for noun, C for conjunction. Whenever it has been possible to realise the aim of 'analysability' and remain consistent with the existing tagset, this has been done. So J, R, and I have been retained for adjectives, adverbs, and prepositions. But because we now want to group all verbs in the same major

class, all verb tags have been relabelled with a common first character of V.

The second character represents subcategorisation within the major class and where the second character repeats the first, this indicates the most general or unmarked subcategory. All common noun tags will start N*N*, and proper nouns N*P*, as in the existing system, but general adverb tags will now start with R*R*, and locative adverbs R*L*. The old pre-qualifier and pre-quantifier tags, ABL and ABN, and the post-determiner AP tags have now been reclassed, together with the determiner DT tags and the WH-determiner WDT, as subcategories of a major determiner class, and labelled D*A* for after-determiner, D*B* for before-determiner, D*D* for determiner, and D*O* for WH-determiner, as appropriate. Modal verbs and *be*, *have* and *do* verb forms will now have tags commencing V*M*, V*B*, V*H*, and V*D* respectively, while other lexical verbs will be tagged V*V*. In a few cases, tags without subcategories have been left unchanged, even though their labels are not strictly consistent with the new system. E.g. EX, UH, TO.

Further distinctions within the major class, such as verb tense or pronoun person, may be indicated in the next character. E.g. VV*G* for '-ing' participle lexical verb PP*Y* for 2nd person pronoun. The meaning of the medial characters will depend on the major class character they follow. Certain subcategorisations, however, are shared by more than one major class. For instance, a singular/plural distinction can be marked on articles, determiners, nouns, and pronouns. These common subcategories will be represented by consistent characters at the end of the tag label. So *a* would be tagged AT*1*, *these* - DD*2*, *cats* - NN*2*, *it* - PPH*1* etc. When a subcategory character is missing from a label, this indicates that the relevant distinction is unmarked. DD alone will be the tag for determiners unmarked for number, e.g. *any*.

It is thought that these new 'analysable' tags will be especially useful at the parsing stage. At the moment, for parsing purposes, tags making identical predictions about higher grammatical constituents are grouped together into 'cover-tags' and the parsing look-up dictionary contains entries relating to these cover-tags. The new tags would allow the look-up dictionary to be generalised even further. It could contain predictions not only for all verb forms, say V**, but for all '-ing' participle forms, V*G, or for all singular or plural forms, **1, **2.

Where new tag distinctions have been brought in, to improve the linguistic basis of the tagging, this generally provides better discrimination of categories significant in syntactic analysis, and will therefore also be useful to the automatic parser. Within the class of adjectives, for instance, in addition to the old tag JJB, relabelled JB and used for adjectives that only occur in attributive position, a new tag JA has also been introduced for adjectives like *tantamount* that can only appear in predicative position. A new leading co-ordinator tag, LE, will now be used for *both* in *both...and* constructions, *either* in *either...or* etc. This means that the old blanket tags ABX and DTX can be abolished, and a tagging distinction made between "*both* the twins", which will be tagged DB2,

and "*both* Mary and John", tagged LE, which should be useful in parsing the problematic coordinated constructions. Subcategorisation has also been introduced for specific conjunctions and prepositions which introduce particular sorts of clauses or clause-like constructions. E.g. *as* CS*A*, *with* IN*W*.

Certain words which share a semantic identity have also been grouped together, particularly within the N class. In addition to the general common noun category, there are now six further subcategories within NN; for example, NN*S* for nouns of style or title. Weekday and month names have been picked out from the NP class as NP*W* and NP*M* respectively. It is felt that this information, apart from its usefulness for later semantic analysis, will prove helpful even at the parsing stage, in enabling the parser to recognize proper names, for instance, or temporal adverbials like *last Monday*.

3 Automatic Tagging System

3.1 Pre-editing

Before automatic tagging starts there is an initial 'pre-editing' phase which deals with textual anomalies of various kinds causing difficulties for automatic tagging. Pre-editing also reorganises the input texts from so-called 'horizontal' format, the normal left-to-right format of written texts, into 'vertical' format, in which only one syntactic unit appears on each line, with a unique reference number, for ease of processing later.

