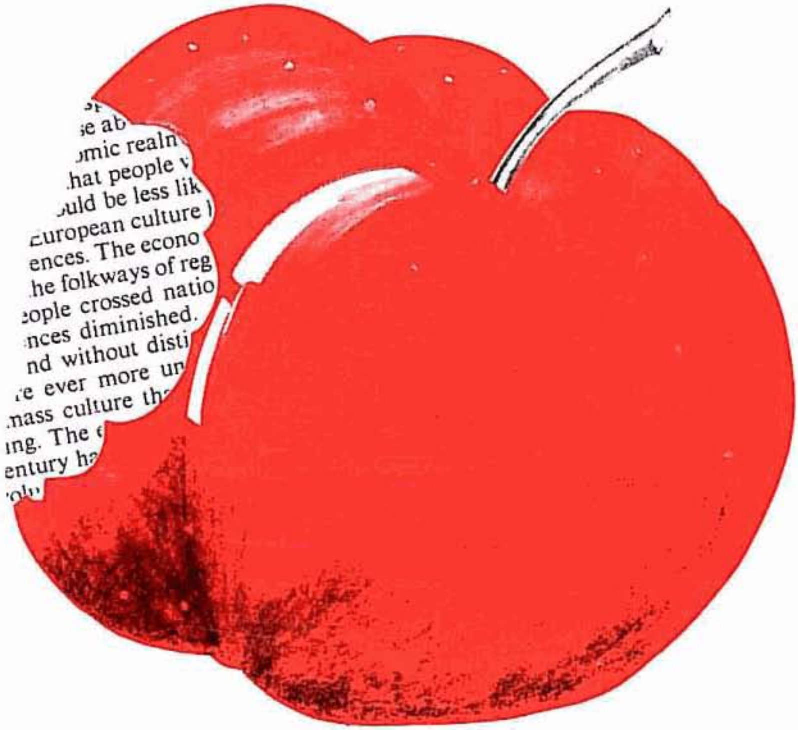
ICAME NEWS

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



NAVF

No. 10

May 1986

CONTENTS

Guest editor: *Bengt Altenberg*, Survey of Spoken English,
Lund University, Sweden

Editor's Foreword

This issue of *ICAME News* is dedicated to W. Nelson Francis, the pioneer of English computational corpus linguistics. There may have been better occasions for doing this - last year, for example, marked his 75th birthday and in 1987 twenty-five years will have passed since the inception of the Brown Corpus. But no special occasion is needed to honour a great scholar.

The importance of Nelson Francis to the work of ICAME and to computerized corpus research in general is well known to the readers of this newsletter. In this number, we pay tribute to him as a person and a scholar, directly or indirectly, in each of its three main parts.

The first part presents a double portrait of Nelson Francis: first, a self-portrait drawn with historical sweep in the 'classical' dinner speech he made at the 5th ICAME Conference at Windermere in May 1984 (reproduced here at the request of many participants and with the speaker's reluctant permission), followed by an appreciation by Jan Svartvik, Lund University.

The second part contains abstracts from the 6th ICAME Conference at Röstånga, Sweden, in May 1985. The varied interests and applications reflected at this conference give a good picture of the present state of computerized corpus research in English and its development since the creation of the Brown Corpus in the early 1960s.

The third and final part contains an updated bibliography of studies related to computerized English corpora, followed by a study of a previously neglected aspect of the Brown Corpus. The bibliography containing close to 300 items should be compared with the first list of 57 works published in *ICAME News* 2 in 1979. This increase in corpus-based research illustrates better than anything else the steadily growing interest in machine-readable English corpora as a basis for language study and software development, and the great importance of Nelson Francis' pioneering work nearly twenty-five years ago.

Bengt Altenberg
Lund University

DINNER SPEECH

given at the 5th ICAME Conference on Computers in English Language Research, Windermere, England, 21 May 1985

W. Nelson Francis
Brown University

You probably can't see it from where you sit, but some of you may have noticed that I am wearing a tie clip in the shape of a monkey wrench - or what I believe is called an 'adjustable spanner' in the curious dialect of this country. The story behind this peculiar piece of jewelry goes back to the early 60s, when I was assembling the notorious Brown Corpus and others were using computers to make concordances of William Butler Yeats and other poets. One of my colleagues, a specialist in modern Irish literature, was heard to remark that anyone who would use a computer on good literature was nothing but a plumber. Some of my students responded by forming a linguistic plumber's union, the symbol of which was, of course, a monkey wrench. The husband of one of them, being a jewelry manufacturer, had a few of these clips made. I cannot say that they have become collectors' items, but I would certainly not part with mine.

I later encountered that colleague on some social occasion, and he had the grace to say "Ah, Nelson, me bhoy, it was not you I was after callin' a plumber; it was them other fellows like Henry Kučera." - I should point out that much reading of Sean O'Casey had had a strange effect on his speech. I don't think he is genuine Irish; if so, he's the only Irishman I ever met named Kraus.

People are more familiar with computers nowadays, and perhaps not so hostile as my colleague David O'Kraus. But corpus-based computational linguistics is rather mysterious to the general public. Just a few days before I left home to come here, I found myself at a cocktail party of the kind university administrators feel obliged to give at the end of term. I got into conversation with a middle-aged lady - at least *I* would call her middle-aged, since she seemed not a day older than I am. She asked the usual question that lay folk ask of academics at this time of year - "What are you going to be doing during the vacation?" I told her I was leaving shortly for England. "And what's taking you to England?" she asked. "I hope it's a 747," I answered, "but you never can tell about British Airways." "That's not what I mean," she said, "why in the world are you going to **England**?"

"Well, there's a conference going on about corpuses. People from all over Europe are going to be there."

"Oh. But what are you doing about corpses?" - (as a good Bostonian she doesn't pronounce postvocalic r's).

"Most of the people are trying to parse them with computers. We have a standard one at Brown."

"Oh, dear. Will you be taking it with you?"

"No, only my wife. They have our corpus there already. The British have made a replica of it."

"Isn't that what they call cloning?"

"Not exactly - cloning means making an exact duplicate. Their corpus is not exactly like ours, because it's British, you see. Whenever we say 'monkey wrench' they say 'adjustable spanner'."

"How odd. But what do you mean by passing it?"

"Well, before you can parse it, you have to segment it. That's pretty hard to do with a computer. But at Brown we have a very sharp hacker to help with that - name of Andy Mackie."

"That's a funny name for a hatchet. But why can't you leave the poor dead corpse in peace?"

"Oh, our corpus isn't dead, it's still living. Or at least it was in 1961 when we collected it."

At that the lady gasped, gave me a frigthened look, and said "Excuse me, I think I need another drink."

"Why don't you let me get it for you?" I offered, politely. But within seconds she had disappeared into the crowd around the bar.

Not long afterward, I saw this same lady talking to my wife. From the way they were looking at me I was sure they were talking about me. As soon as I could I got Nearlene into a corner and asked what the lady had been saying.

"Well," said Nearlene, "she asked me if I knew you. When I said I knew you pretty well, she said "I think there's something wrong with him!"

"I often feel that way too," Nearlene responded.

"He told me he was going to a convention in England where they were all going to chop up this corpse and pass the pieces around. And the corpse isn't even dead!"

"Yes," said Nearlene, "they do that sort of thing all the time. That's why they're called computational linguists."

So here we all are, to talk about our ghoulish hobby. By the way, I would like to make a correction to a remark last evening implying that the origin of this disorganized organization occurred near the fish market in Bergen in April of 1979. I should point out that the origin was two years earlier in the English Department at Oslo. There were five charter members, four of whom are here tonight - Stig Johansson, Geoff Leech, Jan Svartvik, and myself. The fifth, Jostein Hauge, could not be here but has sent his deputy, Knut Hofland. On that occasion was formed the International Computer Archive of Modern English or acronymically ICAME. From that grew the first congregation in Bergen, which was not as widely heralded as it might have been because the nuclear mishap at Three Mile Island crowded us off the front pages. Perhaps if Ronald Reagan and

Maggie Thatcher can keep their warships out of the Persian Gulf, we might have better luck with the corpus this time.

The next organization that appeared was the very small and select International Society of Angry Wives or ISAW. This has never had more than four members, two of whom - Nearlene and Fanny - have been brave enough to show up here. Fanny, in fact, following the old American advice "If you can't lick 'em, join 'em", has bored into our midst. We send our greetings to Gunilla and Faith Ann - we wish you was here.

Now I expect the ultimate organization to be born from this conference, by Caesarean section, will be the International Congress Of New and Quite Unusual Experiments Related to English Discourse.

FOR W. NELSON FRANCIS

Jan Svartvik
Lund University

When recently visiting the People's Republic of China, I had the opportunity of discussing English linguistics with a number of Chinese colleagues representing different universities in the Republic. One of the matters that my hosts invariably tended to bring up for discussion was related to 'corpus' (and I'm not thinking just of queries about whether the plural is *corpuses* or *corpora*), for example by questions such as the following: "How many are there available?", "How can we make our own?", and so forth. Obviously, one reason for these questions may have been Oriental courtesy in that their visitor carried an imprint of 'corpus linguist' and thus might be expected to be able to answer such questions without losing face; another, I think, was that there is a genuine interest in corpus-work. When I suggested it might be a good thing to start by using corpuses that are already available, the usual reply was: "Yes, of course, but we want to make our own". Clearly, there is now status in making an English corpus, and that goes not only for China but also for other countries.

Evidence of this realization of the value of using a corpus in machine-readable form can be found in (1) the number of such corpuses now available for general use by scholars all over the world; (2) the creation of the International Computer Archive of Modern English (ICAME); (3) the publication of *ICAME News*; (4) the annual meetings under the auspices of ICAME; (5) the number of corpus-based studies. While we note with satisfaction the progress made in English computational corpus linguistics, we should not forget that it is very largely the natural continuation of the making of the Brown corpus. That is why this number of *ICAME News* is dedicated to W. Nelson Francis.

ICAME was set up in 1977 with the primary purposes of (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. ICAME is a most extraordinary international organization in that it has no official status: to my knowledge, it is not registered in any country, it has for certain no President and, above all, no Treasurer -- and hence no funds. But there is a coordinating secretary by the name of Stig Johansson and a distributor of corpuses and publisher of the newsletter by the name of the Norwegian Computing Centre for the Humanities in Bergen. I am confident that the success of ICAME can be chiefly attributed to the contributions to our cause by Stig Johansson in Oslo and Jostein Hauge and his colleagues in Bergen.

Our latest conference, ICAME 6th, took place in May 1985 at Röstånga, outside Lund in Sweden. This volume of *ICAME News*, edited by Bengt Altenberg, includes abstracts of papers read at the conference and, in addition, a most useful bibliography including studies based on English computer corpora. It seems to me a rather impressive collection.

One striking thing is the range of uses to which corpuses have been put. I doubt that Nelson Francis's early critics (and perhaps not even the Founding Fathers of the Linguistic Plumber's Union at Brown themselves) could foresee the scholarly ingenuity that is reflected in studies based on electronic corpuses, including such fields as lexicography, lexicology, syntax, semantics, word-formation, parsing, question-answer systems, software development, spelling checkers, speech synthesis and recognition, text-to-speech conversion, pragmatics, text linguistics, language teaching and learning, stylistics, machine translation, child language, psycholinguistics, sociolinguistics, theoretical linguistics, corpus clones in other languages such as Arabic and Spanish -- well, even language and sex.

It is my guess that what we have seen is only the beginning of the linguistic electronic revolution. With the Mighty Micro now within easy reach on the linguist's desk and online links with international databases, we cannot even guess what to expect next. It is right and important that linguists should take an active part in this revolution.

The beginnings of most revolutions can usually be traced back to one leader. In the case of the Electronic English Corpus Revolution, the historical research task is simple: the revolutionary leader is W. Nelson Francis, the originator and begetter of the Standard Sample of Present-Day Edited American English, for Use with Digital Computers, alias the Brown Corpus. I remember as a revolutionary milestone that morning in the early sixties when Nelson, straight off the sun deck of the Queen Elizabeth, walked into the office of Sir Randolph Quirk (then just Randolph Quirk) at the Survey of English Usage in University College London, banged the Brown Corpus writ on magnetic tape on Randolph's desk, saying: "My Sir, Habeas Corpus".

It was a fine act, Nelson, and we are grateful to you for setting us on the revolutionary path. In the words of the Constitution of the United States of America, the privilege of habeas corpus "shall not be suspended, unless when in cases of rebellion or invasion the public safety may require it". Let's hope there will never be a reason to suspend the right to future use of your corpus.

ICAME 6TH

The 6th International Conference on English Language Research on Computerized Corpora at Röstånga, Sweden, 19 - 22 May 1985

1 INTRODUCTION

Bengt Altenberg
Lund University

The Sixth International Conference on English Language Research on Computerized Corpora (ICAME 6th) was held at Röstånga outside Lund, Sweden, on 19-22 May 1985. The conference was organized by Jan Svartvik, Lund University, under the auspices of ICAME, and sponsored by The Swedish Royal Academy of Letters, History and Antiquities, The Swedish Council for Research in the Humanities and Social Sciences, and Lund University.

Nearly fifty participants from thirteen different countries attended the conference. Thirty papers were presented on a variety of subjects, and panel discussions were devoted to the possibility of setting up a corpus of spoken American English and to the future organization of ICAME.

The main theme of the conference was 'Lexicology and parsing', but the papers in fact represented a much wider range of interests. This variety is a good illustration of the vitality of corpus-based research at present and the increasing awareness among linguists of the possibilities that computerized corpora offer for the description of English and other purposes.

The following summary of certain major tendencies reflected at Röstånga is not intended to give an exhaustive picture of the conference, but may serve as an introduction to the abstracts presented below.

1 New corpora

A prerequisite of computer-aided corpus research is the availability of machine-readable corpora. Apart from the existing 'standard' corpora - the Brown Corpus (BC), the Lancaster-Oslo/Bergen Corpus (LOB) and the London-Lund Corpus (LLC) - which have long been accessible to scholars all over the world, machine-readable material of various kinds is being used or collected at many places. The vast Birmingham Collection of English Text continues to grow and there are plans to make available parts of the corpus in concordanced format and perhaps to create modern counterparts of the BC, LOB and LLC corpora (see Renouf).

A new mega-word corpus of contemporary English is being compiled at Nijmegen for syntactic analysis within the TOSCA II project. The software developed at Nijmegen will also be tested for automatic syntactic corpus analysis of Arabic and Spanish (Aarts).

The growing interest in spoken English was evident in several ways. New methods of gathering natural spoken data are being tested in Birmingham, and a corpus of spoken English (careful delivery, BBC sound broadcasts) is in preparation at Lancaster University for the automatic intonation assignment project in progress there (Knowles & Taylor). There are also advanced plans to establish archives of American English (with an emphasis on speech) at Berkeley, California, and Uppsala, Sweden. The various problems connected with the creation of these archives are discussed in the abstracts by Chafe and Tottie.

2 Developments in tagging and parsing

The word-class tagging of the LOB Corpus has now been completed and the results will be published this year (Johansson & Hofland). However, further refinement of the automatic LOB tagging program is in progress at Lancaster University (Blackwell). Various parsing systems for automatic or semi-automatic syntactic analysis are also being developed at Lancaster (Leech), Lund (Svartvik, Eeg-Olofsson) and Nijmegen (Aarts, Oostdijk)

3 Lexicology

Large corpora, especially vast ones like the Birmingham Collection, offer interesting possibilities for lexicological research (Renouf). Machine-readable dictionaries are also well suited for lexicological studies with ramifications into syntax and semantics and applications in lexicography and language learning/teaching. Various projects are going on in Liège, Amsterdam and Jerusalem. One promising approach is to examine the grammatical coding of words in a dictionary and test their systematicity internally (against other words in the dictionary) and externally (against grammars and other dictionaries). The results of such comparisons are of interest not only to grammarians, lexicographers and ordinary dictionary users (Moulin et al, Devons) but also to creators of lexical databases for automatic parsing programs (Akkerman et al). Another way of condensing the description of words is to establish a network of lexical definitions by means of chains of semantic primitives (Meijs). This approach, too, has useful applications, eg in parsing, text characterization and compression and expert systems.

4 New projects and applications

The existence of machine-readable corpora and software developed to analyse them has a healthy tendency to engender new ideas, approaches and applications. Good examples of such spin-off effects are the following recently started projects:

Lancaster: Spelling detection and correction (Elliott)
Lancaster/Winchester: Automatic intonation assignment (Knowles & Taylor)
Lund: Text-to-speech conversion (Svartvik, Altenberg, Stenström, Eeg-Olofsson)
Lund/Gothenburg: Teaching communicative competence (Aijmer)

Projects of this kind are of interest not only for their applicational possibilities, but also - and perhaps primarily - for the theoretical insights they provide and for their contribution to the description of English.

The theoretical and descriptive value of corpus research was also clearly demonstrated in several stylistic or sociolinguistic studies presented at the conference:

Uncovering dimensions of linguistic variation (Biber & Finegan)
Analysing linguistic style (Cheng)
Pronominal manifestations of sex-roles (Kjellmer)

5 Software development

Developing software for various linguistic purposes is an essential part of most of the projects presented at the conference. In addition to the techniques created for the analysis of large corpora, experiments are going on in other areas such as machine-translation and question answering (Sgall) and the production of query languages for semantic network databases (Jones).

However, the rapid development of computational software, inside and outside linguistics, also creates problems. Even within a small organization like ICAME the variation in techniques and software tools for comparatively similar research tasks is considerable. To some extent this is inevitable, but undoubtedly much could be gained by an increased exchange and standardization of software within ICAME (Atwell).

There will be a good opportunity to discuss these and other matters of common interest at the next ICAME conference, which will be arranged by Willem Meijs and his colleagues at the University of Amsterdam in June 1986.

The following abstracts represent the great majority of the papers and progress reports presented at Röstånga. The abstracts have been arranged according to project location or (where this has been more natural) major research field.

2 RESEARCH AT LANCASTER UNIVERSITY

Geoffrey Leech:
Current computer corpus-based research at Lancaster

In addition to other purposes for which a corpus is useful, I would like to concentrate on the purpose of developing natural language processing software. This is the main emphasis of current research at Lancaster, undertaken within the Unit for Computer Research on the English Language (UCREL for short).

There are three main projects, which I list below together with the personnel working on them.

UCREL (co-directors: Roger Garside, Geoffrey Leech)

1 *SERC project* (Andrew Beale, Susan Blackwell, Barbara Booth, Fanny Leech):
Syntactic analysis of the LOB Corpus (1983-6)
Sponsor: Science and Engineering Research Council

2 *ICL project* (Stephen Elliott):
Developing a context-sensitive spelling checker (1983-6)
Sponsor: International Computer Limited

3 *IBM (UK) project* (Gerry Knowles, Lita Taylor):
A corpus of spoken English for speech synthesis research (1984-7)
Sponsor: IBM (UK) Research Centre

We have also benefitted from the collaboration of Geoffrey Sampson and Eric Atwell, both now at the University of Leeds.

In all three projects, there is a common principle: that by systematically analysing a corpus of naturally-occurring text, one can provide essential information for the improvement of software.

We are using CLAWS (= Constituent-Likelihood Automatic Word-tagging System), developed for the grammatical tagging of the LOB Corpus, as a prototype for the methodology of other projects. CLAWS, using probabilistic methods, was fairly successful (achieving a c. 96% correct result). We believe that a similar methodology will be successful in other projects.

The methodology involves a cyclic progression:

1 Analyse corpus (using analytic computer system)
2 Do error analysis of output (to identify weaknesses of the system)
3 Improve analytic system (making use of information derived from 1 and 2)
4 Analyse new corpus (using improved system)
5 etc...

To illustrate this, I will briefly mention two 'deliverables' we have to produce in the SERC Project:

A A new improved version of CLAWS (see Susan Blackwell's paper)
B A parser and a parsed version of the LOB Corpus.

The parser will be non-standard and probabilistic. The theory is that the statistics derived from a manual parse of part of the Corpus will provide the probabilistic basis for the automatic parsing. The methodology will be based on steps 1-4 above.

Susan Blackwell:
Revision of the LOB tagging suite

The LOB Corpus has been tagged by a series of programs collectively known as CLAWS - Constituent-Likelihood Analysis Word-tagging System. However, there are several undesirable aspects of the present system:

1 Pre-editing

The first program in CLAWS is the automatic Pre-Editor, which verticalises the text and tags punctuation. At present the input text has to be manually coded to deal with certain typographical and linguistic information. This is inefficient, since it eliminates the advantages of acquiring text in machine-readable form. Moreover, much of the encoded information is redundant.

The new Pre-Editor, therefore, will run over 'raw' text, and the manual stage will be eliminated. This entails formulating new algorithms to deal with the automatic interpretation of upper- and lower-case letters and punctuation. Potential problem areas are upper-case abbreviations (which could be confused with a full stop and the start of a new sentence) and enclitics like *John's* for 'John is' (which could be confused with a genitive). Most of these new tasks will be carried out by the word-tagging program rather than the Pre-Editor.

2 Tagset

The present tagset is often confusing and does not lend itself easily to applications such as the production of concordances or frequency listings. For example, in order to extract all verbs from a given text, one would have to search for no less than five separate groups of tags: those beginning with VB, BE, HV, DO and MD.

In the new tagset, however, *all* verbs will start with 'V'. In general, the first letter of a tag will indicate its grammatical category, and subsequent letters will represent subcategories. This will make the output text more suitable for the Automatic Parser which is currently being developed.

3 Disambiguation

Chainprobs, the tag disambiguation program, refers to a matrix of tag-pair cooccurrence probabilities (unlike previous systems which used context-free rules). The technique proved to be remarkably successful, but there is still room for improvement. The present matrix is based on statistics which are modelled on the Brown tagset and derived from the Brown Corpus. The new version of Chainprobs will have a LOB-based matrix, which should lead to improved accuracy in tagging.

The proposed changes in the Tagging Suite should simplify the tagging process and render it more transparent to the human user. It should increase the accuracy of the tagging decisions and make it possible to run CLAWS over new raw texts in their original orthography, without the need for manual pre-editing.

The modifications should prove beneficial in several areas of future work, including automatic parsing, semantic analysis and the production of concordances and frequency listings.

Stephen Elliott:
Progress on the LOB-based context-sensitive textual error detector and corrector project

The aim of this project is the detection and correction of spelling errors that form other valid English words. Eric Atwell proposed the ideas and did much of the initial development work.

The spelling error detector program takes its input from the CLAWS programs and uses the tag-pairs probability matrix. Text is processed one word at a time. The program reads the probability of the current pair of word tags from the matrix and multiplies this by the probability of the last pair of word tags to obtain a combined figure.

The CLAWS tagset of 134 word-tags is used, but this contains some distinctions which are unimportant for a spelling checker. There are also some distinctions that could usefully be added for use in a spelling checker, eg between countable and uncountable nouns. Any reductiou in the number of word-tags will speed up the program.

A small number of words are tagged incorrectly; this could be avoided by using a tagged wordlist at the word-tagging stage, but errors are rarely created by this mistagging.

The program produces probabilities of pairs of word-tags held as THISPAIRPROB and LASTPAIRPROB, and the product of these two probabilities. Using 35 sentences each containing one error, in 14 sentences the error had the highest figures of the sentence and in a further 7 sentences it had the equal highest value. Also, in 24 sentences the error was at a peak on both sets of figures. There were only 4 sentences where there was no peak at the error. However, to use this information the program would have to be rewritten to hold several lines at once.