In the current system both automatic and manual pre-editing takes place. The automatic pre-editor puts each word or punctuation mark on a separate line of output. It gives an enclitic marker to the words *cannot* and *I'm* and forms ending *'ve 're, 'd, 'll* and *n't*, and splits these over two lines. Start of sentence markers are recoded as a line of dashes to distinguish sentence divisions clearly, and certain other amendments are made to the text. Abbreviation markers are recoded, and uppercase letters are changed to lowercase equivalents when they are not considered relevant for tagging, at start of sentence or when whole words appear in uppercase for graphical emphasis. The manual pre-editor then checks the output for exceptions to the automatic rules. Proper-names at start of sentence are given back their initial capital, for instance, and mid-sentence word-initial capitals in headings are changed to lower-case. 3rd person enclitic forms of *be* and *have* are split over two lines, but *'s* genitive forms are left attached.

In the new system, the tasks of the manual pre-editor will be divided between an initial formatting program and the tag assignment program, and any residual errors will simply be left for correction at the post-editing stage. Essentially, the formatter will concern itself with the verticalisation task alone; that is, splitting the text into separate syntactic units for tagging. All decisions about the nature of the tags will be left for tag-assignment and beyond. So the formatting

program will handle the task of splitting all enclitic forms, but the problems of capitalisation and abbreviations, except with relevance to splitting the text, will be handled later. The formatter will no longer change uppercase letters to lowercase nor carry out any tagging.

The formatting program will put each word or punctuation symbol on a separate line, and split enclitics over two lines, as does the current system. Apostrophe s (*'s*) will be treated exactly like the other enclitic forms and the remainder of the tagging procedure will decide whether the form is a genitive or reduced auxiliary verb. *S* apostrophe (*s'*) will be handled in the same way, with the apostrophe put on a separate line and marked as enclitic. However, in the revised input format there is no coding distinction between apostrophe and single quote and this implies a tagging ambiguity later. It may be necessary to try to distinguish these at verticalisation if it proves impossible to resolve the ambiguity later.

In the revised input format sentence-boundaries will no longer be explicitly marked in the text. A sentence-boundary will be recognised, and a line of dashes inserted in the output, after a word with 'sentence-terminating' punctuation, if the next word starts with a capital letter, optionally preceded by quote or bracket. Question-mark, exclamation-mark, and full-stop, again optionally followed by quote or bracket, are all considered to be 'sentence-terminating', except that full-stop alone is excluded if certain rules apply. At the moment these rules are intended to capture titles and/or initials before names, or strings of titles following names, but obviously the rules cannot be infallible. A 'sentence-terminating' full-stop will be placed on a separate output line, otherwise it will be left attached to the preceding word as an abbreviation marker.

3.2 Tag Assignment

The tag assignment program and associated dictionaries have needed considerable modification. The tags against the entries in the wordlist and suffixlist needed amending in line with the new tagset, as did the default tag assignments in the program itself. Apostrophe *s* (*'s*) will now be tagged ambiguously as VBZ, VHZ, and (genitive tag); *'* will be tagged *''* (quote); *'* marked enclitic will be ambiguously tagged *''* and $.

The problems of capitalisation and abbreviation will also now have to be handled at this stage. In the current system, manual post-editing ensures that only proper-names and words habitually written with initial capitals are left capitalised in the text. The wordlist contains both lowercase entries and words with initial capitals, so the same word may appear twice with different sets of tags; e.g. *August* NP, *august* JJ. At tag assignment, a word is looked up in the form in which it occurs in the text, and if found receives the appropriate set of tags. If not in the wordlist, a special suffixlist applies to words with initial capitals, which assigns either or both of the tags NNP or JNP, noun or

adjective with word-initial capital. If all else fails, capitalised words are tagged NP, proper name, by default.