The minimum value of the product of THISPAIRPROB and LASTPAIRPROB at an error is 25, whereas the overall minimum is 19. A boolean function makes various tests on the figures around a word and, if they are true, the word is flagged as an error.

As far as correcting errors is concerned, moves are currently being made to acquire EXPERT SPELLERS, developed by Dave Fawthrop at Bradford University. There will be several problems to overcome to get the two programs to work together.

Gerry Knowles and Lita Taylor: Automatic intonation assignment

The project in automatic intonation assignment was begun at the University of Lancaster in the Autumn of 1984. We are working at the linguistic end of text-to-speech processing, and concentrating on the relationship between linguistic structures and prosodic patterns, rather than on the generation of pitch patterns themselves. The objective is to do automatically what a phonetician does when assigning intonation to a written text. The text can subsequently be 'spoken out' with high quality intonation by a speech synthesizer.

The project falls into two parts: (1) the collection of a corpus of contemporary spoken English, and (2) the development of a set of rules to generate the prosodic transcription automatically from a conventional orthographic text. The collection of the corpus is being funded by IBM (UK) Ltd. The work on transcription-by-rule was supported in 1984-85 by the Humanities' research fund of the University of Lancaster, and is currently supported by IBM (UK) Ltd.

The corpus is expected to be of value in speech research in general, but in the first instance it is to be used for research in intonation. Much conventional work in this area involves theorizing from general linguistic principles and using invented examples and intonation patterns. It is difficult in these cases to know how normal the invented patterns really are. Even major theoretical assumptions - eg that it is necessary to make a full parse of a sentence before assigning intonation - might not in fact be valid. Linguists' intuitions about intonation are fallible in this area, and it is essential to base work on real and appropriate data.

The corpus will be rather smaller than written corpora. We have collected about 20,000 words so far, and aim at a total of about 100,000 by 1987. In order to avoid problems of sociolinguistic variation, we are keeping to RP, and because the texts are to be used as models for speech synthesis, the preferred style is the careful delivery of a text, rather than spontaneous speech. The best source of high quality data is in BBC sound broadcasts, for which we have obtained the necessary permission.

In view of the relationship between intonation and punctuation, considerable care is being taken to avoid circularity. The recordings are first written down in ordinary spelling but without punctuation, and are punctuated by some other person. The work of prosodic transcription is shared between two phoneticians, one at Lancaster and one at IBM, who have played no part in the preparation of the orthographic versions. The transcription system being used is fairly close to that used by O'Connor and Arnold in their *Intonation of colloquial English*. The transcribed texts are then used as targets for the intonation assignment rules. Selected extracts of the text are being analyzed instrumentally by IBM, and the Fo contours compared with the marks in the prosodic transcriptions.

Before the intonation rules are applied, the text is grammatically tagged, making use of the LOB word-tagging programs developed at the University of Lancaster. The first stage in intonation assignment is to divide the text into tone groups, and it is assumed (with a number of known exceptions) that the sections of text marked off by punctuation are also prosodic units of some kind. The text is processed a 'chunk' at a time, from one punctuation mark to the next, and this 'chunk' is divided into tone groups.

The term 'tone group division' implies working from the top down, but the most easily inferable rules work bottom up. For instance, grammatical words tend to attach themselves as clitics to lexical words, eg *the* is attached to *North* in the phrase *the North*. An adjective or numeral will tend to combine with a following noun to form a single prosodic unit with no rhythmical or pitch discontinuities, eg *North* combines with *Wind* to form the phrase *North Wind*. These groups may under certain conditions combine to form larger units, eg in the phrase *stronger than the other* the groups *stronger* and *than the other* are collapsed into one. In this way larger units are constructed out of individual words. Conventional tone groups belong to some (possibly arbitrary) level in this process.

At present tone group division is based solely on grammatical and phonological criteria. To deal with 'given' and 'new' information, with parallelisms, and with compounding, we shall need further knowledge about the text. And having identified the tone groups, we shall need some means of predicting appropriate nuclei.

3 RESEARCH IN OSLO AND BERGEN ON THE LOB CORPUS

Stig Johansson and Knut Hofland: Current work on the tagged LOB Corpus

A study of the word-class distribution in the tagged LOB Corpus gave the results shown in the table below. Francis and Kucera's (1982:547) figures for the Brown Corpus are listed for comparison in the column to the right.

	LOB			Brown
	A-J%	K-R%	Total	Total
Nouns	26.9	20.1	254,801	272,984
Verbs	16.4	21.9	179,910	185,393
Determiners	13.0	10.5	124,949	123,321
Prepositions	13.1	9.6	123,342	122,613
Adjectives	7.8	5.7	73,609	72,034
Pronouns	5.0	13.2	71,490	66,879
Adverbs	5.0	7.2	55,946	53,283
Conjunctions	5.5	5.4	55,454	60,328
Quantifiers	2.2	0.9	19,125	20,853
Infinitival *to*	1.5	1.7	15,817	15,030
Wh-words	1.5	1.6	15,697	14,921
Not	0.6	1.1	7,447	6,976
Existential *there*	0.3	0.3	2,793	2,280
Interjections	0.0	0.4	1,113	629
Other (formulae, foreign words, quoted forms, letters, 'ditto tags')	1.1	0.4	9,554	

There are considerable differences between categories A-J (informative prose) and K-R (fiction). In A-J there is a higher frequency of nouns, adjectives, determiners, and prepositions, probably reflecting greater complexity at the noun phrase level; cf Ellegård's (1978:46f) observations on phrase length and phrase depth based on part of the Brown Corpus. In K-R there are more verbs, adverbs, pronouns, interjections, and occurrences of *not*. The last two features clearly reflect the proportion of dialogue; cf Tottie (1982) as regards the greater frequency of negative expressions in conversation. The style in K-R is more verbal. The clauses are shorter and noun-phrase slots are more often occupied by pronouns.

Comparisons of the LOB Corpus and the Brown Corpus should be made with caution, because of some differences in tagging conventions. The rank order of the word classes is almost identical. Some minor differences between the corpora are in part parallel to those pointed out above between categories A-J vs K-R in the LOB Corpus. Note the lower number of nouns in the LOB Corpus and the higher frequency of pronouns, adverbs, interjections, and occurrences of *not*. These differences may be due, at least in part, to a somewhat higher proportion of dialogue in the LOB Corpus than in the Brown Corpus; see also Johansson (1985).

Other current work includes the preparation of a homograph-separated concordance, homograph-separated word lists, and studies of tag combinations and collocations.

References

Ellegård, A. 1978. *The syntactic structure of English texts*. Gothenburg Studies in English 43. Gothenburg: Acta Universitatis Gothoburgensis.

Francis, W.N. & H. Kucera. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Johansson, S. 1985. Some observations on word frequencies in three corpora of present-day English texts. *ITL Review of Applied Linguistics* 67-68: 117-126.

Tottie, G. 1982. Where do negative sentences come from? *Studia Linguistica* 36: 88-105.

4 RESEARCH AT BIRMINGHAM UNIVERSITY

Antoinette Renouf: Progress report on work at Birmingham University

1 Corpus developments

The lexicographic analysis of the 7.3 million word Corpus is well in hand, with about 80 per cent of the resultant database complete. Since the last ICAME conference, work has begun on the correction of errors in this corpus. In five months, a team consisting of two Dutch undergraduates and one Malaysian computer scientist has corrected half of the book component of the Corpus, using a series of strategies designed to achieve accuracy efficiently. A separate report on the details of the process will be made available in due course.

The reserve corpus of text which we also have continues to grow, and now amounts to around 17 million words of written language. An exact count is under way. Certain areas of lexis which were minimally represented in the main Corpus, such as those associated with 'alternative' aspects of society, technology, hobbies, and sports, are now better covered. There will be long term uses for this resource, but in the immediate future we are likely to go ahead with the concordancing of certain items which have occurred with low or zero frequency in the main Corpus.

Over the last year or so, three experimental batches of spoken data have been recorded: in March 1984 (*ICAME News* 9: 12-13), July 1984 and July 1985. Each time, the parameters have been changed in the light of previous experience. The latest of these events has been orchestrated by a postgraduate student, Martin Warren, whose own research shows that natural speech will only emerge in situations where the speakers feel that the conversation is their own, that they are responsible for the outcome. Accordingly, he has minimised task instructions, trying instead to create conditions in which people feel naturally moved to initiate and sustain a conversation. For example, the participants were required jointly to assemble an apparatus, to open a container with unexpected contents, and so on. The results of this experiment are still to be analysed in detail, but we shall be expecting to detect a greater degree of 'naturalness', as defined by criteria which we are evolving, than is evident in the earlier data. A detailed report on the spoken data which we have elicited so far will be produced soon.

Birmingham is now considering how to make available its corpus data to fellow researchers. Two possibilities spring to mind: a one million word corpus in concordance format could be extracted from existing holdings, and made accessible on-line; or a larger corpus, of twenty to twenty-five million words, again in concordanced format, could be transferred to compact disk, and distributed with accompanying software. There was a useful discussion of the data resources required by colleagues present at the conference, and it was generally felt that the time had come for the existing Brown, LOB and London-Lund corpora to be supplemented, for comparative and other purposes, by more modern counterparts. A suggestion was made that Birmingham should create two new million-word corpora, one of spoken English and one of written, which could be made available to everyone. This possibility is now being considered, and it is probable that one of the proposed corpora will involve some degree of collaboration with colleagues at Nijmegen and at the Norwegian Computing Centre for the Humanities (NAVF) at Bergen.

2 Software developments

It is now possible to produce sample concordances for the more frequent word forms in the Corpus, which would otherwise generate unmanageable amounts of data. Samples can be arithmetic, logarithmic or random, and of any size. Extended concordances can now be accessed on-line, for word forms associated with larger discourse patterns. The larger contexts are taken from a one million word subcorpus selected by lexicographers for this purpose, and context size can be specified. Concordances sorted by a variety of criteria, such as that of left-hand context, can also be produced on-line from this subcorpus.

Profiles of collocational frequency immediately adjacent to the node-word have been produced automatically for some time, and non-adjacent patterning is similarly available. Further work is taking place in this important area, however. One immediate development, in view of our present interest in high-frequency word forms, will be to provide statistical help in judging the relevant significance of lexical combinations at the top frequency levels.

Our computer staff have also produced the first of a series of distributional analyses of the lexis in the Corpus texts. These, together with the lemmatized word lists and other statistical data which we now have, will provide a powerful set of criteria for the selection and presentation of lexis for a variety of linguistic and pedagogical purposes.

5 LEXICOLOGY

Eric Akkerman, Pieter Masereeuw and Willem Meijs
University of Amsterdam

ASCOT: One-third of the race

ASCOT is a project that aims at the development of a lexical database and morphological analysing system, which together can provide the coding of words in uncoded corpora (for an introduction, see *ICAME News* 9:19-20). In our paper we presented a survey of our activities and findings in the project's first year. For a detailed report, see Akkerman, Masereeuw & Meijs (1985).

1 A comparison of OALD and LDOCE

Since the basis of the ASCOT lexicon will be the computer-tape version of an existing dictionary, a detailed comparison was made between the *Oxford Advanced Learner's Dictionary* (OALD) and the *Longman Dictionary of Contemporary English* (LDOCE). The following points of difference were established.

1) *Entry structure*: LDOCE's entries are structurally less complex than OALD's entries. Where LDOCE has separate entries for phrasal and prepositional verbs, almost all compounds and fixed collocations and many derived words, OALD often compresses all of these into one entry.
2) *Word class coding*: in a number of cases OALD is less consistent than LDOCE.
3) *Grammatical coding*: in general, LDOCE's grammatical coding system is both more comprehensive and more detailed than that of OALD. Furthermore, LDOCE's coding of verbs is more clearly structured and in some cases grammatically sounder than that of OALD.

On the whole, these advantages make LDOCE more suitable as a basic dictionary for the ASCOT project than OALD.

2 Using dictionaries on computer tape

To be able to extract the necessary information for the ASCOT lexicon by means of a computer program, it is required that the input file is structured very systematically. Therefore we first tried to develop a grammar describing the structure of OALD (which was the first dictionary of which we had a computer-tape version), specifying the order(s) in which the various kinds of information are presented and the typefaces and printing marks that identify these.

Such a grammar can then be automatically transferred into a parsing program. However, it turned out that the various chunks of dictionary information can be ordered in several ways and that many of these are left out in certain situations. As a result the grammar became rather ambiguous (because every possible order has to be accounted for), which caused a dramatic slow-down of the parsing process.

Meanwhile it had gradually become apparent that LDOCE would serve our purposes better, because it is structured very systematically. It is unnecessary to specify anything about its structure in the form of a grammar, because it uses a system in which every kind of information is unambiguously identified. Yet, in order to make the file optimally accessible to the programs that will be used to create the ASCOT lexicon, it was still necessary to develop a special-purpose program that can execute certain transformations (like the reconstruction of abridged derivations and the decompacting of compressed grammatical codes).

3 The morphological component

For the development of the morphological component (which can recognize words that are not in the lexicon, but which are the result of certain morphological processes) two approaches were followed:

1) An algorithm was developed and implemented in the computer language PASCAL. Special attention was paid to a number of specific problems that arose when the function of the algorithm was studied in detail.
2) A beginning was made to develop a grammar representing most of the productive rules of English morphology. Such a formal grammar can be transformed into a general parsing program, which uses it to analyse an input file.

Two parsing systems are available to the ASCOT project:

a) PARSPAT, which is being developed at the Computer Department of the Faculty of Arts of Amsterdam University (see van der Steen 1984);
b) ORACLE, which has been developed at Delft University of Technology (see Honig 1984).

In close cooperation with the researchers involved, a number of simple inflectional grammars were written and presented to both parsers. However, as PARSPAT is not yet completely finished and ORACLE has not been thoroughly tested, progress in this area is slow.

4 The design of the ASCOT codes

Finally, a beginning was made with the development of the ASCOT codes, which will be designed in such a way that the ASCOT lexicon can be used for many different purposes, not only as a basic lexicon for various automatic grammatical analysing systems, but also for specific text-queries. As far as the form of the ASCOT codes is concerned, we have chosen a structure consisting of different information positions. A kind of 'special options mode' must make it possible to choose exactly those types of information that are of interest. The code system will be worked out in detail in the second year of the project.

References

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

Honig, H.J. 1984. Oracle user's manual. A morphological analysis program. Delft University of Technology.

van der Steen, G.J. 1984. On the unification of matching, parsing and retrieving in text corpora. *ICAME News* 8:41-46.

Willem Meijs
University of Amsterdam

Links in the lexicon: The dictionary as a corpus

The first half of the title of this contributon is the same as that of a research project for which it is hoped funding will be provided by the Dutch Research Council (ZWO). At the time of writing prospects look good, which means that the project can probably get under way early in 1986. It is scheduled to take three years and will provide research posts for two English language specialists and one computer-scientist (all three half-time).

The aim of the project is the construction of a coherent system of linguistically usable meaning characterizations (the LINKS system) associated with the words in a comprehensive computerized lexicon (as is being developed in the ASCOT project), and with a specific theoretical linguistic framework.

The theoretical framework is provided by the kind of semantics put forward in Dik (1978), which does not make use of an abstract metalanguage, but reduces the meanings of the words of the language via a stepwise system of definition-chains to a limited number of basic words, which can then be regarded as the 'semantic primitives' of the system. The system must obey a strict maximum-economy principle, which stipulates that definitions must make maximum use of words already defined by other definitions, and thus avoid duplication. Thus if *man* and *person* are already defined (say as 'male adult person' and 'human being' respectively) then the definition for *bachelor* must not be 'unmarried male adult person' or 'unmarried male adult human being', but simply 'unmarried man'. Dik's approach will be combined rather undogmatically with certain insights from inter alia Aarts & Calbert (1979), Levi (1978) and Warren (1978, 1984).

Dik claims that his approach in fact reflects what he calls "standard dictionary practice". In a pilot project called "Natural Primitives?" associated with the ASCOT project, we have looked at whether this is true of the LDOCE. Not surprisingly, it turns out that LDOCE conforms only partially to this assumed dictionary practice, but enough to warrant further exploration and exploitation, as planned in the LINKS project.

In LINKS we want to make maximum use of the fact that LDOCE employs a restricted 'controlled vocabulary' consisting of some 2000 words. In fact we intend to turn the total set of definitions in the dictionary into a tagged corpus by first manually tagging the words of this controlled vocabulary and then automatically projecting the tagging onto occurrences of the words in the definition chains. We are planning to project two kinds of tagging: grammatical and semantic. The grammatical tags will be familiar labels like ART, ADJ, VBN, etc. For the semantic tagging we hope to employ the kind of systematic hyponym-hyperonym relationships indicated in Aarts & Calbert (1979). If the tagging can be successfully inserted into the definition chains, the individual words from the controlled vocabulary in those strings will all have labels. For instance: *travel*: VB*MOVE, *beer*: N*LIQ, *large*: ADJ*SIZE, *mallet*: N*TOOL, etc.

The next stage of the project will be to explore the definition chains by means of QUERY search patterns. Unlike the procedure in the 'Natural Primitives?' project, these can in fact be formulated as syntactic patterns. Thus the definitions of nouns nearly all conform to well-formed noun phrase structures with a central noun as the syntactic head and one or more pre- and/or postmodifying elements. With the use of the search patterns a fairly complete picture should emerge of systematically used definition patterns. By zooming in on the headword elements of those definitions, and 'jumping' from one headword to the next one down, the vertical links that are there can be inspected and systematically occurring definition chains detected.

The work will certainly not be so easy as the picture sketched here might suggest. There will no doubt be many obstacles which we cannot yet foresee. And some of the obstacles we can foresee are already formidable enough. To mention just one: it turns out that the definitions contain quite a few word forms that are derived from the basic vocabulary, like *wood/en*, *cat/like*, *show/y*, *brown/ish*, etc. To deal with such forms (as well as with regular inflected forms) the QUERY system will have to incorporate some version of the REROUTE program developed in ASCOT for morphological decomposition.

The LINKS software package resulting from the project should be useable as an independant syntactic-semantic database or as a component in (semi)automatic syntactic and semantic analysis. There may also be possible applications in text characterization and text compression, and as a baseline subcomponent in artificial intelligence (expert) systems.

References

Aarts, J.M.G. & J.P. Calbert. 1979. *Metaphor and non-metaphor: The semantics of adjective-noun combinations*. Tübingen: Niemeyer Verlag.

Dik, S.C. 1978. *Stepwise lexical decomposition*. Lisse: Peter de Ridder Press.

Levi, J. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

Warren, B. 1978. *Semantic patterns of noun-noun compounds*. Gothenburg Studies in English 41. Gothenburg: Acta Universitatis Gothoburgensis.

Warren, B. 1984. *Classifying adjectives*. Gothenburg Studies in English 56. Gothenburg: Acta Universitatis Gothoburgensis.

André Moulin, Jacques Jansen and Archibal Michiels
University of Liège

Computer exploitation of LDOCE's grammatical codes

The Longman Dictionary of Contemporary English (LDOCE) is a 'dictionary-cum-grammar': the syntactic potential of a word is defined through grammatical codes, which appear at both entry and definition level on the LDOCE computer tape. To make the codes directly accessible we had to devise a complex decompacting procedure which can rewrite the entry fields into the definition fields, disambiguate operators such as semicolons, commas, etc, and clarify the meaning and scope of some word context codes. Once this procedure is completed, it is possible to use such software packages as STAIRS or SAS and produce 'profiles', ie configurations of codes within the same definition, or KWIC indexes of code profiles, which eg clearly bring out the similarities in the syntactic behaviour of the modals and yield what looks like a condensed grammar of English. Using the profiles, one can investigate the mutual attraction or repulsion of codes or check the systematicity and validity of their allocation.

We concentrated on the V3 code (*I want him to go*) and compared our KWIC index with the list that appears in Quirk et al (1972). This enabled us to point out wrong or doubtful code allocations both in LDOCE (*give*, def. 17) and in the Quirk et al list (*claim*) and to show how the LDOCE coders' concern for surface structure accuracy, though justified on pedagogical grounds, often leads to difficulties when it comes to defining transformational potential. Indeed, LDOCE provides interesting indications, often in the form of caveats or restrictions, concerning the transformational potential of the codes assigned to a particular verb. This is illustrated for instance by *make* (05): [V2; (V3 pass.)]: *The pain made him cry out/She was made to wait for over an hour*. This modified pattern, not coded in the introduction, is assigned the code V3, which, of course, is incorrect. In fact, the modified pattern should have been given a code of its own. In addition, neither the ability of a verb to undergo this passive transformation (eg *hear*) nor its inability to do so (eg *watch*) is systematically coded. Similarly, intransitive verbs which can have a passive meaning - the so-called 'deactivatives' - are not coded as such. The transform could have been coded 19, but this code does not exist. Besides, 19 is not valid for examples like *Such books don't sell.* In fact, as Michiels (1982:130) has pointed out, there are two ways of capturing the deactivative conversion in LDOCE: (1) conversion from [T1] to [L9]; (2) conversion from [T1] to [I0]. But both methods are inappropriate. Assigning [L9] to a deactivative verb implies, wrongly, that it is a linking verb. [L9] requires an adverbial, a condition which does not always apply to deactivatives (*This jacket won't button*). The assignment of [I0] implies that the deactivative can be used as any other intransitive and allows sentences like **Such a play acts.*

Alshawi et al (forthcoming) also make some interesting observations about code configurations. They take the example of *believe*, which is assigned codes V3 and T5 in the same sense (03,) and remark that "the presence of the T5 code tells us that *believe* is a 'raising to object' verb and logically two-place under the V3 interpretation. On the other hand, *persuade* is only assigned the V3 code, so we can conclude that it is three-place with object control of the infinitive. By systematically exploiting the collocation of different codes in the same field it is possible to distinguish the raising, equi and control properties of verbs." One could infer from their remark that all verbs that have both V3 and T5 allow the raising-to-object transformation in the sense concerned. But things appear to be more complicated. Take the example of *mean* (02). This sense does not allow raising to object: *He means that his son will succeed* does not signify the same thing as *He means his son to succeed*. In fact, the cause of the difficulty here is LDOCE's sense division. This is confirmed by a comparison with the *Oxford Advanced Learner's Dictionary of Current English*, where the two senses are kept apart and the corresponding verb patterns do not cooccur. The subtle correlation between sense and grammatical code is a difficult problem. From a practical point of view we can distinguish several cases:

(1) *One sense and two or more codes*: in the majority of cases, we have no criticism to make. Recall however the case of *mean* where it was suggested that the two patterns in question represent two senses.

(2) *One sense and a multiplicity of codes*: this is illustrated by a verb like *teach* which has only one sense and 10 different codes.

(3) *Several senses, each with their own code or cluster of codes.* It is a point often dealt with in grammar that a difference in grammatical pattern of the type *I remembered posting your letter* ([T4])/*Remember to post my letter* ([T3]) corresponds to a difference in meaning. In the LDOCE entry for *remember*, [T3] and [T4] are assigned to two different definitions or senses,

which is quite justifiable, even if one finds definition (02) not explicit enough: it should include a notion such as 'fail to'. In *forget*, however, the two patterns are assigned to the same sense. If we go on to examine all the verbs which can have both [T3] and [T4], we notice that they are coded correctly, that the two patterns belong either to the same sense (*start*) or to different ones (*try*) but that, too often, the difference in meaning is not made perfectly clear.