Under the new system all capital letters will be left unchanged in the text, whether at start of sentence or mid-sentence, and the same default value can no longer be assumed. A capitalised word will need the usual lowercase default tags for noun, verb and adjective in addition to the proper name tag, and the relative probabilities of these will depend on whether the word occurs at the beginning or middle of a sentence. Entries in the wordlist will now be entirely in lowercase, but may have two sorts of tags attached, a 'lowercase' set, and an 'uppercase' set which will be distinguished with a colon, and which will only be relevant if a word starts with a capital; e.g. *august* JJ NPM1:. For some words, one or the other of these sets might be empty; e.g. *enquiry* NN1, *england* NPL:. Similarly, the suffixes in the suffixlist may have two sorts of tags. The tag assignment program will reduce words to lowercase before looking them up, but will remember whether or not they occurred at start of sentence, with initial capital, or all in capitals. These three factors will determine which tags will be assigned, and with what probability weightings. In addition, capitalised words will always receive a proper name tag with some degree of probability.

The same sort of approach will be used for handling abbreviations. In the existing system abbreviations are explicitly marked by a special coding device both in the text and in the wordlist. But with the revised input format the system will not always be certain whether a particular form is abbreviatory or not, for instance *in* occurring at the end of a sentence. Tags appearing against entries in the wordlist will be marked with a full-stop if they are only relevant to abbreviatory forms e.g. *approx* .RR, *in* II RL .NNU. When looking up a word in the wordlist, the tag assignment program will remember whether or not the word occurred with a final full-stop, or at the end of a sentence, and these factors will determine whether only abbreviation tags, or only word tags, or both sets of tags should apply.

3.3 Tag Selection

The main task associated with this last phase of the automatic tagging procedure is the derivation of an updated tag-pair transition probability matrix. Two approaches are possible. We could retag a sample from the corpus in accordance with the revised tagset, as far as possible automatically but with some post-editing, and we could then derive a matrix from this retagged sample. Alternatively, we could derive a matrix from the tagged and post-edited corpus, and then adjust the matrix to take account of differences between the new and old tagsets. The latter approach is the more likely, as that was the method used when the current matrix was derived from the tagged Brown corpus.

It is possible that some of the finer distinctions introduced in the revised tagset might be relevant for parsing but not for tag selection. The transition matrix need not contain statistics for such subcategorisations, but only for the

more general categories. Tag assignment could assign the subcategory symbols but they would be ignored in tag selection.

Another enhancement we should like to incorporate is the use of genre-specific statistics. We intend to derive a separate probability matrix for each of the categories and allow the option of using these for tag selection, to see if this will improve the tagging accuracy yet further.

Conclusion

The overall aim behind the current enhancement work has been to improve the tagging system both for human use and for machine-processing. The tagset itself will be easier for a new user to understand and memorise. The revised input format and the elimination of the manual pre-editing phase will make it simpler to run the system over new texts. The improved linguistic basis of the output should prove helpful for later syntactic and semantic analysis.

References

JOHANSSON, S., LEECH, G., and GOODLUCK, H., (1978), *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*, Department of English, University of Oslo.

LEECH, G., GARSIDE, R., and ATWELL, E. (1983), 'The Automatic Tagging of the LOB Corpus', *ICAME News* 7, 13-33.

MARSHALL, I. (1983). 'Choice of Grammatical Word-Class without Global Syntactic Analysis for Tagging Words in the LOB Corpus', *Computers and the Humanities*, XV11, 139-150.

G.E.R.A.S.

Jean-Marie Baïssus, University of Montpellier

GERAS (Groupe d'Etudes et de Recherches sur l'Anglais de Spécialité) is a French national association of approximately 100 members who are interested in ESP teaching and research. The teaching takes place essentially in non-Arts Universities (Sciences, Law, Medicine) and Instituts Universitaires de Technologie.

For research purposes four local teams were set up in 1982 with the active support of the Ministère de l'Education Nationale and the British Council. It must be emphasized that the members of the research teams are essentially self-trained linguists and are currently 'learning about computers'. Their academic interests are not centered on computer textual analysis but they all manage to find the time and the energy necessary to progress in this area. Their intention is to lead the way for full-time specialists in a not too distant future.