(4) *A sense division based on code differentiation*: a careful examination of some entries (eg *know*) suggests that, were it only to help the user, some senses could easily be brought together: this would of course lead to a reassignment of the codes (see Michiels et al 1980).

In conclusion, we have a few suggestions to make. The coding system should be extended to cover idioms. If *pick up* is defined as (01) T1: 'to take hold of and lift up', why not code *take hold of* as T1 too? If one thinks of the possibility of using the dictionary as a parser, this extended coding would certainly improve its efficiency. Another but more difficult extension would be towards 'inheritance rules': to what extent is it possible to establish and code the link between *to know that* (T5) and *the knowledge that* (T5 too?)? We have also raised pedagogical issues: is the code identification system mnemonically and syntactically justified? To what extent do the codes help or mislead the user? Why do so many users (learners and teachers) ignore them? If we try to develop and sophisticate the dictionary further, will we not put learners off for good? This is bound up with the whole question of what a dictionary should be or do. We have for instance shown that codes tend to influence sense division. Should the latter be based on semantic rather than syntactic criteria? To put it in Fillmorean terms, do syntactically-based word senses not contradict the native speaker's cognitive frames and thus give the foreign learner a false idea of the way in which the language he is learning categorizes reality?

References

Alshawi, H., B. Boguraev & T. Briscoe. Forthcoming. Towards a dictionary support environment for real time parsing. Computer Laboratory, Cambridge University.

Fillmore, C.J. 1982. Frame semantics. In *Linguistics in the morning calm*, ed by the Linguistic Society of Korea. Seoul: Hanshin.

Hornby, A.S. & A.P. Cowie (eds). 1974. *Oxford advanced learner's dictionary of current English.* Oxford: O.U.P.

Michiels, A. 1982. Exploiting a large dictionary data base. Unpublished doctoral dissertation, University of Liège.

Michiels, A., A. Moulin & J. Noel. 1980. Working with LDOCE. *ABLA Papers* 4, Brussels, VUB.

Michiels, A. & A. Moulin. 1983. The Longman lexicon of contemporary English: A tentative appraisal. *Grazer Linguistische Studien.*

Procter, P. (ed.). 1978. *Longman dictionary of contemporary English.* London: Longman.

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

Nina Devons
Hebrew University of Jerusalem

Observations on the lexicographic treatment of *one* and the approach adopted in FREQSUCON

The primary classification of the various senses/uses of *one* adopted in most current monolingual English dictionaries is syntactic (Table 1, Nos. 1-8). While this breakdown is useful in the case of most polysemic words, it was argued that this is not so in the case of *one.*

Semantically, the various senses/uses of *one*, with the exception of two or three colloquialisms, lend themselves to a tripartite classification: (a) numerical; (b) replacive, ie anaphoric or ostensive; (c) personal - in the absence of contextual reference *one* is interpreted as a person (or animate being).

The three entries of *one* - adjective, pronoun and noun - in Webster III (Table 1, No. 8), the most authoritative of the dictionaries that adopt a syntactic classification, were examined. The categorization by part of speech is seen to separate semantically related uses, and group together, without adequately distinguishing between them, meanings which are semantically remote.

Inspection of the entry *one* in the O.E.D. (Table 1, No. 11) revealed that, with one minor reallocation (noted below), a grouping of the subentries results in a breakdown which is in line with the tripartite semantic classification suggested above. Thus, subentries I - IV cover *one* as a numeral (adjective/determiner, pronoun and noun) including extensions which are closely associated with the numerical concept of *one*; subentry V defines the contexts (other than those given in subdivision 24 of subentry VI) in which *one* denotes a person or animate being; and subentry VI consists of three subdivisions, two of which (22 and 23) treat anaphoric and ostensive *one.*

Subdivison 24 reads: "after pronominal and other adjectives without contextual reference: = Person, body, persons; as in 'any one, every one, many a one, some one, such a one; little ones, the Holy One, the Evil One,' etc."

The primary semantic breakdown adopted in FREQSUCON follows the overall categorization of the O.E.D., except that the collocations of *one* listed in subdivision 24 of subentry VI, labelled "a pronominal or substantival form of 'a, an'", and as such grouped together with anaphoric/ostensive *one,* are, in FREQSUCON, categorized as a subdivision of the subentry to which they semantically belong, ie *one* = a person.

Table 1. The entry *one* in selected British and American dictionaries. Part of speech labels of separate entries (in dictionaries 3, 7, 8 and 10 below) or of subdivisions of a single entry.

	Dictionaries	Part of speech labels
A	Primary syntactic classification. All senses/uses of a particular part of speech are grouped together.	
1	*The American College Dictionary* (1957)	adj, n, pron
2	*The World Book Dictionary* (1977)	n, adj, pron
3	*Collins English Learner's Dictionary* (1975)	nc, determ, pron
4	*Oxford American Dictionary* (1980)	adj, n, pron
5	*Macmillan Contemporary Dictionary* (1979)	adj, pron
6	*The Heritage Illustrated Dictionary of the English Language* (1973)	adj, n, pron
7	*Longman Dictionary of Contemporary English* (1978)	determ, n/pron
8	*Webster's Third New International Dictionary* (1971)	adj, pron, n, vt
B	Independant classification. Subentries are not syntactically determined.	
9	*The Concise Oxford Dictionary of Current English* (1978)	adj, n, & pron
10	*Oxford Advanced Learner's Dictionary of Current English* (1974)	1) num adj, pron 2) indef pron 3) pers pron 4) impers pron
11	*The Oxford English Dictionary on Historical Principles* (1888-1933) with *Supplements* (1972-1982)	Num adj, pron, etc I) Simple numeral II) Emphatic numeral III) Pregnant senses IV) Particualrizing or partitive sense V) Indefinite pronoun VI) Pronominal or substantival form of *a*, *an* VII-IX) Obsolete uses, etc

6 RESEARCH AT NIJMEGEN UNIVERSITY

Jan Aarts:
TOSCA and after

Both the TOSCA (Tools for Syntactic Corpus Analysis) project and the LDB (Linguistic Database) project have recently been finished; the software designed and implemented in the projects is now operational (though not yet tested in actual use by other users than the Nijmegen researchers).

The TOSCA system is an interactive system for the analysis of corpora. The user of the system should provide a corpus, a formal grammar that can be automatically converted to a parser, and a lexicon, although the lexicon may also be built in the course of the analyzing process. The heart of the system is a computer program called the Linguist's Workbench. The Workbench organizes the communication between the components of the system: the corpus, the grammar, the lexicon, a logbook (in which the user's interventions are recorded) and the database in which the results of the analysis are stored.

The database system which will normally be used in conjunction with the TOSCA system is the LDB. In the LDB, analytic trees with labelled nodes can be stored. A query language enables the user to inspect trees which meet conditions stipulated by him, and also to investigate a wide variety of features: categories, functions and other relations between constituents, word forms, level of structure, etc. Although the LDB was built to be used in conjunction with the TOSCA system, it can also be used independently to store any kind of tree structures, whether linguistic or not. Recently, the LDB has been loaded with the analytic results from a 130,000 word corpus.

With the availability of TOSCA and LDB (see Note), corpus research in Nijmegen is now entering a new phase. Three new projects have started which make use of the TOSCA system and the LDB:

1 *Corpus Analysis of Contemporary English* (informally known as TOSCA II), in which a one million word corpus will be analyzed;
2 *Automatic Syntactic Corpus Analysis of Modern Standard Arabic* (ASCAM-SA), analysis of a corpus of 500,000 words;
3 *Análisis Sintáctico Automatizado de Textos Españoles* (ASATE), analysis of a 750,000 word corpus.

In all three projects, the type of formal grammar used is an Extended Affix Grammar. For each of the three projects a new corpus has been, or is being, compiled.

In addition to these, three other projects are being undertaken:

1 *LDB II*. Aim: (a) transfer of the LDB to microcomputers, (b) adaptation of the LDB for use in teaching;

2 *PG-project*. Aim: adaptation and new development of parser generators for grammars in corpus linguistics;

3 *Formal Grammars*. Aim: study of formal grammars in corpus linguistics and their relation to models in theoretical linguistics.

Note

Both TOSCA and LDB software will be made available to other corpus researchers, under the normal ICAME conditions. Those who are interested are requested to write to Jan Aarts, English Department, University of Nijmegen, Erasmusplein 1, Nijmegen, The Netherlands, stating the type of their own computer facilities and operating system(s).

Nelleke Oostdijk:
Some thoughts on the structure of a formal grammar

Current work on the Nijmegen research project TOSCA (II) includes the writing of an Extended Affix Grammar, to be used for the syntactic analysis of a corpus of contemporary English. In the past we wrote various subgrammars, each describing a particular category such as NP, AJP, PP, etc. Now we are combining these separate modules so as to form one grammar describing the English sentence. Whereas phrase categories were taken as a starting point in the modules, the grammar includes both functional and categorial constituents. It is in this context that we are concerned with coordination and reformulation (apposition). Assigning an appropriate structural analysis to a given string turns out to be quite problematic in case it contains two or more adjacent constituents that, on the same level of analysis, should be assigned the same function.

An attempt has therefore been made to give a basic description of both coordination and reformulation, and the conditions under which they occur. In our conception, the grammar consists of a number of modular components that are concerned with functions and categories, and some processes such as coordination and reformulation. These processes are not necessarily restricted to one particular module; they may be relevant to several or even all modules. They can be looked upon as 'subroutines'. Rule schemata have been devised for coordination and reformulation which, when called upon, will operate according to the rule-generating principle.

7 RESEARCH AT LUND UNIVERSITY

Jan Svartvik: The TESS approach

The aim of the Lund project TESS (Text Segmentation for Speech) is to describe some of the rules that govern the prosodic segmentation of continuous English discourse.

The different phases of the research process can be thought of in terms of a clock's face to illustrate the stages from recording and transcribing natural talk (beginning at 12 noon) to, hopefully, producing synthetic talk (ending at 12 midnight). The 1-3 period is the corpus compilation phase (which took place at the Survey of English Usage, University College London); the 3-9 period is the analysis phase; and the 9-12 period is the synthesis phase (the two latter falling within the research domain of the TESS project at Lund University).

In the analysis phase we attempt to gain a better understanding of how natural speech is segmented by analyzing some texts in the London-Lund Corpus of Spoken English. An important element in this work is the parser, which will help us to account for what is the grammatical content of tone units and also provide information about spoken English grammar and lexis.

In the synthesis phase we plan, first, to set up segmentation rules on the basis of our analysis of genuine speech; second, to 'reverse' those rules by applying them to written texts to be spoken; third, to check the result with the aid of a speech synthesizer.

Our hypothesis is that the natural segment for the analysis of spoken discourse is the prosodic segment we call 'tone unit'. (Other names for what appears to be roughly synonymous concepts are Halliday's and Sinclair's 'tone group', Chafe's 'chunk' and 'intonation unit'.) To take two recent definitions, Chafe considers the intonation unit to be "a sequence of words combined under a single, coherent intonation contour, usually preceded by a pause" (Chafe 1984:3); Brazil defines the tone unit as "the stretch of language that carries the systemically-opposed features of intonation" (Brazil 1985:11-12), and maintains that "each of the meaningful oppositions our description recognises *can* be identified on the basis of pitch treatment alone" (13). There seem to be a number of good reasons for considering the tone unit to be the basic unit of spoken discourse. In particular, "we may speculate, with some plausibility, that the speaker 'plans' the tone unit and the hearer 'decodes' it *as a whole*" (Brazil 1985:12). Another good reason for us is, of course, that the whole of the London-Lund Corpus has been analyzed in terms of tone units.

Outlines of the tagging and parsing systems have been described elsewhere (see Svartvik 1982, Svartvik & Eeg-Olofsson 1982, Eeg-Olofsson & Svartvik 1984). I would now like to report on some other aspects of current and future procedures as a preamble to papers by my colleagues (see contributions by Altenberg, Stenström and Eeg-Olofsson below). Since the Nijmegen conference we have changed from using a mainframe to personal computers, and the parser has been rewritten in Snobol4. The reason for switching to micros is partly financial, partly practical. As for the latter, we envisage speech synthesis to be carried out with a PC, so that it is convenient to have the parser available on the same system.

We have now acquired a Votrax speech synthesizer. There is of course much more sophisticated machinery available on the market but, as it happens, it is not the market of TESS. However, with any machine, there is probably today no way of getting away from the robot-like voice of a synthesizer. The advantage with Votrax is that it is fairly open to manipulation of the parameters of speech rate, pause, pitch (called 'inflection'), and amplitude. Also, we are not going into the phonetics of it but will concentrate on the segmentation into prosodic chunks for which there seem to be no provision made in the available systems.

The assignment of segmentation rules may seem a tall order in view of the fact that there is, clearly, more than syntax involved in speakers' decision to segment or not to segment at any particular place in the discourse. Well, we are not mind-readers and we fully realize the importance of factors beyond the control of a simple syntactic parsing programme. Still, it seems that we have to begin somewhere in order to entangle the mystery of how speakers choose to present their utterances and that such factors as grammar, lexis and speech situation cannot be disregarded. The value of the synthesizer here is that it should provide useful feedback in our attempts to propose segmentation rules.

There is another objection to our procedure which seems decidedly more disturbing: that we are attempting to apply the rules of natural, spoken discourse to written-to-be-spoken discourse. One answer is that, whether we like it or not, synthetic speech is here to stay and we, as linguists, should play some part in trying to improve the quality of it; another that the feedback role of the machine could be a valuable one in providing new linguistic information about various linguistic features of different types of discourse, among which the spoken-written parameter is only one. There seems to be much to learn from a faulty analysis coming out, loud and clear, from a loudspeaker.

References

Brazil, D. 1985. *The communicative value of intonation in English.* Discourse analysis monograph No. 8, English Language Research, University of Birmingham.

Chafe, W.L. 1984. Cognitive constraints on information flow. Berkeley Cognitive Science Report No. 26. Berkeley: Institute of Cognitive Studies, University of California.

Eeg-Olofsson, M. & J. Svartvik. 1984. Four-level tagging of spoken English. *Corpus linguistics*, ed. by J. Aarts & W. Meijs, 53-64. Amsterdam: Rodopi.

Svartvik, J. 1982. The segmentation of impromptu speech. *Impromptu speech: A symposium*, ed. by N. E. Enkvist, 131-145. Publications of the Research Institute of the Åbo Akademi Foundation No. 78. Åbo: Åbo Akademi.

Svartvik, J. & M. Eeg-Olofsson. 1982. Tagging the London-Lund corpus of spoken English. *Computer corpora in English language research*, ed. by S. Johansson, 85-109. Bergen: The Norwegian Computing Centre for the Humanities.

Bengt Altenberg: Speech segmentation in a scripted monologue

A central task for the TESS project is to set up a system of rules that will automatically 'chunk' a written input text into prosodic segments resembling the information units or tone units (TUs) a speaker produces in natural speech. Ideally, such rules will have to satisfy at least the following requirements:

1 they must be based on principles drawn from authentic spoken English, preferably a non-interactive speech style that is somewhat more spontaneous than reading aloud;

2 they must be based on a coherent prosodic and grammatical framework (we use a combination of Crystal 1969 and Quirk et al 1985);

3 though speech segmentation reflects semantic choices, the rules will mainly have to be based on a combination of statistical probabilities and grammatical information produced by the parser;

4 if possible, they should also be sensitive to thematic and pragmatic information;

5 they must have maximal generality (cover all possible cases) and some variability (reflect at least some variation with regard to speed of delivery, communicative purpose, etc).

6 they must have maximal simplicity and efficiency to produce anything approaching real-time speech processing.

The first three requirements are basic. The others are partly in conflict (eg 5 and 6) and perhaps unrealistic (eg 4 and 6), but they are listed here to emphasize the difficulties involved.

There are few systematic studies of speech segmentation that are immediately applicable to automatic text-to-speech conversion. One exception is the fairly detailed model for speech segmentation presented by Crystal (1975:15-22), which assigns TU boundaries to input strings on the basis of their grammatical structure at sentence, clause and phrase level. This model is attractive in several respects (it is empirically founded and uses a grammatical and prosodic framework that is practically identical with our own), but it also has its limitations: it is exclusively based on spontaneous conversation, it is neither exhaustive nor sufficiently explicit, and its rules are too categorical to allow any contextual variation. To be useful for our purposes, it must therefore be extended and supplemented with information from non-interactive speech.

To test the applicability of Crystal's model I have started to examine the principles of TU segmentation in a scripted monologue from the LLC corpus (a public lecture on the life and history of Stoke Poges). Though scripted and delivered at comparatively slow speed, this monologue is partly produced 'off the cuff' and thus represents a semi-spontaneous and fairly 'natural' speech style.

One type of information that can be extracted from this monologue is the statistical distribution of TU boundaries at different points in the clause structure. This information is illuminating in many ways: it tells us, for example, that (at least in this text) TU boundaries are more frequent inside than between clauses, that they tend to set off clause-initial adverbials more often than postverbal elements, that noun phrases are often split internally when coordinated or postmodified, that existential *there* is never separated from a following verb, etc.

However, statistical probabilities, though useful in extreme cases, provide little information about the factors that determine the segmentation in a given context. To find out something about these, it is profitable to look at speech production as a linear process in which the prosodic breaks are related to the distribution of information in discourse. Even if we cannot always predict the location of information foci in a text, there are many interesting correspondences between thematic organization and grammatical structure that can be used as cues for automatic segmentation rules. To identify these correspondences is thus a prerequisite for a satisfactory output.

Exactly what kind of information is needed for automatic TU segmentation, and to what extent this information can be formalized and integrated with the other components of the system (the parser, rules handling tonicity, rhythm, etc) is too early to say. At present, work is in progress to refine and develop the approach outlined here and to test the possibility of incorporating the results in a probabilistic-grammatical framework of the type proposed by Crystal.

References

Crystal, D. 1969. *Prosodic systems and intonation in English.* Cambridge: Cambridge University Press.

Crystal, D. 1975. *The English tone of voice*. London: Arnold.

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

Anna-Brita Stenström:
Pauses in discourse and syntax

This study aims at establishing a preliminary set of optimally predictive rules for pause assignment in synthetic speech. It is based on one text from the London-Lund Corpus of Spoken English, a monologue which consists of a prepared and partly read lecture presented to an audience. The speech rate is slow: 1.7 words/sec, compared with an average speech rate for ten texts of 2.5 words/sec. The total pause rate is consequently high: 6.3 words/pause, 98% of which are silent pauses (SPs) and 2% filled pauses (FPs), compared with an average pause rate of 8.9 words/pause.

I studied the occurrence of pauses between and within tone units (TUs) in relation to the syntactic structure of the monologue, with special emphasis on SPs, and found that the pauses did not only separate syntactic constituents of varying lengths, but also served to mark transitions between paragraphs in the discourse organization.

SPs as discourse markers varied in length from unit to treble (pauses in the LLC texts are 'relative' and marked as 'brief', 'unit', 'double', and 'treble' depending on the individual speaker's actual speed of utterance). The length of pauses reflected the narrative structure of the discourse, in so far as treble SPs separated topic paragraphs, while brief or unit SPs separated prefaces from the immediately following piece of discourse.

SPs as syntactic demarcators were generally brief (63.5%) or unit (27.5%), but rarely longer (9%). The parallelism between pause length and junctures in the syntactic hierarchy was not total. It is true that unit or longer SPs were much more frequent between sentences than between clauses and clause elements, but it is also true that unit SPs occurred more often between clause elements than between clauses and that a much larger percentage of the total number of SPs occurred between clause elements, which indicates that there was no typical clause-by-clause structuring.

Regarding the correlation between TU boundaries with a pause and syntactic junctures, it appeared that sentence junctures were always coterminous with TU boundaries, while this was not always the case with clauses. Pauses between clause elements were found both within and between TUs, while pauses between phrase constituents (single words) always occurred within the TU.

The tentative rules for predicting pause assignment in synthetic speech that were the result of this study will be modified on the basis of a study of additional texts.

Mats Eeg-Olofsson:
Computer processing in the TESS project

File transfer

The project's main machine-readable data base is a subcorpus of the London-Lund Corpus of Spoken English, consisting of 10 texts totalling 50,000 running words, to which word-class tags have been added. Up to now, this corpus has been stored on magnetic tape at the Lund University Computing Centre. After a long period of technical problems, we have finally got access to a version of the KERMIT file transfer program that has enabled us to transfer the subcorpus from the Computing Centre to our own microcomputers.

Database management

The relational data base management system dBASE II has been used extensively to create small databases for special-purpose studies within the project. An updated version of that program, dBASE III, has provided us with the technical possibilities of storing the entire corpus as a single database. In particular, dBASE III can handle up to 10 files at the same time, which makes it possible to access data at all linguistic levels simultaneously. Another useful new feature of dBASE III is the MEMO data type, which can be used to handle large, variable-length text records.

Parsing

An experimental parser, intended to serve the dual purpose of searching and tagging the texts, has been designed as a system of context-sensitive rewrite rules and implemented as a chart parser written in Catspaw SNOBOL4+ on IBM PC XT. Since this SNOBOL implementation turned out to be so slow as to be virtually useless, the parser will eventually be rewritten in TURBO-Pascal.

Karin Aijmer:
Conversational phrases in the London-Lund Corpus

This paper is a report on an ongoing investigation with the aim of identifying and collecting so-called conversational phrases (or pragmatic expressions) in the London-Lund Corpus of Spoken English (LLC). Conversational phrases are expressions of one or several words realizing strategies that are used for certain communicative ends, such as discourse organization, discourse planning and politeness.

The investigation is part of a project carried out in cooperation with the English

Department at the Teachers' Training College in Gothenburg: 'Communicative Competence and the Teaching of English'. The project which is based on an analysis of pragmatic or communicative errors made by Swedish students of English has the aim of developing a model of communication and analytical units which could be used in the teaching of spoken English.

Conversational phrases are important in language teaching. We must however have more information about which they are and how they are used. For that purpose material was collected from the LLC. The collection of material is still going on. At a later stage the material will be further analysed with regard to the categories that may be useful to distinguish from a pedagogical point of view.

Other investigations of conversational phrases have concentrated on a particular lexical form (*you know*, *well*, *I mean*) and described it from various points of view, in particular function or strategy.

The following aspects of the use of conversational phrases will be investigated:

1 The relation between function and form. How is a certain function or strategy realized?

2 The variational aspect. It may be necessary to further constrain a particular discourse function if we want to establish a set of variants that are interchangeable in a communicative situation.

3 Frequency. The investigation of frequency helps us to establish prototypical exponents of a particular class.

4 Intonation. To what extent are conversational phrases associated with a fixed intonation pattern? How much intonational variation is there?

5 Collocation. Collocation is regarded as a salient feature of conversational phrases and part of what the speaker knows about their use. What are the collocations in the corpus?

6 Combinations of conversational phrases. Conversational phrases can combine, resulting in what can be described as contradictions and redundancy. What combinations are possible and how should they be interpreted?

Hedging is one of the discourse functions that has been investigated. It was defined for my purposes as a strategy for 'operating on' a word, an utterance, or a speech act, making it vague or uncertain (cf House & Kasper 1981). Hedges are an interesting class of discourse items simply because there are so many different ways of hedging and because some hedges are very frequent.

In the corpus of informal conversation from the LLC (Svartvik & Quirk 1980) consisting of about 170,000 words there were 263 clause-terminating hedges: 155 *and*-tags (representing 42 types, most frequently *and so on* and *and things*) and 108 *or*-tags (representing 19 types, most frequently *or something*). Other types of hedges with a freer distribution in the utterance included *sort of* (*a sort of*, *sorts of*, *of some sort*), *(a) kind of*, *a type of*, *roughly*, *somehow*, *as it were*, *in so many words*, *in a sense*, *so to speak*, *more of less*, etc.

It is difficult to establish which of the variants that are arrived at in this way are equivalent in the communicative situation. One problem is that there is only a weak tie with semantics. Compare:

He has got his PhD *and everything*
He has got his PhD *and all that sort of thing*

The examples (invented) are semantically (truth-conditionally) equivalent. Their meaning can be described as *and* + the universal quantifier ('all'). The semantic equivalence is not reflected on the discourse level, however. *And everything* 'foregrounds' the information in the first part of the utterance; *and all that sort of thing* signals that the preceding element serves as an illustration only (cf Dines 1980).

A fascinating aspect of the conversational phrases is how they interact with each other and how different devices combine to convey the speaker's intention. An example is:

^oh Y/\ES# it ^/IS a H\OUSE'HOLD god *of S/OME 'sort#*
^\ISN'T it# ^I should !TH\/INK# or ^is it a ':D\ANcer#
I don't ^KN\OW# (1.6.667-672)

Of some sort is a hedge. The other underlined phrases represent different but related strategies.

The speaker can even use devices connected with opposing strategies:

you ^KN/OW# - it's *^just sort of* . :one W/ORD# . [e?]
^every about !ten "!S\ECONDS {^coming \OUT#}#
(3.2.129-131)

The same assertion contains both *just* and *sort of* (cf Aijmer 1985). The effect of *just* is to strengthen the force of the assertion ('I declare emphatically that'). The use of the emphatic form fulfils a social purpose: the speaker shows solidarity and camaraderie. *Sort of* expresses another conversational strategy. It makes the following expressions less precise, thus weakening the force of the assertion on the listener.

References

Aijmer, K. 1985. Just. *Papers on language and literature presented to Alvar Ellegård and Erik Frykman*, ed. by S. Bäckman and G. Kjellmer, 1-10. Gothenburg Studies in English 60. Gothenburg: Acta Universitatis Gothoburgensis.

Dines, E.R. 1980. Variation in discourse - 'and stuff like that'. *Language in society* 9:13-33.

House, J. & G. Kasper. 1981. Politeness markers in English and German. *Conversational routine*, ed. by F. Coulmas, 157- 185. The Hague: Mouton.

Svartvik, J. & R. Quirk. 1980. *A corpus of English conversation*. Lund Studies in English 56. Lund: Gleerup.

8 NEW CORPORA OF AMERICAN ENGLISH

Wallace Chafe
University of California, Berkeley

Options for the archiving of spoken and written data

We are presently engaged in setting up a language archive at Berkeley, to include materials collected by Susan Ervin-Tripp, Lily Fillmore, John Gumperz, Dan Slobin, myself, and others. I will discuss here some of the considerations that have seemed relevant to the establishment of such an archive.

A language archive may serve various purposes. We have been concerned with increasing the availability of data for studies of various kinds: of syntax and discourse; of semantics and pragmatics; of cognitive linguistics and sociolinguistics; of prosody; of oral literature; and so on. The data already collected by the people mentioned above were intended to shed light on such diverse areas as social interaction among adults or in school classrooms, language acquisition, cross-cultural differences in language use, and differences between speaking and writing. We want to facilitate access to data that are relevant to these and other purposes and also, and especially importantly, to preserve records of the present state of the language for future generations.

Our primary target will be American English, but we do not intend to exclude other dialects of English or other languages, as future interests may dictate. Although we have begun with an emphasis on spoken language, we intend to include samples of written language as well. We may eventually try to broaden the coverage to a wide sampling of many genres of both speaking and writing.

In collecting these data, we have in mind the importance of naturalness - of representing language as it is actually used in real situations. In part, therefore, we see the archive as providing a useful corrective to the more artificial kinds of language that are frequently used in syntactic and experimental studies.

We have been confronted with various practical aspects of archiving, to which we do not necessarily have acceptable solutions at the present time. There are practical questions of cataloguing that all archivists face, and other questions that are raised by the special nature of these materials.

All of the spoken materials will have sound recordings as at least one component. In many cases there will be video recordings as well. One needs, of course, to make working copies of the tapes, storing master copies under conditions that will assure their permanent availability. Magnetic storage, however, presents serious problems in this regard, and we may hope that the technology of audio and video storage will sooner or later provide us with more permanent and foolproof options. In the meantime we will have to rely on magnetic tape, extending its life with a controlled environment and periodic rewinding.

We have given much thought to cataloguing, and particularly to the kind of documentation that will be maximally useful to present and future scholars. For each archive entry we will provide several kinds of information, accessible through a computerized catalogue: a descriptive title; a record of the place and date of collection, and the length of the sample; a description of the participants and the situation (whether a dinner table conversation, a job interview, a lecture, etc); something about the genre (whether a narrative, an explanation, an argument, etc.); any conversational or prosodic features that may be noteworthy; remarks on the quality of the recording; and an indication of whether and by whom the tape was transcribed, and the conventions used.

The archive will also include these transcriptions, when they are available. Transcribing is a process that can be applied to the recordings as time and personnel permit. As we all know, transcriptions are always selective, and much remains to be learned regarding formats that are optimal for various purposes. We would like to arrive at a more or less standard format that captures much of the information necessary for most studies, realizing that the original recordings will remain available for those who may wish to submit them to further analysis - for example in detailed studies of intonation. In general, we plan to create transcriptions that capture some of the essential information, hesitations, volume, tempo, and voice quality, without necessarily including everything that might be included in these areas.

Since we are aiming at computer storage, not only of the catalogue but also of the transcriptions, we are somewhat limited with respect to fonts and graphic displays. But the variety of presentations that can be manipulated and retrieved on a computer screen, as well as on printed copy, is constantly improving.

One question of interest is the value of representing written language in something other than its standard form. For example, it seems clear that writing has a covert prosody that is assigned to it by writers and readers, and only partially represented through punctuation. Further research may suggest ways of dividing written language into units resembling the 'tone units' of speech, and thereby facilitating comparisons between written and spoken language.

The kinds of analysis that may be applied to these data are of course limited only by the creativity of present and future researchers. Certainly one obvious step beyond transcribing is the 'tagging' or 'coding' of material. What particular codes are assigned, and even how they are formatted with respect to the transcriptions, needs to remain flexible to accommodate the interests of different scholars. We expect, however, to make available various coding schemes, suitable for different purposes.

Once codes have been assigned, they can be tabulated and subjected to statistical analysis. Again, what is done here needs to be left open to the researcher. Nevertheless, it will be useful to provide various easily applied options for analysis, and we see the provision of relevant statistical programs as another service the archive can provide.

Here, as in data collecting, storage, retrieval, and transcription, our goal will be to provide options that will be maximally useful to ourselves, our collaborators, and those who come after us. We are much interested in exchanging views and experience with others in Europe and America who may be involved in similar enterprises.

Gunnel Tottie
Uppsala University

A corpus of spoken American English - a panel discussion

The background of the panel discussion was the need for a generally available corpus of spoken American English to complete the present set of computer corpora consisting of Brown, LOB, and LLC. Moreover, the establishment of the new Institute for American Studies at Uppsala University made it seem desirable that the English Department at Uppsala should be instrumental in collecting at least part of a new American Corpus. However, as it is not at all clear how such a new American corpus should be structured, a panel was convened to discuss the matter. The members of the panel were Professor John Algeo, the University of Georgia, Professor W. Nelson Francis, Brown University, Professor Wallace Chafe, the University of California at Berkeley, Professor Edward Finegan, the University of Southern California, and Gunnel Tottie, chair, of the University of Uppsala, Sweden. The following are the most important questions that were submitted to the panel and the general membership of ICAME, with some of the reactions of panelists and others.

A *Material*

1) Do we need an entirely new corpus of spoken American English, or should we try to make use of existing material, either collected by other scholars for research purposes or of other kinds (eg FBI transcripts)? Assuming that there is material available, how do we make an inventory and get permission to use it?

John Algeo and W. Nelson Francis pointed to the existence of material in American dialect archives, especially that in the University of Wisconsin. This material has a varied composition, comprising conversational material in addition to narratives, reading passages, and interviews. The material has the advantage of being readily available without copyright problems.

Wallace Chafe announced the creation of an archive of spoken American English at Berkeley (see his report above), comprising recordings made by John Gumperz and Susan Ervin-Tripp, the fifty hours of dinner-table conversation collected by

Chafe, as well as material collected by other reseachers. The availability of material and surveys of existing material at the Center for Applied Linguistics was also signalled. Jan Svartvik mentioned the availability of network English on audiotape.

2) Assuming that we decide to make a new corpus, what kind of corpus do we want? Should we aim for as close a counterpart to the LLC as possible, as we have in Brown and LOB, or do we want to include other types of material as well, other types of speakers, etc?

The general feeling was that one should not strive for an American counterpart of the LLC. W. Nelson Francis advocated the inclusion of regional varieties, and Wallace Chafe suggested including the language of blue-collar workers.

3) Much of the London-Lund material was recorded surreptitiously. Is such a procedure desirable and feasible with regard to American English?

It was unanimously decided that non-surreptitious recording with visible microphones would be the only practicable method of collecting a corpus of American English.

4) All of the above questions have to do with comparability with the LLC material. Another problem is the time factor. If we make new recordings in the late eighties, how comparable will they be to the LLC material from the sixties and early seventies?

Again, there was general consensus that comparability was not of the essence.

B *Transcription and related questions*

1) What level of delicacy is desirable for the transcription? Should the whole material be provided with prosodic analysis, or could part of it be transcribed in conventional orthography only?
2) Which model of prosodic analysis should be chosen? If the LLC model is chosen, which version is to be selected, the original one or the simplified version used in the Svartvik & Quirk book edition?
3) Should any part of the corpus be submitted to instrumental analysis?
4) Should the corpus or part of it be made available in the original soundtrack version?

It was generally felt that the detailed prosodic analysis of the LLC should not be emulated. Jan Svartvik suggested that parts of the material could be analyzed at different levels of delicacy, with most of it rendered in normal orthography. Geoffrey Leech proposed a rudimentary prosodic analysis. John Sinclair underlined the necessity of making available 'clean' texts for eg lexical study, and pointed to the possibility of providing interlinear marking of prosody.

It was suggested that new American material should be made available in parallel versions, ie transcriptions accompanied by soundtracks, and that this would render prosodic analysis unnecessary. Instrumental analysis would also be superfluous. Individual researchers would be able to make the kind of analysis required for their own purposes.

John Sinclair emphasized that transcription should take place at the word-processor, and that standard typefaces for optical scanning should be chosen.

9 STYLISTIC STUDIES

Douglas Biber and Edward Finegan
University of Southern California

Uncovering dimensions of linguistic variation in English: A research report

This paper describes the methodology and major findings of a research project that seeks to identify and characterize the dimensions of linguistic variation among texts in English. We define *dimensions* as underlying functional parameters of linguistic variation among texts. Each dimension comprises a group of linguistic features (eg passives, nominalizations, prepositional phrases) that cooccur with a markedly high frequency in texts. Factor analysis is used to identify these groupings of features, and the combinations of cooccurring linguistic features are taken to define different speech styles (cf Ervin-Trip 1971; Hymes 1972).

Consideration of the contextual characteristics of the speech styles (ie the contextual differences between those texts having markedly high and markedly low frequencies of the cooccurring features defining a dimension) enables us to assign an interpretation of each dimension. In this way, the situational and psycholinguistic parameters associated with a dimension are explored.

The resulting dimensions are used to define *relations* among texts. Each text is situated along each dimension according to its exploitation of the defining features of the dimension. By simultaneously considering the positions of texts along all dimensions, the overall linguistic similarities and differences - the textual relations - within a set of texts can be defined.

A global model of linguistic variation among texts cannot be constructed from analysis of a few texts or text types. Hence, this project utilizes standard computerized corpora as input, computer programs for identifying features, and multivariate statistical techniques for analysis. Standardized corpora, in which texts have been chosen from a wide sampling of text types, help to ensure the wide range of linguistic variation that must be accounted for in a model of textual variation. Computerized corpora provide ready access to a large number of texts, enabling analysis on a scale not feasible by hand counts. Multivariate analysis enables quantitative description of the relation among these texts.

Computer programs written in PL/1 or Pascal are used to identify textual features automatically. So as to avoid difficulties of ensuring comparability across tagged and untagged corpora, only grammatically untagged versions of the Brown, LOB and LLC corpora are used as input. The algorithms for feature identification are written to capture as many tokens of a construction as possible without skewing from one text type to another. (A description of the algorithms is available from the authors.)

Thus far we have investigated relations among oral and written text types, among texts having high and low sociolinguistic prestige, among various genres of fiction, and among British and American written texts. We have discovered that the relations among text types are complex and that no single dimension adequately captures the similarities and differences among text types. We have consistently found that a multi-dimensional model is required (Biber 1985; Finegan & Biber forthcoming (a); Finegan & Biber forthcoming (b); Biber forthcoming).

To date, we have identified three primary dimensions defining linguistic variation among texts in English. To reflect their functional content, we have tentatively labeled these dimensions as follows:

Dimension I	Interactive vs Edited Text
Dimension II	Abstract vs Situated Content
Dimension III	Reported vs Immediate Style

Dimension I is characterized linguistically by features like questions and first and second person pronouns vs, for instance, word length. Dimension II is characterized by features like nominalization and passives vs, for instance, place and time adverbs. Dimension III is characterized by past tense vs present tense features.

Our multi-dimensional model can be illustrated by a consideration of the relations among Academic Prose, Professional Letters, Broadcast, and Conversation along two dimensions. With respect to Dimension I (Interactive vs Edited Text), Conversation and Academic Prose are at opposite extremes; Conversation is characterized as highly 'interactive' and not highly 'edited'; Academic Prose is highly 'edited' but not highly 'interactive'. These characterizations are precisely quantifiable. Along this same dimension, Professional Letters, though written, are more similar to Conversation than to Academic Prose, while Broadcast, though spoken, is more similar to Academic Prose than to Conversation.

With respect to Dimension II (Abstract vs Situated Content), the relations among these four text types are somewhat different. Conversation and Academic Prose are again at opposite poles: Conversation highly 'situated', Academic Prose highly 'abstract'. Contrary to their positions with respect to one another along Dimension I, however, Broadcast is very similar to Conversation: both are highly 'situated'. Similarly, both Professional Letters and Academic Prose are highly 'abstract'.

While consideration of the distribution of texts along any dimension is informative, a fuller picture of the relations among these four text types results from a joint consideration of Dimensions I and II: Conversation is 'interactive' and 'situated'; Professional Letters is 'interactive' and 'abstract'; Broadcast is 'situated' but not markedly 'interactive' or markedly 'edited'; Academic Prose is 'edited' and 'abstract'. Analysis of the positions of all text types along all three dimensions enables a first approximation of a model of textual relations in English.

Earlier work in this project took for granted the basic validity of the text type categorizations, or genres, proposed by the compilers (human!) of the computerized corpora. Texts within each labeled genre were assumed to share characteristics, and genres were regarded as validly distinct from one another. Some of our research called these assumptions into question, however. We found, for example, greater differences among texts within the genres of fiction than across them. As a result, we undertook to identify the text types, or *speech styles*, that are in fact well defined in terms of the previously uncovered dimensions. We propose defining distinct speech styles as groups of texts that are uniformly characterized by the frequent occurrence of functionally related sets of linguistic features.

As a first step towards the identification of the speech styles of English, we investigated the marking of stance in our texts. By *stance*, we mean the overt expression of a writer's or speaker's attitudes, feelings, or frames of reference. We limited ourselves to the adverbial marking of stance in this part of the project. The attitudinal and style disjuncts listed in Quirk et al (1985) served as potential markers of stance, and all occurrences of these adverbials were identified in the LOB and LLC corpora. Using a KWIC listing, we analyzed each adverbial in context to distinguish true markers of stance from adverbials of the same form serving other functions. The adverbials marking stance were divided into six functional categories, and the frequency of occurrence for each category in each text was computed. The six stance categories represent expressions of (1) manner of speaking; (2) generalization; (3) conviction; (4) doubt; (5) assertion of reality; (6) attitude towards content.

Using a statistical technique called cluster analysis, texts that are maximally similar were grouped into clusters on the basis of their exploitation of stance adverbials. We interpreted each cluster by consideration of the characteristic linguistic features and situational contexts of the texts constituting that cluster. Among the eight styles identified were *Faceless*, *Confident/Dogmatic*, and *Cautious*.

While the preliminary analysis of stance presented here invites detailed consideration of the individual texts in each cluster, the findings to date illustrate the potential for empirically grouping texts on the basis of their exploitation of linguistic features rather than on any a priori basis. The resultant groupings are internally coherent while being maximally distinct from one another; they thus provide a firm empirical foundation for the identification and understanding of basic speech styles in English.

References

Biber, D. 1985. Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics* 23:155-178.

Biber, D. Forthcoming. Spoken and written textual dimensions in English: resolving the contradictory findings. To appear in *Language* 62, No. 2.

Ervin-Tripp, S.M. 1971. Sociolinguistics. *Advances in the sociology of language* I, ed by J.A. Fishman, 15-91. The Hague: Mouton. (Reprinted from *Advances in experimental social psychology* 4 (1969), ed by L. Berkowitz, 91-165. New York: Academic Press.)

Finegan, E. & D. Biber. Forthcoming (a). Toward a unified model of sociolinguistic prestige. To appear in *Proceedings of NWAVE* XII (tentative title), ed by D. Sankoff. Washington, D.C.: Center for Applied Linguistics, and New York: Harcourt Brace Jovanovich.

Finegan, E. & D. Biber. Forthcoming (b). Toward a multi-dimensional model of linguistic complexity. To appear in *Southern California Occasional Papers in Linguistics* 11.

Hymes, D. 1972. *Foundations in sociolinguistics: an ethnographic approach.* Philadelphia: University of Pennsylvania Press.

Cheng Yumin
Fudan University, Shanghai

An attempt at analysing linguistic style

Apart from having geographical and social differentiations, a language also includes a set of stylistic varieties whose use is determined by the situational context.

In order to give the stylistic analysis a tangible form, an attempt was made to apply a unified model to stylistically relevant features in such a way that an indexed measurement of each of them is possible.

Stylistically relevant features are those linguistic elements that form socially and contextually relevant groups of 'synonymous' variants. The stylistic differences between these variants constitute the substance of stylistic differentiation. Yet, stylistically differentiated as they are, they carry the same message in their respective contexts.

The concept of 'fullness of expression' is introduced to unite under it all the different kinds of stylistic substance.

A stylistical rating of (+1) or (-1) is assigned to each of those linguistic features which, in a particular context, apparently play a role either in raising the style upward (+1) or dragging it downward (-1), while the majority of linguistic items are considered stylistically neutral and not included in the count. The basic idea is that any instance of language use remains neutral until stylistically marked features begin to influence it by moving it in the upward or downward direction. This scale of stylistic differentiation, as distinguished from 'professional' varieties, represents the stylistic command that all speakers of English possess in varying degrees.

A working list of 32 stylistic markers, all of a very general nature, was prepared to provide a basis for a quantitative study of stylistic features in various texts.

Experiments were made with different texts by applying the stylistic markers. The results showed that an investigation of the stylistic markers may help to reveal the stylistic nature of a text.

A further investigation of samples of language use by Professor W. Diver of Columbia University showed that the + and - markers correlate with the situational context.

The investigation suggests: (1) stylistic differentiations represent subdivisions of language varieties, as is implied in Labov's investigation of New York English; (2) style may be envisaged as a scale along which texts are differentiated.

10 SOFTWARE DEVELOPMENT

Petr Sgall
Charles University, Prague

Experiments with machine translation and question-answering in Prague

On the basis of a theoretical approach towards linguistic description, using dependency syntax and levels of representation ordered from meaning (underlying structure) to sound and graphemics (see Sgall et al, in press), the linguistic research group of the Faculty of Mathematics and Physics, Charles University, has prepared two experimental systems for automatic comprehension of natural language.

1 The APAC system (Automatic Translation from English to Czech) is designed to translate texts in electronics. Grammatical rules for a syntactic-semantic analysis of English and for a synthesis of Czech were formulated by Kirschner (1982) in a rather detailed way in Colmerauer's Q-language. Several hundreds of rules handled morphemics, the structure of the noun phrase, verbal valency, coordination, apposition, several types of embedded clauses, constructions with infinitives, participles and gerunds, as well as the topic-focus articulation. The main dictionary comprises only about 1000 words, but a transducing algorithm translates thousands of international technical terms, changing their orthographic form and their end segments, so that also many newly coined words can be handled.

2 TIBAQ (Text-and-Inference-Based Answering of Questions) is a linguistically oriented method suitable for a natural-language front-end system of contact with an expert system, an intelligent robot, or for an automatic compilation of a knowledge base from an input text and for question-answering systems over such knowledge bases. The first experiment concerns two short pieces of Czech texts on electronics.

The main procedures of the system are:

a) a syntactic-semantic analysis (see Panevová & Sgall 1980; Panevová & Olivia 1982), by means of which Czech sentences are transduced to representations of their meanings, based on dependency graphs and specifying whether a given lexical occurrence functions as an Actor, Addressee, Objective, Instrument, Manner adverbial, Time interval, Cause, Condition, and so on, as well as the values of the morphological categories (number, definiteness, tense, aspect, modality, etc); moreover, the topic and the focus of each sentence, clause and syntagm are identified; also the questions formulated by users undergo this analysis, and their meanings are then compared with a concordance compiled from the stock of assertions (representations of meanings gained from the input texts), so that a set of relevant assertions is delimited (ie those assertions that have autosemantic occurrences in common with the question);

b) rules of inference (ranging from general rules concerning conjunction and disjunction to specific, linguistically based rules extracting *why*-clauses and relative clauses under specific conditions, conjoining two assertions with the same Actor, etc), which are defined on the set of meaning representations; also definitions are made use of by this procedure;

c) the procedure of the choice of answer checks whether a complete answer to the given question has been found (ie an assertion that fully corresponds to the question, including specific correspondence rules for the articulation of topic and focus, and also some equivalence classes for the syntactic and morphological values), or whether just a partial answer can be given (eg concerning some more or less general information than that requested by the question);

d) the procedure of synthesis, transducing the assertion(s) chosen as answer(s) into the outer form of Czech sentences.

The first experiment with question-answering on the basis of the TIBAQ method was successfully carried out in 1984; its main ingredients are described by Hajicová & Sgall (1981).

References

Hajicová, E. & P. Sgall. 1981. Text-and-inference based answering of questions. *Prague Bullentin of Mathematical Linguistics* 36:5-23; completed in Sgall (1984:291-320).