The following tabulated data give the essential characteristics of the four teams in alphabetical order:

Name & Address of Team Leader	Research Projects	Hardware
Prof. Michel PERRIN Vice-Pres. of GERAS Univ. of Bordeaux II 3, Pl. de la Victoire 33 076 Bordeaux CEDEX Dept. de Langues Vivantes Pratiques Tel. (56)91.34.24 Ext. 623.	Analysis of scientific discourse leading to computer assisted production of scientific articles (medical). Production of CALL programs for medical students.	1 Apple LISA 1 MacIntosh + ancillaries.
Prof. J.-M. Baïssus Vice-Pres. of GERAS Univ. of Montpellier III I.L.S.E.R. Rte de Mende, BP5043 34032 Montpellier CEDEX Tel. (67)63.44.91	Computer analysis of surface markers Analysis of weather forecasts leading to automatic production of weather bulletins. Specialised corpus processing.	1 ACT VICTOR Sirius + ancillaries including Winchester disk drive.
Prof. Michèle RIVAS Vice-Pres. of GERAS C.E.R.L.A.C.A. Univ. of Paris IX (Dauphine) Pl. de Lattre de Tassingny 75 775 Paris CEDEX 16 Tel. (1)505.14.10 Ext. 23.37	CALL in Economics Production of teaching programs. Computer assisted corpus analysis of international conferences and meetings.	GOUPIL 3 10 Megabyte Winchester disk drive

Prof. F. COSTA President of GERAS Univ. of Toulouse I Institut des Langues et Civ. Etrangères, Pl. Anatole France, 31 042 Toulouse CEDEX Tel. (61)23.11.45 Ext.391	Lexical Data bank in Law and International Trade. Computer assisted analysis of ill-formed texts by students for corrective and CALL purposes.	SIL Z3 + ancillaries.

Whereas Bordeaux and Toulouse work informally in the course of their day to day teaching activities, Montpellier and Paris-Dauphine have set up officially recognized research teams affiliated to GERAS. GERAS itself is affiliated to the SAES (Société des Anglicistes de l'Enseignement Supérieur).

In Montpellier III, ILSER (Institut de la Langue de Spécialité Enseignement et Recherche) was created in 1983 thanks to the drive and iniative of Tony LATTES (who is also the national Secretary of GERAS). It is a local association of university teachers. They are not all ESP specialists which explains why two kinds of seminars are held through the year: pedagogy of ESP and theoretical research on specialised discourse/ communication. No. 3 of *Les Cahiers de l'ILSER,* their bi-yearly periodical, has just been published. The computer research programs set up with the active cooperation of Ian SEDWELL (formerly member of the COBUILD Birmingham team) are now operational as described above.

In Paris-Dauphine the C.E.R.L.A.C.A. Research Group is currently setting up an ESP computer corpus of oral English. The Dauphine corpus will emphasize the type of (formal) spoken English used by native speakers in international congresses and meetings, negotiation sessions and plenary sessions of international organizations. The data used will come from:

1. Official transcripts of F.A.O., UNESCO and European Parliament meetings.
2. Their own transcriptions of other data drawn from recorded committee meetings of other similar organizations.

The other research area concerns the creation of CALL software based on economic texts already used in English courses at Dauphine. CERLACA has bought the programs written by Professor YANG of Shanghai Jiao Tong University designed to produce word frequency lists and concordances of words in context. Software with an emphasis on teaching procedures of reading comprehension are being developed through two specific authoring langues, P.I.G.E. and Arlequin.

WORD FREQUENCIES IN INDIAN ENGLISH: A PRELIMINARY REPORT*

S.V. Shastri, Shivaji University, Kolhapur

1. Introduction

This is a preliminary report of work in progress, whose object is to produce a comprehensive study of word frequencies in Indian English. This will be a sequel to the already available study of word frequencies in British and American English (Hofland and Johansson 1982). It is expected to reveal some broad features of Indian English and serve as a starting point for more detailed studies of Indian English.

Indian English (IE) has now come to be recognized as a variety of English in its own right. There was a time when IE was a derogatory term used to indicate a 'sub-standard' variety and all sorts of labels were employed: 'butler English', 'babu English', 'chee chee English', and so on. What is more, even serious-minded linguists argued that there was no such thing as IE on elaborate theoretical grounds. They argued that there could not possibly be *a* variety of IE; there might be as many Indian Englishes as there are IE speech communities, regional, social, and so on. Some even held that these could at best be pidgins and at worst a mass of chaotic idiolects.

Happily that phase has now passed. Scholars have not only sharpened their methods of analysis, but are finding evidence to show that IE is *one* national variety of English sharing a very large part of the common core of English. This has been found to be specially true of the written variety (Kachru 1983).

2. Material and method

The projected study is based on a million-word corpus of IE English comparable with the existing British (LOB) and American (Brown) Corpora; see Shastri (1980). In the last twenty years after the building of the first computer corpus, i.e. the Brown Corpus, it has become amply clear that such a corpus is good enough for all sorts of lexical, syntactic and semantic studies. Many major studies of aspects of British (BE) and American (AE) English have appeared. It has also become possible to compare the two varieties in a more systematic and objective way, after the building of the LOB Corpus of British English. It is now hoped that the IE Corpus will make it possible for us to describe IE as compared with BE and AE.

*Paper presented at the 7th AILA World Congress in Brussels, Belgium, August 1984.

A detailed description of the rationale behind the building of these corpora, the sampling techniques, the coding systems etc is beyond the scope of this paper. However, it should be stated that the corpora consist of 500 texts of 2000 running words distributed over 15 genres of writing representing different styles. Table No. 1 shows the break-up of the samples.

TABLE 1: The basic composition of the American, British and the projected Indian Corpus

Text Categories	*Number 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 & hobbies	36	38	38
F.Popular lore	48	44	44
G.Belles lettres	75	77	70
H.Miscellaneous (gov. 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	
L.Mystery & detective fiction	24	24	
M.Science fiction	6	6	126
N.Adventure & western fiction	29	29	
P.Romance & love story	29	29	
R.Humour	9	9	
Total	500	500	500

The coding system used in the IE Corpus differs slightly from those used in the LOB and the Brown Corpora. The codes are more of an interpretive than a descriptive nature. Details of graphic features have largely been ignored and only their significance has been preserved. For example *type shifts* have been coded almost entirely in interpretive terms. Two important codes, one for Indian words and the other for Indian expressions, have been devised. All in all, we have attempted to record as many apparent features of IE as possible at the stage of manual pre-editing so that, by suitable programming devices, we may later be able to do automatic analysis by computer.

Software for word-frequency counts has been developed. One programme gives three output files: (1) frequencies of English words, (2) frequencies of

Indian words, (3) frequencies of expressions - both Indian and foreign. Another programme tabulates the data showing the frequencies across categories and the distribution among the text samples.

The observations made in this paper are based on the analysis of word frequencies in the Press materials, which account for about one fifth of the corpus. The details of the composition of the samples in this category are given in Table 2.

TABLE 2 : Composition of the Press materials

	National Daily/Weekly		Regional Daily/Weekly		Total
Press Reportage Cat.A.					
Political	6	2	5	-	13
Sports	2	2	2	1	7
Spot news	4	1	4	1	10
Society/Culture	3	3	2	2	10
Financial	2	1	1	-	4
Editorial Cat.B.					
Institutional editorial	6	3	6	1	16
Personal editorial	2	1	-	1	4
Letters to the editor	3	1	2	1	7
Press Reviews Cat.C.	5	9	2	1	17
Total:	33	23	24	8	88

3. Results and discussion

I would like to present, and briefly comment on, two significant sets of statistics resulting from an analysis of the Press materials in the IE corpus as compared with the LOB and Brown Corpora: (1) comparative frequencies of words in the three corpora and (2) a break-up of the kinds of words in the IE Corpus. Table 3 shows the relative ranks of the fifty most frequent words in Brown (American), LOB (British), and the IE Corpus (Press materials only).

It is interesting to note that, even in a relatively small sample of the IE Corpus, the figures are so close. Maybe they will turn out to be even closer when the entire corpus is analysed. But even as it is, 40 out of the first 50 words in the IE rank list occur among the fifty most frequent words in the LOB Corpus; the overlap between the IE Corpus and the Brown Corpus rank list is 39 out of 50. What is more, the absolute frequencies of some of the items in the IE and LOB

TABLE 3 : Table showing the relative ranks of the fifty most frequent words in the BROWN, LOB and Indian English (Press materials only) Corpora.