Kirschner, Z. 1982. A dependency-based analysis of English for the purpose of machine translation. *Explizite Beschreibung der Sprache und automatische Textbearbeitung* 9. Prague: Matematicko-fyzikální fakulta UK.

Panevová, J. & K. Oliva. 1982. On the use of Q-language for syntactic analysis of Czech. *Explizite Beschreibung der Sprache und automatische Textbearbeitung* 8:108-117. Prague: Matematicko-fyzikální fakulta UK.

Panevová, J. & P. Sgall. 1980. On some issues of syntactic analysis of Czech. *Prague Bulletin of Mathematical Linguistics* 34:11-32.

Sgall, P. (ed). 1984. *Contributions to functional syntax, semantics, and language comprehension*. Prague: Academia, and Amsterdam/Philadelphia: Benjamins.

Sgall, J., E. Hajicová & J. Panevová. In press. *The meaning of the sentence in its semantic and pragmatic aspects*. Prague: Academia, and Dordrecht: Reidel.

Val Jones
University of Stirling

A query language for semantic network databases

Methodology

The software design methodology being developed by researchers at Stirling University in collaboration with ICL SETC at Kidsgrove as part of Alvey SE project 029 involves the following steps:

1 devise abstract objects
2 devise abstract operations upon objects
3 devise representation of objects
4 design operations using the executable formal specification language 'me too' (Henderson 1985)
5 build prototype
6 iterate

Semantic network databases

Relations between objects can be modelled by semantic networks, where labelled nodes are linked by labelled directed arcs.

A database recording information about courses is used as an example. Nodes of the type 'students', 'courses', 'periods', and 'lecturers' are linked by the relations 'attends', 'timetabled' and 'teaches', eg:

```
{Jane} attends {archaeology, welding}
{archaeology} timetabled {period2}
{Grimm} teaches {embroidery, welding}
```

Netcalc operations

Image and inverse image operations can be used to query the database:

```
im({Jane}, attends, db)
```

returns {archaeology, welding}

```
inv(teaches, {welding}, db)
```

returns {Grimm}

Netcalc expressions may be nested to arbitrary depth. Whilst giving powerful querying facilities, this may make expressions difficult to construct and understand.

Explain

Explain is a collection of operations which give an English explanation of a Netcalc expression. Operations are specified in 'me too' and input to the 'me too' preprocessor which translates them into LispKit source code (Henderson, Jones & Jones 1983) which is then compiled. The 'signature' of the operation is:

```
explain : Nq x Db -> Text
```

An interaction with Explain follows:

```
nq = inv(attends, im({Grimm}, teaches, db), db)
explain(nq, db)
(students that attend courses taught by Grimm)
```

The reverse process might also be useful, ie an operation that accepts a query in English, maps it into the corresponding Netcalc query, and evaluates it.

Nialpxe

Nialpxe allows the user to query the database using (a very restricted subset of) English. The top level operation of the Nialpxe program is 'giveme':

```
giveme : Text x Db -> set(Na)
```

Text is a query in English, Na is the answer to the corresponding Netcalc query. An interaction with Nialpxe follows.

```
giveme((students that attend courses during period2), db)
{{David, Sam, Jane, Stuart, Arthur}}
```

A text has three components, and is recursively defined:

```
Text -> Nodesettype Reltext Text | Node
```

These first prototypes of Explain and Nialpxe do not perform any linguistic processing. They rely on meta-information stored in the database. Nodes are already stored; we must also record the type of each node, and some text associated with each relation. The latter will differ, depending on whether we are looking forward along the arc (an 'im' expression) or back (an 'inv' expression). For the 'attends' relation, we add the meta-information:

```
{attends} domain {students}
{attends} range {courses}
{attends} imreltext {attended by}
{attends} invreltext {that attend}
```

For each node, we record its type:

```
{Jane} type {students}
```

The grammar of the 'English' query language is:

```
Text -> Nodesettype Reltext Text | Node
Nodesettype -> courses | students | lecturers | periods
Reltext -> Imreltext | Invreltext
Imreltext -> attended by | taught by | occupied by
Invreltext -> that attend | that teach | during
Node -> embroidery | archaeology | ...
  | Jane | Sam | ...
  | Grimm | Heep | ...
  | period1 | period2 | ...
```

Nialpxe parses the input text, producing a set of valid parses which are passed to the translate operation. The translate operation returns a pair for each valid parse, consisting of the literal Netcalc query Nq, and its evaluation Na (the answer to the query). Giveme picks out the answer part of each pair. The 'me too' specification of 'giveme' is:

```
giveme(q,db) ≡
{2(dbqa) | dbqa <- translate(getvalidparses(parse(q,db)),db)}
```

Future extensions

The first prototypes of Explain and Nialpxe are trivial programs designed to demonstrate the use of the methodology. No linguistic processing is involved, and the 'English' interface is as sytactically restricted and inflexible as the Netcalc interface. However, the prototypes serve as a basis for more interesting projects. Three different approaches are currently under way:

1 With an appropriate lexicon (containing morphological variants), the text to be associated with the relations can be automatically generated rather than stored explicitly.

2 The parse operation can be replaced by a proper parsing algorithm, parameterised with respect to a grammar. The interface builder may then define a query language of arbitrary linguistic complexity.

3 Certain kinds of grammar can be generated automatically from the contents of the database. A program called Induce has already been designed and prototyped using the methodology.

It is hoped that the future extensions suggested above will begin to exercise the software design methodology in relation to natural language processing proper.

References

Henderson, P. 1985. 'me too' - a language for software specification and model building. Preliminary report, Stirling University Computing Science Department FPN-9.

Henderson, P., G.A. Jones & S.B. Jones. 1983. The LipsKit manual. Oxford University OUCS Technical Monograph PRG-32.

Eric Steven Atwell
Leeds University

Software tools for English language analysis

I attended previous ICAME conferences as a member of the UCREL research team from Lancaster University; but I have recently moved from Lancaster to Leeds University, along with Geoffrey Sampson, another collaborator on the UCREL research team. We are keeping in touch with the Lancaster group, and we have also become involved in linguistic computing research at Leeds. I have been investigating various software tools which could be useful in corpus-based research and linguistic computing generally. National and international initiatives such as the Alvey programme in the UK and the ESPRIT programme in the European Community aim to foster collaboration in research, and to promote the principle of a *common base* of research tools and facilities available to (and used by) all researchers in a particular field. ICAME should consider the idea of adopting a common base of standard software and other facilities for corpus-based research, so we can readily exchange useful software.

Lexicographical databases

Researchers at several sites (including Leeds) have developed or are developing lexical databases of different kinds, using English dictionaries in machine-readable form (such as LDOCE, OALDCE, Collins English Dictionary) together with data extracted from a corpus or corpora. We should exchange information on the detailed structure and contents of these lexical databases, with a view to agreeing on a generalised standard structure, and standard dictionary search and retrieval routines.

Programming languages

In the many research reports published in ICAME News and elsewhere, very little mention is made of the programming language(s) used in corpus-based research. If the various corpus researchers could adopt a common programming language, we could exchange programs, subroutines, etc, much more readily, and cut down on reduplication of effort. I suspect that often the programming language used on a project is chosen because it happens to be readily available and/or is favoured by the Computer Service, rather than because it has been objectively evaluated and shown to be particularly suited to the task in hand.

General-purpose procedural languages: BASIC, PASCAL, ADA

Several research teams have opted for widely-available general-purpose procedural languages such as Basic or Pascal. However, Basic in particular comes in many different versions and dialects, so interchange of software would be difficult even if we all used some version of Basic. Pascal is more standardized, but it has very poor facilities for representing and manipulating strings, lists, and trees. Various institutions (particularly the US and NATO Defence Departments) are pressing for Ada as a standard successor for all their current procedural programming languages; unfortunately, as the Ada language is very rich and complex, Ada compilers are slow, and Ada also has poor string-, list- and tree-handling capabilities.

LISP, PROLOG

Much research in artificial intelligence and computational linguistics is done using the functional language Lisp or the logic language Prolog. These have the advantage of straightforward string- and list- (ie tree-) processing facilities, and also the ability to evaluate a string or list as a piece of program code. Prolog in particular encourages the *declarative* style of programming: a 'program' is simply a list of rules to be applied in the production of the desired output, and the sequential organization of processing is dealt with 'behind the scene' by the Prolog interpreter. This reduces the time taken by the programmer to write programs; but unfortunately programs tend to rely on recursion and backtracking, which can make execution inefficient and slow. As corpus-based research typically involves wading through very large datafiles, declarative/functional languages may be inappropriate for fast 'low-level' processing.

Icon, Pop-11 and POPLOG

Icon and Pop-11 are procedural languages with extra features. Icon has predefined string, character-set, and list data types, and a backtracking facility if required (although the language can be used purely procedurally). Pop-11 also has string and list-processing facilities and evaluation functions; furthermore, it uses an incremental compiler, which allows programs to use a large library of utility procedures. Pop-11 comes as part of an integrated programming environment, POPLOG, which includes a powerful screen editor, VED. VED is in fact simply another library routine written in Pop-11, and can be altered to taste by the programmer. The complete POPLOG environment also includes Lisp and Prolog interpreters and compilers (also written in Pop-11), allowing code in any combination of the three languages to be intermixed; for example, it may be convenient to state the overall processing strategy declaratively in Prolog, but to write potentially time-consuming low-level functions such as input-output in Pop-11, thus getting the best of both languages. This freedom of the programmer to 'mix and match' has led to increasing use of POPLOG as an artificial intelligence research and development environment. Introductions to Lisp, Prolog, Pop-11, and POPLOG are included in O'Shea & Eisenstadt (1984).

UNIX: A common base operating system

It may be difficult for us all to agree on a common programming language, and to convert existing programs to one language is a very big job. However, it is still desirable for us to be able to use each other's software with the minimum of conversion problems, and one way of achieving this would be to adopt a common operating system. The artificial intelligence and computing research community in general seem to be thinking this way; many people are putting forward UNIX as a common base operating system. This has several advantages:

1) Although the UNIX trademark is owned by Bell Labs, UNIX is available on a wide range of machines from most manufacturers, and a wide range of languages and tools have already been developed on it; there is also plenty of introductory literature available, eg Bourne (1982).
2) UNIX is particularly well suited to the integration of tools, since processes can be *pipelined* together to run concurrently, passing results from one process as input to the next automatically.
3) Many tools useful to corpus researchers come built in with the standard system, eg PARTS and STYLE for part-of-speech and stylistic analysis of English texts; GREP and AWK for pattern-searching and concordancing; LEX and YACC parser-generator tools to automatically convert a rewrite-rule grammar into a parsing program; SORT to sort files; WC to count words/letters/lines, etc.

I therefore encourage ICAME participants who are considering future computing research requirements to adopt the UNIX standard.

References

Bourne, S. 1982. *The UNIX system*. Addison-Wesley.

O'Shea, T. & M. Eisenstadt. 1984. *Artificial intelligence: tools, techniques, and applications*. Harper and Row.

ICAME BIBLIOGRAPHY

The following list contains works based on or related to computerized English corpora, with special emphasis on the Brown Corpus, the Lancaster-Oslo/Bergen Corpus, and the London-Lund Corpus. Where possible, we have added a code specifying which material has been used or referred to. Abbreviations:

BCE = The Birmingham Collection of English Text
BUC = The Brown University Corpus of American English
LEU = The Leuven Drama Corpus
LLC = The London-Lund Corpus of Spoken English
LOB = The Lancaster-Oslo/Bergen Corpus of British English

Some early works based on spoken material from the Survey of English Usage, University College London, have also been included. Since it is unrealistic to distinguish this material from its computerized descendant, the LLC Corpus, both have been given the same code [LLC].

* * *

Aarts, J. 1984. The description of the English language. *English language research: The Dutch contribution*, I, ed J. Lachlan Mackenzie & H. Wekker, 13-32. Amsterdam: Free University Press.

Aarts, J. 1984. The LDB: A linguistic data base. *ICAME News* 8: 25-30.

Aarts, J. & T. van den Heuvel. 1980. The Dutch Computer Corpus Pilot Project. *ICAME News* 4: 1-8.

Aarts, J. & T. van den Heuvel. 1982. Grammars and intuitions in corpus linguistics. In Johansson (ed) 1982: 66-84.

Aarts, J. & T. van den Heuvel. 1983. Corpus-based syntax studies. *Gramma* 7: 153-173.

Aarts, J. & T. van den Heuvel. 1984. Linguistic and computational aspects of corpus research. In Aarts & Meijs (eds) 1984: 83-94.

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

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

Aarts, J. & W. Meijs (eds). Forthcoming. *Corpus linguistics* II.

Aijmer, K. 1983. Emotional adjectives in English. *Papers from the 7th Scandinavian Conference of Linguistics*, ed. F. Karlsson, 199-220. Department of Linguistics, Helsinki University. [BUC, LOB]

Aijmer, K. 1984. *Go to* and *will* in spoken English. In Ringbom & Rissanen (eds) 1984: 141-157. [LLC]

Aijmer, K. 1984. *Sort of* and *kind of* in English conversation. *Studia Linguistica* 38: 118-128. [LLC]

Aijmer, K. 1985. *Just*. In Bäckman & Kjellmer (eds) 1985: 1-10. [LLC]

Aijmer, K. 1985. What happens at the end of our utterances? The use of utterance-final tags introduced by *and* and *or*. *Papers from the Eighth Scandinavian Conference of Linguistics*, ed. O. Togeby, 117-127. Institut for Nordisk Filologi, University of Copenhagen. [LLC]

Aijmer, K. Forthcoming. Direct and indirect speech in different styles of English. To appear in *Papers from the Tenth Scandinavian Conference of Linguistics*. [LLC, LOB]

Ajmer, K. Forthcoming. Why is *actually* so frequent in spoken English? To appear in Tottie & Bäcklund (eds), forthcoming. [LLC]

Akkerman, E. 1984. Verb and particle combinations: Particle frequency ratings and idiomaticity. *ICAME News* 8: 60-70. [BUC]

Altenberg, B. 1984. Causal linking in spoken and written English. *Studia Linguistica* 38: 20-69. [LLC, LOB]

Altenberg, B. 1984. Lexical and sex-related differences in spoken English: Some results of undergraduate research at Lund University. In Ringbom & Rissanen (eds) 1984: 279-298. [LLC, LOB]

Altenberg, B. 1984. Speech rate and tone-unit length in ten spoken texts. TESS Report, Survey of Spoken English, Lund University. [LLC]

Altenberg, B. 1985. Correlations between prosody and word classes in a spoken monologue. TESS Report, Survey of Spoken English, Lund University. [LLC]

Altenberg, B. 1985. Prosodic patterns in a prepared monologue. TESS Report, Survey of Spoken English, Lund University. [LLC]

Altenberg, B. 1985. Towards a probabilistic onset rule. TESS Report, Survey of Spoken English, Lund University. [LLC]

Altenberg, B. Forthcoming. Causal ordering strategies in English conversation. To appear in *Perspectives in discourse analysis*, ed. J. Monaghan. [LLC, LOB]

Altenberg, B. Forthcoming. Contrastive linking in spoken and written English. To appear in Tottie & Bäcklund (eds), forthcoming. [LLC, LOB]

Altenberg, B. & G. Tottie. 1984. Will there be texts in this class? Writing term papers within a research project. In Ringbom & Rissanen (eds) 1984: 265-278. [LLC, LOB]

André, E. 1974. Studies in the correspondence between English intonation and the noun-phrase in English grammar. Liège: Université de Liège. [LLC]

André, E. 1975. English tone contrasts in relative and adverbial clauses and in enumerative statements. *Revue de Phonétique Appliquée* 35: 107-118. [LLC]

Atwell, E. 1982. LOB Corpus tagging project: Manual postedit handbook. Department of Linguistics and Modern English Language and the Department of Computer Studies, University of Lancaster. [LOB]

Atwell, E. 1983. Constituent likelihood grammar. *ICAME News* 7: 34-67. [LOB]

Atwell, E. 1986. Beyond the micro: Software tools for research and teaching from computer science and artificial intelligence. *Computers in English language teaching and research*, ed. G. Leech and C. N. Candlin, 168-183. London: Longman.

Atwell, E., G. Leech & R. Garside. 1984. Analysis of the LOB Corpus: Progress and prospects. In Aarts & Meijs (eds) 1984: 41-52. [LOB]

Bald, W-D. & R. Ilson (eds). 1977. *Studies in English usage: The resources of a present-day English corpus for linguistic analysis*. Frankfurt/M.: Peter Lang. [LLC]

Baron, N.S. 1977. *Language acquisition and historical change*. North-Holland Linguistic Series 36. Amsterdam: North-Holland. [BUC]

Beale, A. 1985. Grammatical analysis by computer of the Lancaster-Oslo/Bergen Corpus. *Proceedings of the Association for Computational Linguistics*. Chicago. [LOB]

Beale, A. 1985. A probabilistic approach to grammatical analysis of written English. *Proceedings of the European Chapter of the Association for Computational Linguistics*. Geneva. [LOB]

Bergenholtz, H. & B. Schaeder (eds). 1979. *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora*. Königstein: Scriptor Verlag. [BUC, LLC, LOB]

Biber, D. 1985. Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses. *Linguistics* 23:155-178. [BUC, LLC, LOB]

Biber, D. Forthcoming. On the investigation of spoken/written differences. To appear in *Studia Linguistica*. [BUC, LLC, LOB]

Biber, D. Forthcoming. Spoken and written textual dimensions in English: Resolving the contradictory findings. To appear in *Language* 62. [BUC, LLC, LOB]

Biber, D. Forthcoming. A textual comparison of British and American writing. [BUC, LOB]

Biber, D. & E. Finegan. Forthcoming. Styles of stance: A cluster analysis of texts by adverbial use. [BUC, LLC, LOB]

Black, M. 1977. An investigation into factors influencing the choice between the syllabic and contracted form of *is*. In Bald & Ilson (eds) 1977: 171-182. [LLC]

Blackwell, S.A. 1985. A survey of computer-based English language research. *ICAME News* 9:3-28. [BCE, BUC, LLC, LOB]

Boardman, G.M. 1977. A study of certain kinds of anacolutha in a corpus of spoken English. In Bald & Ilson (eds) 1977: 183-221. [LLC]

Booth, B.M. 1985. Revising CLAWS. *ICAME News* 9: 29-35. [LOB]

Burton, D.M. 1968. *Respice finem* and the *tantum quantum*: An essay review of computational studies for 1967-68. *Computers and the Humanities* 3: 41-48. [BUC]

Bybee, J.L. & D.L. Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language* 58: 165-189. [BUC]

Bäcklund, I. 1984. *Conjunction-headed abbreviated clauses in English.* Studia Anglistica Upsaliensia 50. Stockholm: Almqvist & Wiksell. [BUC, LOB]

Bäcklund, I. Forthcoming. *Beat until stiff.* Conjunction- headed abbreviated clauses in spoken and written English. To appear in Tottie & Bäcklund (eds), forthcoming. [LLC, LOB]

Bäckman, S. & G. Kjellmer (eds). 1985. *Papers on language and literature presented to Alvar Ellegård and Erik Frykman.* Gothenburg Studies in English 60. Gothenburg: Acta Universitatis Gothoburgensis. [BUC, LLC, LOB]

Card, W. & V. McDavid. 1966. English words of very high frequency. *College English* 27: 596-604. [BUC]

Carroll, J.B., P. Davies & B. Richman (eds). 1971. *The American Heritage word frequency book.* New York: American Heritage Publishing Co./Boston: Houghton Mifflin. [BUC]

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

Coates, J. & G.N. Leech. 1980. The meanings of the modals in modern British and American English. *York Papers in Linguistics* 8: 23-34. [BUC, LOB]

Cobussen, W. 1980. The identity of the cleft and pseudo-cleft construction: An argument in favour of base-generation. Graduate thesis, Engels Seminarium, University of Amsterdam. [BUC]

Cooper, D., M.A. Emly, M.F. Lynch & A.R. Yeates. 1979. Compression of continuous prose texts using variety generation. Mimeo. Postgraduate School of Librarianship and Information Science, University of Sheffield. [BUC]

Cresswell, T.J. 1975. Usage in dictionaries and dictionaries of usage. *Publication of the American Dialect Society,* Nos. 63-64. Montgomery, Ala.: Alabama University Press. [BUC]

Crystal, D. 1969. *Prosodic systems and intonation in English.* Cambridge: Cambridge University Press. [LLC]

Crystal, D. 1972. The intonation system of English. *Intonation,* ed. D. Bolinger, 110-136. Harmondsworth: Penguin. [LLC]

Crystal, D. 1975. *The English tone of voice. Essays on intonation, prosody and paralanguage.* London: Arnold. [LLC]

Crystal, D. & R. Quirk. 1964. *Systems of prosodic and paralinguistic features in English.* Janua Linguarum, Series Minor 39. The Hague: Mouton. [LLC]

Dahl, H. 1979. *Word frequencies of spoken American English.* Essex, Conn.: Verbatim. [BUC]

Davidson, B.D. 1977. Aspects of information structuring in modern spoken English. In Bald & Ilson (eds) 1977: 157-170. [LLC]

Davy, D. & R. Quirk. 1969. An acceptability experiment with spoken output. *Journal of Linguistics* 5: 109-120. [LLC]

Dubois, B.L. 1972. Meaning and distribution of the perfect in present-day American English prose. Unpublished Ph.D. dissertation, University of New Mexico. DAI 33/12-A, 6892f. [BUC]

Dubois, B.L. & I.M. Crouch. 1979. *Man* and its compounds in recent profeminist American English published prose. *Papers in Linguistics* 12: 261-269. [BUC]

Eeg-Olofsson, M. 1985. A probability model for computer-aided word-class determination. *ALLC Journal* 5: 25-30. [LLC]

Eeg-Olofsson, M. & J. Svartvik. 1984. Four-level tagging of spoken English. In Aarts & Meijs (eds) 1984: 53-64. [LLC]

Ehrman, M. 1966. *The meanings of the modals in present-day American English.* Janua Linguarum, Series Practica 45. The Hague: Mouton. [BUC]

Ellegård, A. 1978. *The syntactic structure of English texts.* Gothenburg Studies in English 43. Gothenburg: Acta Universitatis Gothoburgensis. [BUC]

Elsness, J. 1981. On the syntactic and semantic functions of *that*-clauses. *Papers from the First Nordic Conference for English Studies, Oslo, 17-19 September 1980*, ed. S. Johansson & B. Tysdahl, 281-303. Department of English, Oslo University. [BUC]

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

Elsness, J. 1984. *That* or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies* 65: 519-533. [BUC]

Engels, L.K. 1982. Testing and mastery learning of English vocabulary at university level. *Studia Anglica Posnaniensia* 15: 129-138. [BUC, LEU, LOB]

Engels, L.K., B. van Beckhoven, Th. Leenders, & I. Brasseur. 1981. *Leuven English teaching vocabulary-list based on objective frequency combined with subjective word-selection.* Department of Linguistics, Catholic University of Leuven. [BUC, LEU, LOB]

Enkvist, N.E. (ed). 1982. *Impromptu speech: A symposium.* Publications of the Research Institute of the Åbo Akademi Foundation 78. Åbo: Åbo Akademi. [LLC]