Word	*IE* rank	frequency*	*LOB* rank	frequency*	*BROWN* rank
the	1	14287	1	11940	1
of	2	7080	2	5969	2
to	3	5029	4	4507	4
and	4	4979	3	4407	3
in	5	4183	6	3820	6
a	6	3561	5	4177	5
is	7	2532	8	2466	8
that	8	1832	7	1905	7
for	9	1567	11	1791	11
it	10	1434	10	1594	12
on	11	1296	16		16
by	12	1262	20		19
be	13	1257	15		17
was	14	1196	9		9
not	15	1144	23		23
as	16	1132	13		14
with	17	1089	14		13
has	18	972	42		44
have	19	970	26		28
this	20	921	22		21
are	21	909	27		24
at	22	800	19		18
but	23	759	24		25
which	24	724	28		31
from	25	722	25		26
an	26	650	34	676	29
his	27	679	18		15
their	28	677	-		-
he	29	661	12		10
been	30	620	37		43
they	31	602	33		30
or	32	578	31		27
had	33	573	21		22
there	34	526	36		38
were	35	519	35		34
who	36	517	50		46
will	37	497	48		47
one	38	415	38		32
all	39	464	-		-
if	40	376	45		50
no	41	357	-		-
more	42	356	-		-
other	43	355	-		-
some	44	351	-		-
only	45	342	-		-
would	46	337	-		-
also	47	336	-		-
so	48	313	46		-
can	49	308	-		-
when	50	289	44		45

*These figures pertain to Press materials only

Corpora (Press materials only) are very close. It may be hazardous at this stage to comment on the possible implications of the differences in the figures pertaining to articles (*the, a*) prepositions (*of, for*) conjunctions *and, that*), and so on. These must be analysed in detail at a later stage. They might tell us something about the nature of IE as reflecting the conceptual, cultural and social framework of the people using it.

As to the entire frequency list, a close look reveals that almost all the 4710 words that occur more than ten times in the LOB Corpus also occur in the IE materials. The 370 that do not occur in the IE materials are more than compensated for by the 580 Indian words and expressions. One of the conclusions we might draw is that the *COMMON CORE* vocabulary is very much the same.

TABLE 4 : A break-up of the graphic words in the IE Press materials

Sr.No.	Strings	Absolute frequency	Percentage
1	Graphic words	14377	70.96
2	Unique names	4213	20.85
3	Indian words and expressions	580	2.87
4	Numerals and numeral compounds	558	2.76
5	Abbreviations	352	1.74
6	Initials	117	0.58
7	Foreign words & expressions	46	0.23
	TOTAL	20203	99.99

A break-up of the 20,203 graphic words in the IE materials is given in Table 4. Obviously the very high percentage of unique names (proper names), is an index of the vastness and diversity of the sub-continent. The extent of code-mixing with Indian items, almost 3%, is perhaps a measure of Indianness of IE. I will come back to this later. If we take a close look at these items, we find that they fall into two broad categories:

(1) Indian items used where (near) equivalents are available in the native language(s), e.g. *acharci* = cook; *azadi* = freedom; *bhook* = hunger; *cutcheri* = concert; Film *otsav* = Film festival; *hartal* = strike; *Jhompadappti* = shanties (slums); *kala* = art; *kisan* = farmer; *malik* = landlord/proprietor/

master; *mandir* = temple; *mela* = fair; *natak* = play; *sammelan* = conference.

The motivation for this seems to be a deliberate attempt at asserting the Indian identity. However, their use as a strategy for mass communication cannot be ruled out. Some of them are stylistically marked or emotionally loaded.

(2) Indian items referring to peculiarly Indian concepts, institutions, flora and fauna. These are by far the largest in number. They may be classified according to register as follows:

A. Religion and Philosophy: *advaita* = non-dualist; *atman* = soul; *dharmic* = religious;

B. Fine Arts: (a) Music: *abhanga*; *alapa*; *bhava*; *karnatik dholki*; *dhrut*; *gharana*; *Hindustani*; *Halagi*; *Khayal*; *lalit*; *matra*; *quawali*; *rasa*; *sarod*; *shuti*; *sitar*; *sutra*; *sarangi*; *sangeet*; *taal*; *thumtri*; *tabla*; *veena*;

(b) Dance and Drama: *Bharatanatyam*; *chawka*; *odissi*; *Kathak(ali)*; *natak*; *natya*;

(c) Folk Art and Mythology: *apsara*; *Bailata*; *gopis*; *Gokul*; *Gokulasthami*; *mohinis*; *lavani nantanki*; *Tamasha*; *Yakshagana*;

C. Administration: *challan*; *gauda*; *daphtary*; *lokpal*; *parishad*; *prabandhak*; *panchayat*; *rajya*; *samiti*; *swaraj*; *Tuluka*; *Tehsil*;

D. Flora & Fauna: *Hangul*; *markher*; *palak*;

E. Society: *arrack*; *bhakri*; *bhoodan*; *bandh*; *bhopa*; *brahmin*; *crore*; *chappals*; *chit*; *Diwali*; *dalit*; *dharna*; *Doordarshan*; *goonda*; *ghat*; *Harijan*; *Holi*; *Hamal*; *jatra*; *jat*; *Khadi*; *Khandsari*; *lakh*; *mahout*; *maidan*; *mela*; *matka*; *peria*; *paisa*; *paan*; *satta*; *sari*; *satyagraha*; *vanaspati*; *shri*; *shrimati*;

It appears that the extent of code-mixing depends on the genre of writing. I have posited elsewhere a cline of Indianness of IE and have suggested that IE lies along a cline with the heavily Indian at one end and the hardly Indian at the other. The heavily Indian writings would belong to Religion, Philosophy, Fine Arts and the hardly Indian to Scientific and Technical Writings. Between the two lie the other genres of writing. I have also said that code-mixing is a *transparent* feature of IE as contrasted with other syntactic, semantic and stylistic features which I have called *opaque* (Shastri forthcoming).

4. Conclusions

From a preliminary analysis of the Press materials it appears, then, that the IE Corpus shares the common-core vocabulary with its British and American counterparts. A peculiar feature of IE is that it has a significant percentage of

Indian words. This *transparent* 'Indianness' is especially characteristic of certain genres of writing, which I have called heavily Indian, e.g. Religion and Fine Arts. Other genres, e.g. Scientific and Technological Writings, are hardly Indian and their 'Indianness' is *opaque*. Detailed investigation of opaque features of IE is the most interesting and challenging task facing researchers.

REFERENCES

Hofland, Knut and Stig Johansson (1982), *Word Frequencies in British and American English*. Bergen : The Norwegian Computing Centre for the Humanities. London : Longman.

Kachru, Braj B. (1983), *The Indianization of English. The English Language in India*. Oxford University Press.

Shastri, S.V. (1980), 'A Computer Corpus of Present-Day Indian English', *ICAME News* 4, 9-10.

Shastri, S.V. (forthcoming), Code-Mixing in the Process of Indianization of English.

MATERIAL AVAILABLE FROM BERGEN

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

Brown Corpus, text format I (without grammatical tagging): 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, text format II (without grammatical tagging): This version is identical to text format I, but typographical information is reduced and the line division is new.

Brown Corpus, KWIC concordance (also on 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.

LOB Corpus, text: 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).

LOB Corpus, KWIC concordance (also on 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. The text of the LOB Corpus is not available on microfiche.

LOB Corpus, tagged versions: see p. 47.

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

London-Lund Corpus, KWIC concordance I: A complete concordance for the 34 texts representing spontaneous, surreptitiously recorded conversation (text categories 1-3), made available both in computerised 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: A complete concordance for the remaining 53 texts of the London-Lund Corpus (text categories 4-12).

The material has been described in greater detail in previous issues of *ICAME News*. Prices and technical specifications are given on the order forms which

accompany this newsletter. *Note that the concordances are now also available on higher-density tapes at a lower price.*

A printed manual accompanies tapes of the LOB Corpus. 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 concordances for the corpus. Users of the London-Lund material are, however, recommended to order the recent book by Svartvik *et al., Survey of Spoken English: Report on Research 1975-81*, Lund Studies in English 63, Lund: C.W.K. Gleerup, 1982. The grammatically tagged version of the Brown Corpus can only be obtained from: Henry Kučera, TEXT RESEARCH, 196 Bowen Street, Providence, R.I. 02906, U.S.A. The Syntax Data Corpus, which consists of part of the Brown Corpus, with detailed syntactic tagging, can only be obtained from: Alvar Ellegård, Department of English, University of Gothenburg, Lundgrensgatan 7, S-412 56 Göteborg, Sweden.