Enkvist, N.E. 1973. *Linguistic stylistics.* Janua Linguarum, Series Critica 5. The Hague: Mouton. [BUC]

Enkvist, N.E. 1985. A note on the definition and description of true anacolutha. Research Institute of the Åbo Akademi Foundation. Pre-publication copy. Åbo: Åbo Akademi. [LLC]

Enkvist, N.E. & M. Björklund. 1985. Toward a taxonomy of structure shifts. Research Institute of the Åbo Akademi Foundation. Pre-publication copy. Åbo: Åbo Akademi. [LLC]

Erman, B. Forthcoming. Some pragmatic expressions in English conversation. To appear in Tottie & Bäcklund (eds), forthcoming. [LLC]

Filipovic, R. 1969 -. *The Yugoslav Serbo-Croatian/English contrastive project.* Zagreb: Institute of Linguistics/Washington: Center for Applied Linguistics. [BUC]

Finegan, E. 1984. Review of Johansson (ed) 1982. *Language* 60: 190-191. [BUC, LLC, LOB]

Finegan, E. & D. Biber. Forthcoming. Toward a multi-dimensional model of linguistic complexity. To appear in *Southern California Occasional Papers in Linguistics* 11. [BUC, LLC, LOB]

Finegan, E. & D. Biber. Forthcoming. Toward a unified model of sociolinguistic prestige. To appear in *Proceedings of NWAVE XII*, ed. D. Sankoff. Washington, D.C.: Center for Applied Linguistics/New York: Harcourt Brace Jovanovich. [BUC, LLC, LOB]

Fjelkestam-Nilsson, B. 1983. Also *and* too: *A corpus-based study of their frequency and use in modern English.* Stockholm Studies in English 58. Stockholm: Almqvist & Wiksell. [BUC, LLC, LOB]

Flognfeldt, M.E. 1984. The semantics and pragmatics of deverbal nouns ending in *-ee*: A report on work in progress. In Ringbom & Rissanen (eds) 1984: 57-67. [BUC, LOB]

Forsheden, O. 1983. Studies on contraction in the London-Lund Corpus of Spoken English. ETOS Report 2, Survey of Spoken English, Lund University. [LLC]

Francis, W.N. 1964. *A standard sample of present-day English for use with digital computers.* Report to the U.S. Office of Education on Cooperative Research Project No. E-007. Providence, R.I.: Brown University. [BUC]

Francis, W.N. 1965. A standard corpus of edited present-day American English for computer use. *Literary Data Processing Conference Proceedings, 9-11 September 1964*, ed. J.B. Bessinger, S.M. Parrish & H.F. Arader, 79-89. Armonk, N.Y.: IBM Corporation. Slightly revised version in *College English* 26: 267-73. [BUC]

Francis, W.N. 1967. The Brown University Standard Corpus of English: Some implications for TESOL. *On teaching English to speakers of other languages*, ed. B.W. Robinett, 131-35. Washington, D.C.: TESOL. [BUC]

Francis, W.N. 1975. Problems in assembling, describing, and computerizing corpora. *Papers in Southwest English: Research techniques and prospects*, ed. B.L. Dubois & B. Hoffer. San Antonio, Texas: Trinity University. [BUC]

Francis, W.N. 1979. Problems of assembling and computerizing large corpora. Revised version of Francis (1975). In Bergenholtz & Schaeder (eds) 1979: 110-123. Reprinted in Johansson (ed) 1982: 7-24. [BUC]

Francis, W.N. 1980. A tagged corpus: Problems and prospects. *Studies in English linguistics for Randolph Quirk*, ed. S. Greenbaum, G. Leech, & J. Svartvik, 192-209. London: Longman. [BUC]

Francis, W.N. 1982. More verbs in *-alize*. *American Speech* 57: 231-233. [BUC]

Francis, W.N. & H. Kučera. 1979. *Manual of information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers.* Original ed. 1964, revised 1971, revised and augmented 1979. Providence R.I.: Department of Linguistics, Brown University. [BUC]

Francis, W.N. & H. Kucera. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin. [BUC]

Fåhraeus, A-M. 1984. *Two kinds of syntactic-semantic value-loading in English*. Studia Anglistica Upsaliensia 54. Stockholm: Almqvist & Wiksell. [BUC, LOB]

Garnham, A., R. Shillcock, G. Brown, A. Mill & A. Cutler. 1981. Slips of the tongue in the London-Lund Corpus of spontaneous conversation. *Linguistics* 19: 805-17. Reprinted in *Slips of the tongue and language production*, ed. A. Cutler, 251-263. Amsterdam: Mouton, 1982. [LLC]

Garside, R. & G. Leech. 1982. Grammatical tagging of the LOB Corpus: General survey. In Johansson (ed) 1982: 110-117. [LOB]

Garside, R.G. & F.A. Leech. 1985. A probabilistic parser. *Proceedings of the European Chapter of the Asscociation for Computational Linguistics*. Geneva. [LOB]

Garside, R.G., G.N. Leech & G.R. Sampson (eds). Forthcoming. The computational analysis of English. Longman. [BCE, LOB]

Geens, D. 1978. On measurement of lexical differences by means of frequency. *Glottometrica* 1, ed. G. Altman, 46-72. Bochum: Studienverlag Dr. N. Brockmeier. [BUC, LEU]

Geens, D. 1984. Semantic analysis automated for large computer corpora and their exploitation. In Aarts & Meijs (eds) 1984: 211-229.

Geens, D., L.K. Engels, L.K. & W. Martin. 1975. *Leuven Drama Corpus and frequency list*. University of Leuven: Institute of Applied Linguistics. [LEU]

Granger, S. 1983. *The* be *+ past participle construed in spoken English with special emphasis on the passive*. Amsterdam: North-Holland. [LLC]

Greene, B.B. & G.M. Rubin. 1971. Automatic grammatical tagging of English. Providence, R.I.: Department of Linguistics, Brown University. [BUC]

Gustafsson, M. 1982. Textual aspects of topicalization in a corpus of English. *ICAME News* 6: 46-76. [BUC]

Gustafsson, M. 1983. Fronting of adverbials in four genres of English. In Jacobson (ed) 1983: 7-17. [BUC]

Gustafsson, M. 1985. Topicalizations revisited. *Working Papers in English Studies* 6, ed. J. Tommola & T. Virtanen, 43-50. Publications of the Department of English, University of Turku. [BUC]

Haan, P. de. 1984. Problem-oriented tagging of English corpus data. In Aarts & Meijs (eds) 1984: 123-139.

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

Haegeman, L. 1984. Pragmatic conditionals in English. *Folia Linguistica* 13: 485-502. [LLC]

Halteren, H. van. 1984. User interface for a linguistic data base. *ICAME News* 8: 31-40.

Hargevik, S. 1983. Various factors influencing the choice of the auxiliary *need* in present-day English. In Jacobson (ed) 1983: 19-30. [LOB]

Haskel, P.I. 1971. Collocations as a measure of stylistic variety. *The computer in literary and linguistic research*, ed. R.A. Wisbey, 159-168. Cambridge: Cambridge University Press. [BUC]

Hauge, J. & K. Hofland. 1978. Microfiche version of the Brown University Corpus of Present-day American English (text and concordance). Bergen: Norwegian Computing Centre for the Humanities. [BUC]

Hedström, K. 1984. A study of repairs in speech. *Stockholm Papers in English Language and Literature* 4: 69-101. Department of English, Stockholm University. [LLC]

Hermerén, L. 1978. *On modality in English: A study of the semantics of the modals*. Lund Studies in English 53. Lund: CWK Gleerup. [BUC]

Hermerén, L. 1978. Testing the meanings of modals. *Studia Anglica Posnaniensia* 10: 137-140. [BUC]

Hermerén, L. Forthcoming. Modalities in spoken and written English. An inventory of forms. To appear in Tottie & Bäcklund (eds), forthcoming. [LLC, LOB]

Hockey, S. 1980. *A guide to computer applications in the humanities*. London: Duckworth. [BUC]

Hofland, K. & S. Johansson. 1979. LOB Corpus: KWIC concordance. Microfiche. Bergen: Norwegian Computing Centre for the Humanities. [LOB]

Hofland, K. & S. Johansson. 1982. *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities/London: Longman. [BUC, LOB]

Hofland, K. & S. Johansson. 1986. The tagged LOB Corpus: KWIC concordance. Microfiche. Bergen: Norwegian Computing Centre for the Humanities. [LOB]

Householder, F.W. 1971. *Linguistic speculations* (Chapter 13: 'The primacy of writing'). Cambridge: Cambridge University Press. [BUC]

Hurk, I. van den, L. Kager, L. Kemp & M. Masereeuw. 1984. To strand or not to. *ICAME News* 8: 71-83. [LOB]

ICAME News. 1978- . Newsletter of the International Computer Archive of Modern English. Bergen: Norwegian Computing Centre for the Humanities.

Isitt, D. 1983. Crazic, menty, *and* idiotal: *An inquiry into the use of suffixes* -al, -ic, -ly *and* -y *in modern English*. Gothenburg Studies in English 52. Gothenburg: Acta Universatis Gothoburgensis. [BUC]

Jacobson, S. 1982. Modality nouns and the choice between *to* + infinitive and *of* + *ing*. *Studia Anglica Posnaniensia* 15: 61-71. [BUC, LOB]

Jacobson, S. (ed). 1983. *Papers from the Second Scandinavian Symposium on Syntactic Variation*. Stockholm Studies in English 57. Stockholm: Almqvist & Wiksell. [BUC, LLC, LOB]

Jacobson, S. 1985. Form vs. meaning in noun phrases with an *of*-construction. *Papers from the Eighth Scandinavian Conference of Linguistics*, ed. O. Togeby, 426-436. Institut for Nordisk Filologi, University of Copenhagen. [BUC, LOB]

Jahr, M-C. 1981. The *s*-genitive with non-personal nouns in present-day British and American English. *ICAME News* 5: 14-31. [BUC, LOB]

Johannesson, N-L. 1982. On the use of post-modification in English noun phrases. *The Eighth LACUS Forum 1981*, ed. W. Gutwinski & G. Jolly, 187-195. Columbia, S.C.: Hornbeam Press. [LLC, LOB]

Johansson, S. 1978. A computer archive of modern English texts. What? How? Why? When? *Språk og språkundervisning* (Oslo) 11.4: 70-73. [BUC, LLC, LOB]

Johansson, S. 1978. *Some aspects of the vocabulary of learned and scientific English*. Gothenburg Studies in English 42. Gothenburg: Acta Universitatis Gothoburgensis. [BUC]

Johansson, S. 1978. Two corpora of modern English texts. *Et norsk datamaskinelt tekstkorpus. Rapport fra en konferanse i Bergen, 19-20 oktober 1978*, 33-45. Bergen: Norwegian Computing Centre for the Humanities. [BUC, LOB]

Johansson, S. 1979. The use of a corpus in register analysis: The case of learned and scientific English. In Bergenholtz & Schaeder (eds) 1979: 281-293. [BUC]

Johansson, S. 1979. Three systems of grammatical tagging of English text corpora. *Rapport fra den nasjonale konferanse om EDB i språk- og litteraturforskning, 4-5 januar 1979*, 33-46. Bergen: Norwegian Computing Centre for the Humanities. [BUC, LOB]

Johansson, S. 1980. Corpus-based studies of British and American English. *Papers from the Scandinavian Symposium on Syntactic Variation, Stockholm, 18-19 May, 1979*, ed. S. Jacobson, 85-100. Stockholm Studies in English 52. Stockholm: Almqvist & Wiksell. [BUC, LOB]

Johansson, S. 1980. *Plural attributive nouns in present-day English*. Lund Studies in English 59. Lund: C.W.K. Gleerup. [BUC, LOB]

Johansson, S. 1980. Some thoughts on the use of computers in linguistic research. *Humanistiske data* (Bergen) 1: 31-39. [BUC, LLC, LOB]

Johansson, S. 1980. The LOB Corpus of British English Texts: Presentation and comments. *ALLC Journal* 1: 25-36. [LOB]

Johansson, S. 1980. Word frequencies in British and American English: Some preliminary observations. *ALVAR: A Linguistically Varied Assortment of Readings. Studies presented to Alvar Ellegård on the occasion of his 60th birthday*, ed. J. Allwood & M. Ljung, 56-74. Stockholm Papers in English Language and Literature 1. Department of English, Stockholm University. [BUC, LOB]

Johansson, S. 1981. Word frequencies in different types of English texts. *ICAME News* 5: 1-13. [LOB]

Johansson, S. (ed). 1982. *Computer corpora in English language research*. Bergen: Norwegian Computing Centre for the Humanities. [BUC, LLC, LOB]

Johansson, S. 1982. Studying British and American English by computer. *Språk og språkundervisning* (Oslo) 13.4: 48-53. [BUC, LOB]

Johansson, S. 1985. Grammatical tagging and total accountability. In Bäckman & Kjellmer (eds) 1985: 208-220. [LOB]

Johansson, S. 1985. Some observations on word frequencies in three corpora of present-day English texts. *ITL Review of Applied Linguistics* 67-68: 117-126. [BUC, LEU, LOB]

Johansson, S. 1985. Word frequency and text type: Some observations based on the LOB corpus of British English texts. *Computers and the Humanities* 19: 23-36. [LOB]

Johansson, S. Forthcoming. Some observations on the order of adverbial particles and objects in the LOB Corpus. [LOB]

Johansson, S., G. Leech & H. Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers.* Department of English, Oslo University. [BUC, LOB]

Johansson, S., E. Atwell & R. Garside. Forthcoming. *The tagged LOB Corpus: Users' manual.* [LOB]

Johansson, S. & M-C. Jahr. 1982. Grammatical tagging of the LOB Corpus: Predicting word class from word endings. In Johansson (ed) 1982: 118-146. [LOB]

Johansson, S. & K. Hofland. Forthcoming. *Computational analysis of the LOB Corpus.* [LOB]

Johnson, D.D. 1971. The Dolch List reexamined. *The Reading Teacher* 24: 449-457. [BUC]

Kajita, M. 1968. *A generative-transformational study of semi-auxiliaries in present-day English.* Tokyo. [BUC]

Kelly, E. & P. Stone. 1975. *Computer recognition of English word senses.* North-Holland Linguistic Series 13. Amsterdam: North-Holland. [BUC]

Keulen, F. Forthcoming. The Dutch Computer Corpus Pilot Project. Some experiences with a semi-automatic analysis of contemporary English. To appear in Aarts & Meijs (eds), forthcoming.

Kiyokawa, H. 1978. A statistical analysis of American English (1). *Shukutoku Daigaku Kenkyu Kiyo*, No. 13 (in Japanese). [BUC]

Kjellmer, G. 1979. On clause-introductory *nor* and *neither*. *English Studies* 60: 280-295. [BUC]

Kjellmer, G. 1980. *Accustomed to swim: accustomed to swimming.* On verbal forms after *to*. *ALVAR. A Linguistically Varied Assortment of Readings. Studies presented to Alvar Ellegård on the occasion of his 60th birthday*, ed. J. Allwood & M. Ljung, 75-99. Stockholm Papers in English Language and Literature 1. Department of English, Stockholm University. [BUC]

Kjellmer, G. 1980. *There is no hiding you in the house*: On a modal use of the English gerund. *English Studies* 61: 47-60. [BUC]

Kjellmer, G. 1981. *Literally*: A case of harmful polysemy? *Studia Neophilologica* 53: 275-282. [BUC, LOB]

Kjellmer, G. 1982. *Each other* and *one another*. On the use of the English reciprocal pronouns. *English Studies* 63: 231-254. [BUC]

Kjellmer, G. 1982. Some problems relating to the study of collocations in the Brown Corpus. In Johansson (ed) 1982: 25-33. [BUC]

Kjellmer, G. 1982. *What to do*? On non-finite direct questions in English. *English Studies* 63: 446-454. [BUC]

Kjellmer, G. 1983. A contemporary semantic clash. *English Studies* 64: 460-466. [BUC]

Kjellmer, G. 1983. *He is one of the few men in history who plays jazz on a violin*. On number concord in certain relative clauses. *Anglia* 101: 299-314. [BUC, LOB]

Kjellmer, G. 1984. A preposition vanishes. In Bäckman & Kjellmer (eds) 1984: 233-244. [BUC, LOB]

Kjellmer, G. 1984. On the grammatical number of relative *what*. *English Studies* 65: 256-273. [BUC, LOB]

Kjellmer, G. 1984. Some thoughts on collocational distinctiveness. In Aarts & Meijs (eds) 1984: 163-171. [BUC]

Kjellmer, G. 1984. Why *great: greatly* but not *big: *bigly*? On the formation of English adverbs in *-ly*. *Studia Linguistica* 38: 1-19. [LOB]

Kjellmer, G. 1985. *Help to/Help ø* revisited. *English Studies* 66: 156-161. [BUC, LOB]

Krogvig, I. & S. Johansson. 1981. *Shall*, *will*, *should* and *would* in British and American English. *ICAME News* 5: 32-56. [BUC, LOB]

Krogvig, I. & S. Johansson. 1984. *Shall* and *will* in British and American English: A frequency study. *Studia Linguistica* 38: 70-87. Revision of Krogvig & Johansson (1981). [BUC, LOB]

Kučera, H. 1968. Some quantitative lexical analyses of Russian, Czech, and English. *American contributions to the Sixth International Congress of Slavists I*, ed. H. Kučera, 1-44. The Hague: Mouton. [BUC]

Kučera, H. 1969. Computers in language analysis and in lexicography. *The American Heritage Dictionary of the English Language*, ed. W. Morris, xxxviii-xl. Boston: Houghton Mifflin. [BUC]

Kučera, H. 1980. Computational analysis of predicational structures in English. *Proceedings of the Eighth International Conference on Computational Linguistics, Tokyo, 30 Sept. - 4 Oct. 1980*, 32-37. [BUC]

Kučera, H. 1982. The mathematics of language. *The American Heritage Dictionary*. Second college edition, 37-41. Boston: Houghton Mifflin. [BUC]

Kučera, H. & W.N. Francis. 1967. *Computational analysis of present-day American English.* Providence, R.I.: Brown University Press. [BUC]

Leech, G. 1986. Automatic grammatical analysis and its educational applications. *Computers in English language teaching and research,* ed G. Leech & C.N. Candlin, 205-16. London: Longman. [LOB]

Leech, G. & A. Beale. 1984. Computers in English language research. State of the art article in *Language Teaching and Linguistics: Abstracts* 17: 216-229. [BCE, BUC, LEU, LLC, LOB]

Leech, G. & J. Coates. 1980. Semantic indeterminacy and the modals. *Studies in English linguistics for Randolph Quirk*, ed. S. Greenbaum, G. Leech & J. Svartvik, 79-90. London: Longman. [BUC, LOB]

Leech, G., R. Garside & E.S. Atwell. 1983. Recent developments in the use of computer corpora in English language research. *Transactions of the Philological Society*, 32-40. [BUC, LLC, LOB]

Leech, G., R. Garside & E.S. Atwell. 1983. The automatic grammatical tagging of the LOB Corpus. *ICAME News* 7: 13-33. [LOB]

Leech, G. & R. Leonard. 1974. A computer corpus of British English. *Hamburger Phonetische Beiträge* 13: 41-57. [BUC, LOB]

Leonard, R. 1977. The Computer Archive of Modern English Texts. *Computational and mathematical linguistics. Proceedings of the International Conference on Computational Linguistics, Pisa, 27 Aug - 1 Sept 1973*, ed. A. Zampolli & N. Calzolari, 417-428. Florence: Leo S. Olschki Editore. [BUC, LOB]

Lynch, M.F. & S.D. Rawson. 1976. Equifrequent character strings - A novel text characterization method. *The computer in literary and linguistic studies*, ed. A. Jones & R.F. Churchhouse, 47-58. Cardiff: University of Wales Press. [BUC]

McCarthy, M. Forthcoming. Interactive lexis: Prominence and paradigms. To appear 1986. [LLC]

McCarthy, M. Forthcoming. Some vocabulary patterns in conversation. To appear in *Vocabulary in language learning and teaching*, ed. M.J. McCarthy & R. Carter. London: Longman. [LLC]

Marshall, I. 1983. Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB Corpus. *Computers and the Humanities* 17: 139-50. [LOB]

Meijs, W. 1982. Exploring Brown with QUERY. In Johansson (ed) 1982: 34-48. [BUC]

Meijs, W. 1984. *You can do so if you want to* - some elliptic structures in Brown and LOB and their syntactic description. In Aarts & Meijs (eds) 1984: 141-162 [BUC, LOB]

Meijs, W. 1984. Data and theory in computer corpus research. *English language research: The Dutch contribution*, I, ed J. Lachlan Mackenzie & H. Wekker, 85-99. Amsterdam: Free University Press. [BUC, LOB]

Meyer, C.F. 1983. A descriptive study of American punctuation. Unpublished Ph.D. thesis, University of Wisconsin-Milwaukee. [BUC]

Meyers, W.E. 1972. A study of usage items based on an examination of the Brown Corpus. *College Composition and Communication* 23: 155-169. [BUC]

Michiels, A. 1982. Exploiting a large dictionary data base. Ph.D. thesis, Université de Liège.

Monroe, G.K. 1965. Phonemic transcription of graphic post-base affixes in English: A computer problem. Unpublished Ph.D. dissertation in linguistics, Brown University. DA 26/08, 4648. [BUC]

Nässlin, S. 1984. *The English tag question: A study of sentences containing tags of the type* isn't it? *and* is it?. Stockholm Studies in English 60. Stockholm: Almqvist & Wiksell. [LLC]

Olofsson, A. 1981. *Relative junctions in written American English.* Gothenburg Studies in English 50. Gothenburg: Acta Universitatis Gothoburgensis. [BUC]

Oostdijk, N. 1984. An extended affix grammar for the English noun phrase. In Aarts & Meijs (eds) 1984: 95-122.

Oreström, B. 1977. Why */dhi/ book*? SSE Report, Survey of Spoken English, Lund University. [LLC]

Oreström, B. 1982. When is it my turn to speak? In Enkvist (ed) 1982: 267-276. [LLC]

Oreström, B. 1983. *Turn-taking in English conversation.* Lund Studies in English 66. Lund: Liber/Gleerup. [LLC]

Oreström, B., J. Svartvik & C. Thavenius. 1976. Manual for terminal input of spoken English material. SSE Report, Survey of Spoken English, Lund University. [LLC]

Oreström, B. & C. Thavenius. 1978. Auditory and acoustic analysis: An experiment. SSE Report, Survey of Spoken English, Lund University. [LLC]

Pearson, C. 1978. Quantitative investigations into the type-token relation for symbolic rhemes. Mimeo. Georgia Institute of Technology. [BUC]

Phillips, M. Forthcoming. *Lexical structure of text.* English Language Research, Monograph No 12. University of Birmingham. [BCE]

Quirk, R. & D. Crystal. 1966. On scales of contrast in connected English speech. *In memory of J.R. Firth*, ed. C.E. Bazell et al, 359-369. London: Longman. [LLC]

Quirk, R., A. Duckworth, J. Rusiecki, J. Svartvik & A. Colin. 1964. Studies in the correspondence of prosodic to grammatical features in English. *Proceedings of the Ninth International Congress of Linguists*, 679-691. The Hague: Mouton. Reprinted in R. Quirk, *Essays on the English language: Medieval and modern*, 120-135. London: Longman, 1968. [LLC]

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

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A comprehensive grammar of the English language.* London: Longman. [BUC, LLC, LOB]

Quirk, R. & J. Svartvik. 1978. A corpus of modern English. In Bergenholtz & Schaeder (eds) 1978: 204-218. [LLC]

Recktenwald, R.P. 1975. The English progressive: Semantics and history. Unpublished PhD dissertation in linguistics, Brown University. [BUC]

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

Renouf, A. Forthcoming. The elicitation of spoken English. To appear in Tottie & Bäcklund (eds), forthcoming. [BCE]

Ringbom, H. 1973. *George Orwell as essayist: A stylistic study.* Acta Academiae Aboensis, Ser. A, Humaniora, 44:2. Åbo: Åbo Akademi. [BUC]

Ringbom, H. 1975. The style of Orwell's preface to *Animal Farm*. *Style and Text: Studies presented to Nils Erik Enkvist*, ed. H. Ringbom, 243-249. Stockholm: Språkförlaget Skriptor. [BUC]

Ringbom, H. & M. Rissanen (eds). 1984. *Proceedings from the Second Nordic Conference for English Studies, Hanasaari/Hanaholmen, 19-21 May 1983*. Publications of the Research Institute of the Åbo Akademi Foundation 92. Åbo: Åbo Akademi. [BUC, LLC, LOB]

Rissanen, M. 1979. On the position of *only* in present day written English. *Papers from the Scandinavian Symposium on Syntactic Variation*, ed. S. Jacobson, 63-76. Stockholm Studies in English 52. Stockholm: Almqvist & Wiksell. [BUC, LLC]

Rycker, T. de. 1984. Imperative structures: Form and function in conversational English. *Antwerp Papers on Linguistics* 38. Antwerp: University of Antwerp. [LLC]

Sahlin, E. 1979. Some *and* any *in spoken and written English*. Studia Anglistica Upsaliensia 38. Uppsala: Almqvist &Wiksell. [BUC, LLC]

Sampson, G.R. 1983. Fallible rationalism and machine translation. *Proceedings of the First Chapter of the Association for Computational Linguistics*, 86-89. Menlo Park, California.

Schaeder, B. 1976. Maschinenlesbare Text-corpora des Deutschen und des Englischen. *Deutsche Sprache* 4: 356-370. [BUC, LLC, LOB]

Schaeder, B. 1979. Maschinenlesbare Text-Corpora des Deutschen und Englischen: Eine Dokumentation. In Bergenholtz & Schaeder (eds) 1979: 325-336. [BUC, LOB, LLC]

Sedelow, S.Y. & W.A. Sedelow, Jr. 1967. Stylistic analysis. *Automated language processing*, ed. H. Borko, 181-213. New York: Wiley. [BUC]

Shastri, S.V. 1980. A computer corpus of present-day Indian English. *ICAME News* 4: 9-10. [BUC, LOB]

Shastri, S.V. 1985. Word frequencies in Indian English: A preliminary report. *ICAME News* 9: 38-44. [BUC, LOB]

Sherman, D. 1977. A computer archive of language materials. *Computing in the humanities: Proceedings of the Third International Conference on Computing in the Humanities*, ed. S. Lusignan & J.S. North, 283. Waterloo, Ontario: University of Waterloo Press. [BUC]

Sinclair, J. McH. 1970. English lexical studies. Final report to OSTI on Project C/LP/08 for January 1967 - September 1969. Department of English, Birmingham University.

Sinclair, J. McH. 1980. Computational text analysis at the University of Birmingham. *ICAME News* 4: 13-16.

Sinclair, J. McH. 1982. Reflections on computer corpora in English language research. In Johansson (ed) 1982: 1-6.

Sinclair, J. McH. 1985. First throw away your evidence. L.A.U.D.T., Series A, Paper No. 151. Duisburg: Linguistic Agency, University of Duisburg. [BCE]

Sinclair, J. McH. 1986. Basic computer processing of long texts. *Computers in English language teaching and research,* ed. G. Leech & C.N. Candlin, 185-203. London: Longman. [BCE]

Sinclair, J. McH. 1986. Sense and structure in lexis. *Linguistics in a systemic perspective,* ed J. Benson, W. Greaves & M. Cummings. University of York, Toronto. [BCE]

Smith, D.A. 1971. An automatic parsing procedure for simple noun- and verb-phrases. M.A. thesis in Linguistics, Brown University. [BUC]

Smith, R.N. 1973. *Probabilistic performance models of language*. Janua Linguarum, Series Minor 150. The Hague: Mouton. [BUC]

Solso, R.L., P.F. Barbuto, Jr. & C.L. Juel. 1979. Bigram and trigram frequencies and versatilities in the English language. *Behavior Research Methods & Instrumentation* 11: 475-484. [BUC]

Solso, R.L. & C.L. Juel. 1980. Positional frequency and versatility of bigrams for two- through nine-letter English words. *Behavior Research Methods & Instrumentation* 12: 297-343. [BUC]

Solso, R.L. & J.F. King. 1976. Frequency and versatility of letters in the English language. *Behavior Research Methods & Instrumentation* 8: 283-286. [BUC]

Steen, G.J. van der. 1982. A treatment of queries in large text corpora. In Johansson (ed) 1982: 49-65.

Steen, G.J. van der. 1984. On the unification of matching, parsing and retrieving in text corpora. *ICAME News* 8: 41-46.

Stenström, A-B. 1982. Feedback. In Enquist (ed) 1982: 319-340. [LLC]

Stenström, A-B. 1983. Questioning strategies in English and Swedish conversation. *Cross-language analysis and second language acquisition* 2, ed. K. Sajavaara, 67-78. Jyväskylä Cross Language Studies 10. Jyväskylä: University of Jyväskylä. [LLC]

Stenström, A-B. 1984. Discourse tags. In Aarts & Meijs (eds) 1984: 65-82. [LLC]

Stenström, A-B. 1984. *Questions and responses in English conversation*. Lund Studies in English 68. Lund: Gleerup/Liber. [LLC]

Stenström, A-B. 1985. English in speech and writing. *Papers and studies in contrastive linguistics*, ed. J. Fisiak, 115-130. Poznan: Adam Mickiewicz University. [LLC, LOB]

Stenström, A-B. Forthcoming. Questions and conversational analysis. To appear in *Questions and questioning*, ed. M. Meyer. [LLC]

Stenström, A-B. Forthcoming. What does *really* really do? Strategies in speech and writing. To appear in Tottie & Bäcklund (eds), forthcoming. [LLC, LOB]

Sundbye, N.W., N.J. Dyck & F.R. Watt. 1980. Essential sight words program, level 2: Guide. Hingham, Mass.: Teaching Resources Corporation. [BUC]

Svartvik, J. 1968. Plotting divided usage with *dare* and *need*. *Studia Neophilologica* 40: 130-140. [BUC]

Svartvik, J. 1980. Interactive parsing of spoken English. *Proceedings from the 8th International Conference on Computational Linguistics, Tokyo, 30 Sept-4 Oct 1980*. [LLC]

Svartvik, J. 1980. Tagging spoken English. *ALVAR. A Linguistically Varied Assortment of Readings. Studies presented to Alvar Ellegård on the occasion of his 60th birthday*, ed. J. Allwood & M. Ljung, 182-206. Stockholm Papers in English Language and Literature 1. Department of English, Stockholm University. [LLC]

Svartvik, J. 1980. *Well* in conversation. *Studies in English linguistics for Randolph Quirk*, ed. S. Greenbaum, G. Leech & J. Svartvik, 167-177. London: Longman. [LLC]

Svartvik, J. 1982. On speaking terms. Inträdesföredrag, *Kungl Vitterhets Historie och Antikvitets Akademien, Årsbok 1982*, 110-118. Stockholm: Almqvist & Wiksell. [LLC]

Svartvik, J. 1982. The segmentation of impromptu speech. In Enkvist (ed) 1982: 131-145. [LLC]

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

Svartvik, J. & M. Eeg-Olofsson. 1982. Tagging the London-Lund Corpus of Spoken English. In Johansson (ed) 1982: 85-109. [LLC]

Svartvik, J., M. Eeg-Olofsson, O. Forsheden, B. Oreström & C. Thavenius. 1982. *Survey of Spoken English: Report on research 1975-81*. Lund Studies in English 63. Lund: Gleerup/Liber. [LLC]

Svartvik, J. & R. Quirk (eds). 1980. *A corpus of English conversation*. Lund Studies in English 56. Lund: Gleerup/Liber. [LLC]

Svartvik, J. & A-B. Stenström. 1985. Words, words, words: The rest is silence? In Bäckman & Kjellmer (eds) 1985: 342-353. [LLC]

Svindland, A.S. 1981. *Both - and*, a re-evaluation. Series B, No 4: 1-56. Department of Phonetics, University of Bergen. [BUC, LOB]

Sørheim, M-C. J. 1981. The genitive in a functional sentence perspective. *Papers from the First Nordic Conference for English Studies, Oslo, 17-19 September 1980*, ed. S. Johansson & B. Tysdahl, 405-423. Department of English, Oslo University. [BUC, LOB]

Taglicht, J. 1977. Relative clauses as postmodifiers: Syntax and intonation. In Bald & Ilson (eds) 1977: 73-107. [LLC]

Taglicht, J. 1983. *Message and emphasis*. London: Longman. [LLC]

Tanaka, H. 1971. A statistical study on selectional features of transitive verbs in present-day American English. Unpublished Ph.D. dissertation, Brown University. DAI 32/10-A, 5769. [BUC]

Thavenius, C. 1979. Referential *it* in spoken English. *Actes du 5ème Congrès de l'Association International de Linguistique Appliquée, Montréal, août 1979*, ed. J-G. Savard & L. Laforge. Québec: Les Presses de l'Université Laval. [LLC]

Thavenius, C. 1982. Exophora in English conversation. A study of third person pronominal reference. In Enkvist (ed) 1982: 291-305. [LLC]

Thavenius, C. 1983. *Referential pronouns in English conversation*. Lund Studies in English 64. Lund: Liber/Gleerup. [LLC]

Thavenius, C. 1984. Pronominal chains in English conversation. In Ringbom & Rissanen (eds) 1984: 209-219. [LLC]

Thavenius, C. & B. Oreström (eds). 1979. Konkordanser: Föredrag från 2:a svenska kollokviet i språklig databehandling i Lund 1979. SSE Report, Survey of Spoken English, Lund University. [BUC, LOB]

Tottie, G. 1980. Affixal and non-affixal negation. Two systems in (almost) complementary distribution. *Studia Linguistica* 34: 101-123. [LLC]

Tottie, G. 1981. Negation and discourse strategy in spoken and written English. *Variation omnibus*, ed. H. Cedergren & D. Sankoff, 271-284. Edmonton, Alberta: Linguistic Research. [LLC]

Tottie, G. 1982. Where do negative sentences come from? *Studia Linguistica* 36: 88-105. [LLC]

Tottie, G. 1983. *Much about* not *and* nothing: *A study of the variation between analytic and synthetic negation in contemporary American English*. Lund: CWK Gleerup. [BUC]

Tottie, G. 1983. The missing link? or, Why is there twice as much negation in spoken English as in written English? In Jacobson (ed) 1983: 67-74. [LLC, LOB]

Tottie, G. 1984. Is there an adverbial in this text? (And if so, what is it doing there?) In Ringbom & Rissanen (eds) 1984: 299-315. [LLC, LOB]

Tottie, G. Forthcoming. *No*-negation and *not*-negation in spoken English. To appear in *The Thirteenth Colloquium on New Ways of Analyzing Variation in English and Other Languages*, ed. S. Ash. [LLC]

Tottie, G. Forthcoming. The importance of being adverbial. Focusing and contingency adverbials in spoken and written English. To appear in Tottie & Bäcklund (eds), forthcoming. [LLC, LOB]

Tottie, G., B. Altenberg & L. Hermerén. 1983. English in speech and writing. A manual for students. ETOS Report, Departments of English, Lund and Uppsala Universities. [BUC, LLC, LOB]

Tottie, G. & I. Bäcklund (eds). Forthcoming. *English in speech and writing. Proceedings from the Symposium on English in Speech and Writing, Uppsala, 5 Oct. 1984*. Stockholm: Almqvist & Wiksell. [LLC, LOB]

Tottie, G., M. Eeg-Olofsson & C. Thavenius. 1984. Tagging negative sentences in LOB and LLC. In Aarts & Meijs (eds) 1984: 173-184. [LLC, LOB]

Tottie, G. & C. Paradis. 1982. From function to structure. Some pragmatic determinants of syntactic frequencies to impromptu speech. In Enkvist (ed) 1982: 307-317. [LLC]

Tottie, G. & G. Övergaard. 1984. The author's *would*. A feature of American English? *Studia Linguistica* 38: 148-165. [BUC, LOB]

Viitanen, O. Forthcoming. On the position of *only* in English conversation. To appear in Tottie & Bäcklund (eds), forthcoming. [LLC]

Warren, B. 1978. *Semantic patterns of noun-noun compounds*. Gothenburg Studies in English 41. Gothenburg: Acta Universitatis Gothoburgensis. [BUC]

Warren, B. 1984. *Classifying adjectives*. Gothenburg Studies in English 56. Gothenburg: Acta Universitatis Gothoburgensis. [BUC]

Westney, P. 1983. Review of Svartvik & Quirk (eds) 1980. *IRAL* 21: 336-338. [LLC]

Wieser, E. Forthcoming. The extra oomph: Given and new information in English pronouns referring to speaker and addressee in conversation and monologue. To appear in *Studia Linguistica*. [LLC]

Wikberg, K. 1984. Some critical observations on present-day English lexicology. In Ringbom & Rissanen (eds) 1984: 103-116. [BUC, LOB]

Yang, H.-J. Forthcoming. Automatic identification of technical terms. To appear in *ALLC Journal*.

Yates, A.R. 1977. Text compression in the Brown Corpus using variety-generated keysets, with a review of the literature on computers in Shakespearean studies. M.A. dissertation, University of Sheffield. [BUC]

Zettersten, A. 1968. Current computing activity in Scandinavia relating to language and literature research. *Computers and the Humanities* 3: 53-60. [BUC]

Zettersten, A. 1969. *A statistical study of the graphic system of present-day American English*. Lund: Studentlitteratur. [BUC]

Zettersten, A. 1969. *A word-frequency list of scientific English*. Lund: Studentlitteratur. [BUC]

Zettersten, A. 1978. A word frequency list based on American English press reportage. *Publications of the Department of English, University of Copenhagen*, vol. 6. Copenhagen: Akademisk Forlag. [BUC]

PUNCTUATION PRACTICE IN THE BROWN CORPUS

Charles F. Meyer
Western Kentucky University

Except for Summey's (1949) treatment of American punctuation, we know very little about American punctuation beyond the lists of prescriptive rules given in style manuals or usage books. While these books provide information about how we should punctuate, they say very little about actual punctuation practice. This paper reports the results of a corpus-based study that sought to investigate American punctuation practice. The study focused exclusively on "structural punctuation" (Summey 1949:3): periods, question marks, exclamation marks, commas, dashes, semicolons, colons, and parentheses. It did not deal with paragraph indentations (or separation) or apostrophes and hyphens, nor did it focus on brackets, ellipsis dots, quotation marks, and underlining, or the use of commas and colons in dates, times, etc. These are marks of punctuation whose uses have been fairly rigidly conventionalized by style manuals.

So that the study accurately reflected current usage, it was based on sections of three contrasting styles of the Brown Corpus: journalism, learned writing, and fiction. Twelve samples from each style were selected; the corpus therefore totaled approximately 72,000 words. All of the frequency counts cited in the study were based on this corpus.[1]

The study was functionally based. That is to say, it focused on the functions of punctuation in the Brown Corpus. It was found that punctuation had three primary linguistic functions: a syntactic function, a semantic function, and a prosodic function.

Syntax and punctuation

Punctuation and syntax are related in the following manner: the marks of punctuation form a hierarchy, with the period, question mark, and exclamation mark highest in the hierarchy and the comma lowest; the hierarchical nature of the marks enables them to indicate whether a constitutent is superordinate or subordinate to some other constituent; and as constituents become more lengthy and complex, it becomes more necessary to punctuate them with a mark higher up in the punctuation hierarchy.

Structural punctuation forms a hierarchy (Quirk et al. 1985:1611-1613 and Limaye 1983:30), whose structure is determined by the particular grammatical unit that a mark of punctuation is used to separate or enclose: the sentence, the clause, or the phrase. Highest in the hierarchy (Level 1) are the period, question mark, and exclamation mark, marks of punctuation that set off sentences; lowest on the hierarchy (Level 3) is the comma, a mark that can set off only clauses and phrases; and in-between these two extremes (Level 2) are the colon, parenthesis, dash, and semicolon, marks setting off a variety of syntactic constructions, from the sentence down to the phrase.

The most common marks of punctuation were the comma and the period, which constituted 47% and 45% of the marks in the corpus, respectively. Occurring much more infrequently were the remaining marks: dashes (2%), parentheses (2%), semicolons (2%), question marks (1%), colons (1%), and exclamation marks (1%).

Periods and question marks most frequently set off declarative and interrogative sentences, respectively. Exclamation marks, however, did not set off exclamatory sentences most frequently, that is, sentences that began with *how* or *what* and in which there was no subject-verb inversion:[2]

> She had cried, she had implored, she had been miserable at this refusal, and finally he had relented – and now how happy she was, how expectant! (K29 1000-1020)

In fact, only one of the twenty-five exclamation marks that occurred in the corpus set off an exclamatory sentence; the majority set off declarative sentences:

> In the family's own words (during the third of twelve visits), they had "reached the crisis peak – either the situation will give or we will break!" (J24 210-240)

The marks of punctuation at Level 1 were distributed differently across the three styles of the corpus. Most orthographic sentences (97%) were separated by periods, a mark of punctuation that occurred more frequently in the journalistic and fictional styles than in the learned style. This distribution of periods is a direct reflection of the tendency for sentences to be longer and more complex in learned styles than in journalistic or fictional styles, styles that have to appeal to a wide range of readers with varying reading abilities and that therefore have to contain shorter and less complex sentences.

Relatively few orthographic sentences (3%) were separated by question marks or exclamation marks, and 80% of these sentences occurred in the fictional style. The restricted occurrence of question marks and exclamation marks in the corpus simply reflects the fact that these marks set off constructions that are unlikely to occur in most styles of writing. Questions occur mainly in conversation because their function is to enable the speaker to request information from the listener (Quirk et al. 1985:803-4). Exclamations are highly emotional and thus would be distracting in learned or journalistic styles.

One mark at Level 2, the semicolon, had a restricted occurrence in the corpus. Most instances (53%) of this mark were found in the learned style; only 17% and 30% of the semicolons in the corpus occurred in the journalistic and fictional styles, respectively. This distribution suggests that semicolons are markers of formal style.

Lowest on the hierarchy is the comma, a mark occurring at Level 3 and restricted to setting off clauses and phrases. The comma is the most versatile mark of punctuation and can set off a variety of different syntactic constructions. Because of the versatility of the comma, situations occasionally arise in writing when two or more commas with distinct syntactic functions could occur close together. In the example below, for instance, one comma is used to separate two clauses coordinated by *and*, two other commas nearby are used to enclose an adverbial phrase, and within this phrase are commas separating elements of a series:

> No attempts to measure the radio emission of the remaining planets have been reported, and, because of their distances, small diameters, or low temperatures, the thermal radiation at radio wave lengths reaching the earth from these sources is expected to be of very low intensity. (J01 370-410)

Two problems arise in situations like the above: if commas are used in all instances, (1) an overpunctuated and visually unattractive construction may result (the two commas around *and* in the above example), or (2) potential misinterpretation may occur (the comma following *distances* being misinterpreted as enclosing the *because of* phrase rather than as separating the first element of a series from the second). Because the marks of punctuation are hierarchical, however, there were various ways in the corpus that they interacted to mark grammatical hierarchies in examples such as above, to distinguish superordinate from subordinate boundaries.

If two or more boundaries occurred close together and both could be optionally punctuated, then the superordinate boundary was generally marked and the subordinate boundary unmarked. In the sentence below, an adverbial follows the coordinating conjunction in a compound sentence. Although current practice allows the superordinate boundary in such examples to be marked with a semicolon and the subordinate boundaries of the adverbial enclosed with commas, this type of punctuation produces an overpunctuated sentence, one which is visually unattractive. Hence, only 2% of the sentences of this type in the corpus were thus heavily punctuated.

> The capital budget, for construction of permanent improvements becomes an appropriating document instead of just a calendar of pious promises; but, as a second-look safeguard, each new product must undergo a Board of Estimate public hearing before construction proceeds. (B07 750-780)

The majority of these types of sentences (85%) contained the superordinate boundary marked and either one or none of the subordinate boundaries marked:

> Bushes and vines abetted the rocks in forming thorny detours for the struggling stranger, and without the direct light of the

> sun to act as a compass, Pamela could no longer be positive of her direction. (N08 1030-1050)

The two sentences above contain boundaries that can be optionally punctuated. Other sentences, however, contain superordinate and subordinate boundaries occurring close together that have to be punctuated. In the corpus, sentences of this type contained boundaries that were both marked. However, the superordinate boundary was set off with a mark higher up on the punctuation hierarchy: a dash, a pair of parentheses, a period, or a semicolon. In the example below, dashes instead of commas were used so that the superordinate boundaries of the appositive were distinguished from the subordinate boundaries of the series:

> Simultaneously, a variety of environmental supports – a calm but not too motherly homemaker, referral for temporary economic aid, intelligent use of nursing care, accompaniment to the well-baby clinic for medical advice on the twin's feeding problem – combined to prevent further development of predictable pathological mechanisms. (J24 370-430)

If a series contained internal punctuation and occurred within a sentence rather than within an appositive, semicolons were used to distinguish the superordinate boundaries of the series from the subordinate boundaries of the units within each part of the series:

> As the historic process of modernization gradually gains momentum, their cohesion will be threatened by diverse forces; the gaps between rulers and subjects, town and country will widen; new aspirants for power will emerge whose ambitions far exceed their competence; and old rulers may lose their nerve and their sense of direction. (J22 10-60)

Length and complexity were the final syntactic considerations that affected punctuation. There were two types of syntactic constructions whose punctuation was affected by their complexity and length: coordinated constructions and adverbials occurring at the beginning of main clauses.

The complexity of the particular coordinated construction directly affected its punctuation. Non-elliptical compound sentences were punctuated quite frequently (85% of the time):

> The revolution was well under way before 700 B.C., and premonitory signs go back virtually across the century. (J54 1150-1170)

Punctuated far less frequently were syntactically less complex constructions: compound sentences with subject-ellipsis at the beginning of the second clause were punctuated only 17% of the time; compound subordinate clauses only 23% of the time; and compound phrases only 13% of the time.