THE TAGGED LOB CORPUS

The tagged LOB Corpus is the product of cooperation between researchers at the University of Lancaster, the University of Oslo, and the Norwegian Computing Centre for the Humanities at Bergen. Each word is provided with a tag indicating its grammatical class (number of different tags: 134), assigned through a combination of dictionary look-up, probabilistic tag-selection rules, and manual post-editing. The tagged LOB Corpus exists in two formats:

I: a horizontal format, with a running text where each word is immediately followed by its associated tag;

II: a vertical format, where each word is on a separate line together with its associated tag and a reference.

Analyses of word and tag statistics and a concordance, sorted by word and tag, are in preparation. The price for a magnetic-tape copy of the tagged corpus will be approximately 2,000 Norwegian kr. For more information, tick the appropriate box on the order form enclosed with this issue and return it to the Norwegian Computing Centre for the Humanities.

References:

Leech, Geoffrey, Roger Garside, and Eric Atwell (1983). 'The Automatic Grammatical Tagging of the LOB Corpus', *ICAME News* 7, 13-33.

Marshall, Ian (1983). 'Choice of Grammatical Word-Class without Global Syntactic Analysis: Tagging Words in the LOB Corpus', *Computers and the Humanities* 17, 139-50.

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.

HUMANITIES RESEARCH CENTER

Brigham Young University

The Humanities Research Center at Brigham Young University (Provo, Utah) has agreed to act as a distribution agent for information on ICAME. The following information on the Center was supplied by its Director, Dr. Randall Jones:

The Humanities Research Center (HRC) was established in 1981 to provide research and technical support to the College of Humanities. Its roots go back to the former Language Research Center and the Translation Sciences Institute.

The HRC is made up of three divisions: 1) Humanities Computing, 2) Audio-Visual and Technical Support, and 3) Language Instruction and Testing Research. Each division has a director who is also a member of the Humanities College faculty. There are also other full-time and student HRC employees.

The HRC is fortunate to have one of the finest equipped humanities computer research centers in the world. It owns an IBM 370/138 mainframe computer, a high-speed line printer, several other special printers for foreign language alphabets, a Data General Nova3 minicomputer, a Kurzweil KDEM-II Omnifont scanner, and numerous Apple and IBM microcomputers. Many faculty members in the college use the facilities for their research, especially in linguistic and literary analysis. Several books and numerous articles have resulted from work done at the HRC. BYU is also involved in several joint projects with other universities, including Cornell University, the University of Bonn (Federal Republic of Germany), the University of Bergen (Norway), and the University of Grenoble (France).

The Audio-visual and Technical Support Division boasts one of the finest learning laboratories anywhere. In addition to a 40 carrel Tandberg IS9 console lab, it also has a 32 carrel self-study lab, each carrel equipped with a Tandberg cassette deck and either a Sony color TV monitor or a Singer Caramate. There are also several mobile VCR's, a color VCR camera, a tri-standard VCR for playing video cassettes from other parts of the world, and a projection TV system. In addition there is a microcomputer lab for computer-assisted language instruction and testing. During the past decade BYU has earned a reputation for being a leader in this area.

The HRC is continuing to develop new programs in order to provide better support to the college. The priorities for the near future include assistance in training faculty in the use of the IBM/PC, the installation of a cable to the new university satellite system in order to receive foreign language television broadcasts, and the improvement of word processing and printing of foreign language texts, especially Chinese, Japanese, Korean, Arabic, Hebrew and Greek.

EDITORIAL NOTE

Further ICAME newsletters will appear irregularly and will, for the time being, be distributed free of charge. The Editor is grateful for any information or documentation which is relevant to the field of concern of ICAME.

ICAME NEWS 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, N-5014 Bergen-University, Norway

ISSN 0800-6806