In non-elliptical compound sentences, there was a tendency for heavier punctuation to be used as the length and complexity of the main clauses increased. If the main clauses were short (ten words or less in length), they were unpunctuated about one-third of the time:

> The Rev. Richard Freeman of Texas City officiated and Charles Pabor and Mrs. Marvin Hand presented music. (A17 1550-1580)

However, if one or both clauses were longer than ten words, they were unpunctuated only 16% of the time, and punctuated in the remainder of situations with commas, semicolons, dashes, or periods:

> In addition, the neocortical-hypothalmic relations play a great role in primates, as Mirsky's interesting experiment on the "Communication Affects" demonstrates. But even in relatively primitive laboratory animals such as the rat, sex activity closely identified with the hypothalamus and visceral brain is enhanced by neocortex. (J17 160-210)

If the main clauses were highly complex (that is, if they contained two or more subordinate clauses), they were punctuated quite frequently (97% of the time):

> Their world, again, was a still simple, traditional age which was only slowly beginning to appreciate the complexity of life. And perhaps an observer of the vases will not go too far in deducing that the outlook of their makers and users was basically stable and secure. (J54 90-130)

However, length may still be a factor in sentences of this type, since main clauses become lengthier as they become more complex.

Length and complexity also affected the punctuation of adverbials when they occurred initially in a sentence. Adverbs are short and simple, and therefore were punctuated only about one-third of the time:

> Guns were going off all over Washington City these days, because of the celebrations, and the theater was not soundproof. *Then* the audience saw a small, dim figure appear at the edge of the presidential box. (K05 1180-1210)

Adverbial phrases, on the other hand, are non-complex but vary considerably in length. Hence, if the phrase was short (two or three words in length), it was punctuated only about half of the time:

> Dr. Hester, of Princeton, N.J., is a native of Chester, PA. He joined N.Y.U. in September, 1960. *Prior to that* he was associated with Long Island University in Brooklyn. (A24 1970-1980)

On the other hand, if the phrase was longer than two or three words, it was punctuated about 80% of the time:

> We congratulate the entire membership on its record of good legislation. *In the interim between now and next year*, we trust the House and Senate will put their minds to studying Georgia's very real economic, fiscal and social problems .... (B01 300-340)

Adverbial clauses were virtually always punctuated (about 98% of the time). While they vary in length considerably, they are syntactically subordinate and for this reason were almost always punctuated:

> *Touring Africa*, the new U.S. assistant secretary observed "Africa should be for the Africans" and the British promptly denounced him. (B01 920-940)

Semantics and punctuation

The relationship between punctuation and meaning works in two directions. On the one hand, punctuation can be used to create a semantic effect. In the example below, the question mark is alone responsible for indicating that the sentence is an interrogative sentence. Without the question mark, the sentence would be interpreted as a declarative sentence.

> Ralph is a beachcomber? (Baldwin and Coady 1978:374)

On the other hand, punctuation can be used to simply reinforce a semantic effect. In contrast to the question mark in the example above, the question mark in the sentence below does not alone indicate that the sentence is an interrogative sentence. Rather, it merely reinforces the meaning of the sentence that has already been conveyed by the syntax.

> There were times now, like this, when she lost control of the count and moved freely back and forth into three generations. Was it a birthday ball? (K16 230-260)

Although all punctuation is to a certain extent semantically motivated, some punctuation is chosen primarily for semantic reasons. However, punctuation is a relatively weak semantic cue: it only rarely and ineffectively creates semantic effects and is most frequently used to reinforce semantic effects.

There were very few examples in the corpus of sentences in which the punctuation alone was responsible for conveying the meaning of the sentence: even though punctuation was used to distinguish restrictive from non-restrictive modifiers, to differentiate homonyms, and to prevent potential ambiguities, frequently pragmatic factors, such as the linguistic context in which a mark occurred, made the mark unnecessary.

In the corpus, some punctuation was crucial for distinguishing restrictive from non-restrictive constructions. In the example below, lack of punctuation suggests that the postmodifier should be interpreted as a restrictive rather than a non-restrictive modifier: if the relative clause *who are on duty* .... were enclosed with commas, it would appear that all attendants were being discussed in this sentence, not merely those working 65 hours:

> The practice of charging employes [sic] for meals whether they eat at the hospital or not should be abolished. The work week of attendants *who are on duty 65 hours or more per week* should be reduced. (B01 1120-1150)

In other constructions, on the other hand, the presence or absence of punctuation did not affect restrictiveness or non-restrictiveness. In the example below, the reference of *these women* is clear from the context. Hence, the relative clause that follows it is still non-restrictive, even though the clause is not enclosed with commas:

> The League of Women Voters, 40 now and admitting it proudly, is inviting financial contributions in the windup of its fund drive ... These women *whose organization grew out of the old suffrage movement* are dedicated to Thomas' dictum that one must cherish the people's spirit but "keep alive their attention." (B01 370-400)

In the corpus, a number of these types of modifiers (10%) were unpunctuated.

In theory, there are a variety of different ways that punctuation can be used to distinguish homonyms or to prevent ambiguities. For instance, punctuation can be used to distinguish various types of homonymous adverbs, specifically adjuncts that can function also as conjuncts or disjuncts. In the examples below, punctuation distinguishes the manner adjunct *naturally* from the attitudinal disjunct *naturally*:

> I expect my dog to behave naturally.
> I expect my dog to behave, naturally. (adapted from Greenbaum 1969:183)

In the corpus, however, distinctions such as the above were not regularly maintained by punctuation. One reason that punctuation did not regularly distinguish homonyms is that frequently the homonymous forms do not occur in the same position. The disjunct *clearly* in the following sentence was not punctuated, simply because the manner adjunct *clearly* will rarely, if ever, appear initially in a sentence; manner adjuncts in general tend to occur in the final position of the clause (Quirk et al. 1985:495):

> All evening Anthea favored him with odd, coy looks. *Clearly* she had been instructed "not to say a word." (P28 570-590)

Since there is no chance of confusing the adjunct *clearly* with the disjunct *clearly* in the above example, a comma following *clearly* would be redundant.

Other homonyms, on the other hand, regularly occur in the same position. Yet in the corpus, punctuation did not always distinguish them. Temporal *while*, for instance, was usually not punctuated, whereas concessive *while* usually was. However, there were instances when concessive *while* was not punctuated and when temporal *while* was:

> Since electrical stimulation of the posterior hypothalamus produces the effects of wakefullness *while stimulation of the anterior hypothalamus induces sleep*, it may be said that the reactivity of the whole organism is altered by a change in the autonomic reactivity of the hypothalamus. (J17 640-680)

> The callous marines had laughed at each other's retching, *while stacking bodies*. (N25 440-450)

Similarly, there were instances when the conjunct *now* was unpunctuated:

> Suppose, says Dr. Lyttleton, the proton has a slightly greater charge than the electron ... This would give the hydrogen atom a slight charge excess. *Now* if one hydrogen atom were placed at the surface of a large sphere of hydrogen atoms, it would be subject both to the gravitation of the sphere and the charge-excess of all those atoms in the sphere. (C13 580-650)

and the adjunct *now* was punctuated:

> [Dr. Conant's] earlier reports considered the American public schools basically sound and not in need of drastic change. *Now*, a close look at the schools in and around the ten largest cities, including New York, has shattered this optimism. (B07 1370-1400)

While punctuation in the corpus rather ineffectively created semantic effects, it far more effectively reinforced them. Punctuation was especially effective at indicating the degree to which constituents were semantically integrated. Punctuation indicated the degree of semantic integration between syndetically coordinated clauses; asyndetically coordinated clauses and phrases; and adverbials and the clauses in which they occurred.

The particular coordinator used to conjoin two main clauses directly affected their punctuation. Since the conjunction *and* indicates that the clauses it conjoins are rather closely integrated, sentences containing clauses joined by *and* were punctuated 79% of the time with a light mark of punctuation (a comma) or with no punctuation:

> Everybody fell in love with Amy again last night at the Warwick Musical Theater, and Shelly Berman was to blame. (C04 690-710)
>
> Public school children have adopted the fund as one of their favorite Christmas charities and their pennies, nickels, dimes and quarters aid greatly in helping Santa to reach the fund's goal. (A24 1140-1170)

On the other hand, the conjunction *but* indicates that the clauses it conjoins are less closely integrated. Hence, main clauses linked by this conjunction were separated 87% of the time by heavier marks of punctuation, commas or periods:

> The visceral brain as well as the neocortex is known to contribute to memory, but this topic is beyond the scope of this paper. (J17 300-320)
>
> Seeming to have roots in the soil, they actually have none in life. They dwell, in short, in the doltish twilight in which peasants and serfs of the past are commonly reported to have lived. But this is a theme which does not take so much time to state as Mr. Wisker dedicates to it. (J66 280-310)

Although the other coordinators (*or*, *for*, *yet*, *so*, *neither*, and *nor*) did not occur as frequently in the corpus as *and* and *but* did, they nevertheless were punctuated according to the semantic relationships that they expressed.

The coordinators *and* and *but* were punctuated somewhat differently in the individual styles of the corpus. The percentage of unpunctuated sentences coordinated by *and* was much lower in the learned style (11%) than in the journalistic and fictional styles (32% and 27%, respectively). The reason for this difference is that the compound sentences in the journalistic and fictional styles tended to contain shorter main clauses than those in the learned style, a style known for having lengthy and sometimes overly complex sentences. And since length and complexity are reasons for punctuating the clauses of a compound sentence, it is not surprising that so few sentences in the learned style were left unpunctuated.

The learned style also differed from the other styles because it contained fewer instances of periods used to separate the main clauses of sentences conjoined by *and* and *but*. Only 7% of the sentences conjoined by *and* and 25% conjoined by *but* were punctuated with a period. The figures are considerably higher in the other styles. In the journalistic style, *and* and especially *but* were preceded by a period 17% and 66% of the time respectively; in the fictional style, they were preceded by a period 20% and 48% of the time, respectively. The learned style is generally a very formal style and is perhaps more likely to adhere to the prescriptive rule that orthographic sentences should not begin with *and* or *but*.

The particular coordinator chosen in a compound sentence is only one way of indicating the degree to which the clauses of the sentence are semantically integrated. There are three other markers of integration in coordinated constructions: whether the coordinated construction contains main clauses whose subjects are co-referential, whether it contains main clauses whose subjects are ellipted, and whether it consists of phrases or clauses that are asyndetically coordinated.

The main clauses of a non-elliptical sentence will be more closely integrated if they contain subjects that are co-referential. Consequently, in sentences of this type, Quirk et al. (1972:1060) hypothesize that punctuation will be less likely, since punctuation will indicate that the clauses are less integrated rather than more closely integrated. However, in the corpus, just the opposite was true: sentences with co-referential subjects were more likely to be punctuated:

> Here's an idea for a child's room that is easy to execute and is completely charming, using puppets for lamp bases. *Most children* love the animated puppet faces and their flexible bodies, and *they* prefer to see them as though the puppets were in action, rather than put away in boxes. (A30 1500-1540)

This finding suggests a difference between British and American usage.

The clauses of a compound sentence will be most closely related if their subjects are ellipted:

> *The Couperin "La Steinkerque"*, with its battle music, brevity, and wit and refined simplicity, already shakes off Corelli and [subject ellipted] points toward the mid-century elegances that ended the baroque era. (C07 1300-1320)

In the corpus, the clauses of this type of compound sentence were punctuated 13% of the time if the coordinator was *and* and 64% of the time if the coordinator was *but*. Those clauses containing *but* were more frequently punctuated because the adversative nature of *but* overrode the effects of ellipsis. Those containing *and* were most frequently not punctuated, because the semantics of *and* and the effects of ellipsis make the clauses more closely integrated, an integration better reflected with no punctuation.

All of the coordinated constructions discussed thus far have contained syndetic coordination: clauses overtly conjoined by a coordinating conjunction. It is also possible, however, for the main clause of a compound sentence not to be overtly conjoined by a coordinating conjunction. This type of coordination is known as asyndetic coordination, and in the corpus the missing coordinator was usually replaced by a semicolon and less frequently by a comma.

Most style manuals that discuss the punctuation of juxtaposed main clauses state that a semicolon can be used to replace any missing coordinator. However, in the corpus, the semicolon replaced a potential *and* most frequently (65% of the time) and *for*, *but*, and *so* only rarely (35% of the time):

> Siepi was, as always, a consummate actor; [and] with a few telling strokes he characterized Alvise magnificently. (C07 590-600)

The remaining coordinators – *or*, *yet*, *neither*, and *nor* – were never replaced by semicolons.

Many style manuals state that an additional use of the semicolon is to juxtapose compound sentences whose second clauses contain conjuncts such as *thus*, *consequently*, or *however*:

> We shall not be able to entirely pass over these connections to the East as we consider Ripe Geometric pottery, the epic and myth, and the religious evolution of early Greece; the important point, however, is that these magnificent achievements, unlike those of later decades, were only incidentally influenced by Oriental models. (J54 1390-1440)

However, in the corpus, only five sentences of this type (6%) were juxtaposed with semicolons. The majority were simply separated with periods.

There are two further types of constructions that can be asyndetically coordinated: a series of three or more constituents (the first example below; hereafter Series 1) or a series of two or more adjectives premodifying nouns (the second example below; hereafter Series 2):

> Their collaboration in the Beethoven Second Symphony was lucid, intelligent and natural sounding. (C07 90-110)
>
> The Vanguard album *Madrigal Masterpieces* ... is a good sample of the special, elegant art of English madrigal singing. (C07 700-720)

In Series 1, sometimes the comma before the final coordinator was omitted (as in the first example above), but this depended on the style in which the construction occurred. The A, B and C pattern occurred in the learned style; the A, B and C pattern occurred in the journalistic style; and both patterns occurred in the fictional style.

Style manuals specify that adjectives in Series 2 be separated by a comma only if the adjectives could be conjoined by *and* or have their order reversed. However, these types of adjectives were not punctuated in the corpus according to the prescriptions of style manuals. That is to say, while many adjectives (85%) were separated by a comma when they could be conjoined by *and* or have their order reversed:

> Then the audience saw a small, dim figure appear at the edge of the presidential box. (K05 1200-1210)
>
> Then the audience saw a small and dim figure appear ....
>
> Then the audience saw a dim, small figure appear ....

there was a high percentage of adjectives (37%) separated by commas that could not be conjoined by *and* or have their order reversed. In the example below, both *the vast and dungeon kitchens* and *the dungeon and vast kitchens* are impossible:

> The stained glass windows may have developed unpremeditated patinas, the paneling may be no more durable than the planks in a political platform. The vast, dungeon kitchens may seem hardly worth using except on occasions when one is faced with a thousand unexpected guests for lunch. (C01 1110-1150)

It appears that the rule for punctuating adjectives in a series has been extended to cover all adjectives, not just those in a series.

The degree to which an adverbial was integrated into clause structure affected its punctuation in two positions: initial position (IP) and final position (EP). In medial position (MP), the punctuation of an adverbial was more affected by prosody (cf. the next section). Adjuncts that were words or phrases were unpunctuated 54% of the time because they were closely integrated into the clauses in which they occurred:

> Dolores smiled; she let the interpretation stand. *Now* Martin heard himself give a snort of mock good nature. (P28 1010-1030)

Disjuncts and conjuncts, on the other hand, are sentential adverbs. Hence, they were unpunctuated only 36% of the time, their relatively frequent punctuation reflecting their loose connection to the clauses of which they were members:

> *Ideally*, brief treatment should be arrived at as a treatment of choice rather than as a treatment of chance. (J24 1780-1800)

Adverbial clauses were almost always punctuated, regardless of their degree of clausal integration, because they are lengthy and complex. However, of the 22 adverbial clauses in IP that were unpunctuated, all were adjunct clauses like the following:

> He was fuzzy in his mind and, for a moment, helpless on the lobby floor, but he was conscious, and free of the weight of Roberts' body. *When his vision cleared* he saw the taller one scramble upward, reaching. (L06 170-200)

The degree to which an adverbial was integrated also affected its punctuation in EP. Subordinate clauses headed by *when*, *because*, conditional *if*, etc. can be adjuncts, and when they were, they were unpunctuated 79% of the time in EP:

> Either way [the budget increase] sounds like a sizable hunk of money and it is. But exactly how far it will go toward improving conditions is another question *because there is so much that needs doing*. (B01 1090-1120)

Subordinate clauses headed by causal *since*, concessive *while*, *whereas*, *although*, etc., on the other hand, are disjuncts and hence were unpunctuated only 25% of the time in EP:

> "Much Ado [About Nothing]" turned serious while the insipid Claudio rejected Hero at the altar, *although some umbrellas were opened*. But the rain came more heavily, and men and women in light summer clothes began to depart. (C13 1530-1560)

Prosody and punctuation

In the previous section, it was demonstrated that punctuation was a relatively weak marker of semantic relations because it was far better at reinforcing semantic relations than at creating them. The relationship of punctuation to prosody is very similar: while punctuation can impose a prosodic structure on the written text, it is far better at reinforcing the positions in the written text where some prosodic juncture would occur in speech. In short, punctuation is at best a rather crude reflection of the complexities of prosody, and although there exist instances of punctuation that are prosodically motivated, the relationship between punctuation and prosody is weak and unsystematic.

The relationship is unsystematic because not all instances of punctuation have some prosodic correlate. There are numerous instances where we pause but do not punctuate and, conversely, where we punctuate but do not pause. In the examples below, pauses occur following the subject in the first example and preceding the *that*-clause in the second example. Yet current practice prohibits our punctuating these junctures (Quirk et al. 1985:1619).

> *Those who are fond of sleeping late, make unreliable workers.

> *It should soon become quite apparent, that current U.S. policy in Central America is a failure.

In the next examples, punctuation occurs where no pause in speech would. In the first example, no pause would occur between *that* and *if*. In the second example, a pause would occur after *too* but not before it.

> Perhaps it was insane, Pamela thought. Perhaps it was all a vividly conceived dream. But she was caught in it, and she faced the terrible possibility *that*, *if* it were a dream, it was one from which she might never awaken. (N08 440-470)

> I was delighted with Paula Prentiss' comedy performance, which was as fresh and unstilted as one's highest hopes might ask. A couple of the males made good comedy, *too* – Jim Hutton and Frank Gorshin. (C04 1760-1790)

In the corpus, only one construction, the adverbial, was punctuated according to its prosodic manifestations in speech. In speech, adverbials have various prosodic patterns that in the corpus were mirrored by punctuation: those that optionally constituted separate tone units in speech were either punctuated or unpunctuated; those that generally did not constitute separate tone units were usually not punctuated; and those that always constituted separate tone units were usually punctuated.

Many adverbials, such as *thus*, *therefore*, and *in fact*, can optionally be tonally integrated into the clauses in which they occur. In the corpus, these adverbials were either punctuated or unpunctuated, punctuation having the effect of imitating in writing the situations in speech when these adverbials constituted separate tone units:

> It is our belief that his readiness to relinquish some control was evidenced by the Kohnstamm-positive subjects in some of the other experimental situations to be discussed below. *Thus*, this readiness to relax controls, evidenced in the Kohnstamm situation, appears to be a more general personality factor. (J28 1630-1680)

> Since rococo music tends to be pretty and elegant above all, it can seem rather vacuous to twentieth-century ears that have grown accustomed to the stress and dissonances of composers from Beethoven to Boulez. *Thus* there was really an excess of eighteenth-century charm as one of these light-weight pieces followed another on Saturday night. (C07 1520-1570)

In contrast to adverbials that can optionally occupy a separate tone unit are those that rarely or never do so. Ninety-one percent of these adverbials were not punctuated in the corpus:

> Some other good bills were also lost in the shuffle. *Certainly* all can applaud passage of an auto title law, the school bills, the increase in teacher pensions .... (B01 150-210)

They were punctuated only to imitate in writing the equivalent of emphatic intonation in speech:

> At the end of the monologue, the audience would applaud ... There was always a pause here, *before the next line*. (K05 1120-1160)

The punctuation in the example above is quite conspicuous. Consequently, this use of punctuation occurred in the corpus only when writers wished to create stylistic effects with punctuation.

The final group of adverbials includes those that obligatorily occupy a single tone unit. This group consists of a wide variety of adverbials occurring in all positions in a sentence or clause, adverbials which were punctuated 76% of the time:

> Freedom of the press was lost in Cuba because of decades of corruption and social imbalances. In such conditions all freedoms are lost. This, *in more diplomatic language*, is what Adlai Stevenson told the newspapermen of Latin America yesterday .... (B07 980-1020)

> The external signs of his approach to it would be covered by the snow, *probably by the next day*. (L06 1420-1440)

Conclusions

This study focused on the practice of American punctuation in the Brown Corpus and demonstrated that punctuation involves the complex interaction of syntax, semantics, and prosody. Since this study was restricted to a discussion of American punctuation practice, what is needed is a study, based on the LOB Corpus, that compares American practice with British practice and that determines the extent to which British punctuation has syntactic, semantic, and prosodic functions.

Notes

[1] Frequency counts alluded to but not discussed in this paper are described in greater detail in Meyer (1983).

[2] In some examples (unless otherwise indicated), sections of sentences under discussion have been italicized.

References

Baldwin, R. and J. Coady. 1978. Psycholinguistic approaches to a theory of punctuation. *Journal of Reading Behavior* 10:363-375.

Greenbaum, S. 1969. *Studies in English adverbial usage*. London: Longman.

Limaye, M. 1983. Approaching punctuation as a system. *The ABCA Bulletin*, 28-32.

Meyer, C. 1983. A descriptive study of American punctuation. Unpublished Ph.D. thesis, University of Wisconsin-Milwaukee.

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

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

Summey, G. 1949. *American punctuation*. New York: The Ronald Press Co.

MATERIAL AVAILABLE FROM BERGEN

The following material is currently available on computer 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, untagged version, 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). The text of the LOB Corpus is not available on microfiche.

LOB Corpus, untagged version, 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.

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

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

LOB Corpus, tagged version, KWIC concordance (also on microfiche): A complete concordance for all the words in the corpus, sorted by key word and tag. At the beginning of each graphic word there is a frequency survey giving the following information: (1) total frequency of each tag found with the word, (2) relative frequency of each tag, and (3) absolute and relative frequencies of each tag in the individual text categories.

London-Lund Corpus, text: The London-Lund Corpus contains samples of educated spoken British English, in orthographic transcription with detailed prosodic marking. It consists of 87 'texts', each of some 5,000 running words. 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 computerized and printed form (J. Svartvik and R. Quirk (eds.) *A Corpus of English Conversation*, Lund Studies in English 56, Lund: C.W.K. Gleerup, 1980).

London-Lund Corpus, KWIC concordance II: 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 tagged versions of the Brown Corpus cannot be obtained from Bergen.*

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

CONDITIONS ON THE USE OF ICAME CORPUS MATERIAL

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

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

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

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

1 No copies of corpora, or parts of corpora, are to be distributed under any circumstances without the written permission of ICAME.

2 Print-outs of corpora, or parts thereof, are to be used for bona fide research of a non-profit nature. Holders of copies of corpora may not reproduce any texts, or parts of texts, for any purpose other than scholarly research without getting the written permission of the individual copyright holders, as listed in the manual or record sheet accompanying the corpus in question. (For material where there is no known copyright holder, the person(s) who originally prepared the material in computerized form will be regarded as the copyright holder(s).)

3 Commercial publishers and other non-academic organizations wishing to make use of part or all of a corpus or a print-out thereof must obtain permission from all the individual copyright holders involved.

4 The person(s) who originally prepared the material in computerized form must be acknowledged in every subsequent use of it.

EDITORIAL NOTE

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. Write to: Stig Johansson, Department of English, University of Oslo, P.O. Box 1003, Blindern, N-0315 Oslo 3, Norway.

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