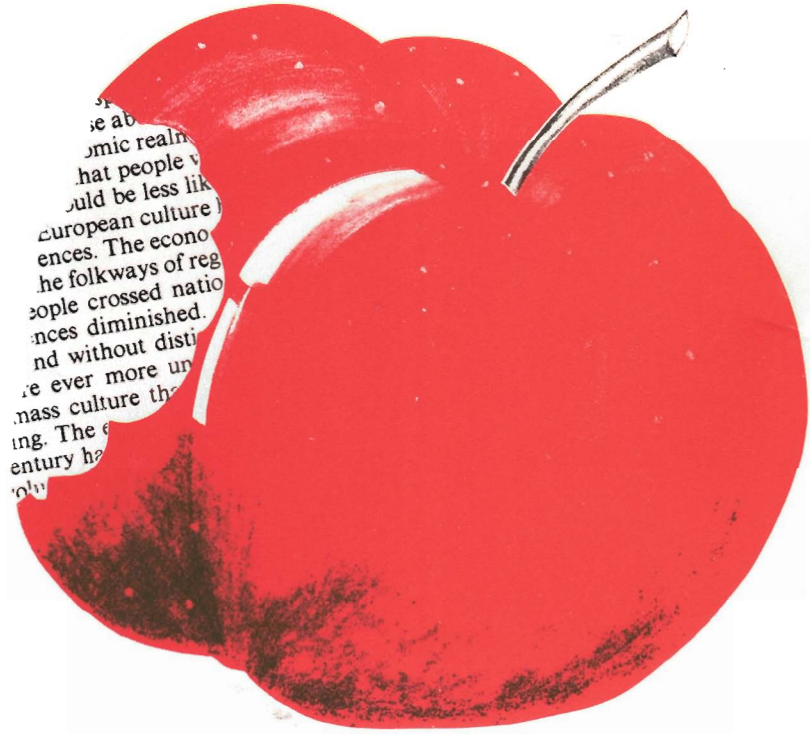
ICAME Journal

International Computer Archive of Modern English

No. 13

April 1989



NORWEGIAN COMPUTING CENTRE
FOR THE HUMANITIES

CONTENTS

Articles:

Conference report:

Reviews:

Shorter notices:

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

The Shakespeare Dictionary Database

H. Joachim Neuhaus
Westfälische Wilhelms-Universität, Münster

1. The Shakespeare Dictionary Database project at Münster has been using structured databases since 1983. The databases are implemented on a dedicated PR1ME computer using standard CODASYL-DBMS software and related tools (cf. Neuhaus 1985a, 1986, 1989). Initially, the project was part of the "Sonderforschungsbereich 100 Elektronische Sprachforschung" sponsored on the national level by the Deutsche Forschungsgemeinschaft. The project will publish first results in 1989. A first volume of a series of reference works is tentatively entitled: *Shakespeare's Wordforms. Inflection, Word-Formation, and Etymology.* There are also plans to publish certain sections of the database on CD-ROM later on.

2. The database integrates available material in computer-readable form, but it may perhaps better be characterized as a second major step in the computational analysis of Shakespeare after the earlier successful use of computers for concordancing and collating in the late sixties (Spevack 1968, Howard-Hill 1969, Widmann 1971). Most importantly, the new research environment of a comprehensive database has already done away with the virtual compartmentalization of textual studies, editing, annotation, and lexical or syntactical analysis. Instead of separate projects and isolated investigations, the database environment provides a consolidated perspective. Spevack's *Complete and Systematic Concordance to the Works of Shakespeare* (1968-1980) and Finkenstaedt, Leisi and Wolff's *A Chronological English Dictionary* (1970), both in machine-readable form, were used in the initial computer-assisted lemmatization procedure which produced the needed dictionary entries for further analysis (Spevack, Neuhaus, and Finkenstaedt 1974).

Spevack's concordances are based on the Riverside text (Evans 1974). In order to present the complete Shakespearean vocabulary and to disengage the database from dependence on a single edition of Shakespeare, the textual data were expanded to include stage directions and speech-prefixes in all quartos up to and including the First Folio (cf. Spevack, *Complete and Systematic Concordance to the Works of Shakespeare*, Volume VII), and the "bad" quartos (Volume VIII), as

well as substantive variants (Volume IX), producing a composite Shakespearean vocabulary in modern and old spelling.

The *Chronological English Dictionary* was the first of its kind. The chronological information is based on the *Shorter Oxford English Dictionary* and various published collections of antedatings. Its entries are sorted according to the year of first occurrence in print. It is thus possible to "stop" the development of the recorded English vocabulary at any desired moment and to compare, for instance, Shakespeare's vocabulary with a large corpus of English words recorded up to 1623, when the First Folio appeared (Neuhaus 1978). The data of the *Chronological English Dictionary* have in the meantime been supplemented by information from the *Michigan Early Modern English Materials* (Bailey et al. 1975) and the original *Oxford English Dictionary* itself.

3. As was to be expected, the lexical part of the database has transcended the familiar paradigm of an author dictionary as a separate reference work, such as Schmidt/Sarrazin (1902) or Onions/Eagleson (1986). The database will further the study of Shakespeare's position in Early Modern English since it also makes accessible large parts of the complement vocabulary – i.e. from other written sources – attested for his time. The set of words in Shakespeare can be compared with the complement set of words available in Elizabethan English but not attested in Shakespeare's works. In this way there is a systematic integration into the total vocabulary. As a result, the database model can easily be expanded to cover other Elizabethan playwrights or various other registers of Early Modern English.

The database has proved to be a valuable editorial tool as well since its data and relations go beyond the formal concerns of both old-spelling and modern-text concordances. The database has access to variant readings of the texts and ultimately to electronic facsimiles of printed pages from early Shakespeare editions on graphics terminals (cf. Neuhaus 1985b). It will then be possible to scrutinize and collate citations in quarto and First Folio pages in a way not practicable in mechanical collation (e.g. Hinman 1968).

4. A primary interest in database design is, of course, to make previously inaccessible information accessible. More important still is the goal of processing information in such a way that new, previously unexpected properties and relations become available for systematic analysis. In the understanding of database information-handling the notion of a value-added chain has proved to be quite helpful. In such a chain

certain key steps can be distinguished, such as collecting, storing, processing, transforming, and disseminating. The processing part is the place for database design. Unlike most other databases, the Shakespeare Dictionary Database has certain peculiarities at both ends of the information chain. The collecting of raw data, namely early editions, has been completed, and the database itself will be a fairly stable information source with only very rare updatings.

5. In the analysis of the Shakespeare data a strict differentiation between database entities has been observed. All entities are systematically linked. It is thus possible to "navigate", as a terminological metaphor quite expressively suggests, from a First Folio citation to a respective modern edition, and ultimately to a dictionary entry, or lemma. Such a chain of links begins with the identification of text tokens on the lower levels of copy text and edited text, which may, of course, be quite different. But this is a strictly formal process. Lexical parsing is necessary on the next level in order to disambiguate text tokens and obtain correctly tagged wordforms. This process is often called lemmatization because it is a prerequisite for a grouping of wordforms under the lemma they belong to. For verbs and nouns, for example, such a grouping of wordforms under a lemma has traditionally been understood as an inflectional paradigm. Of course, it is also possible to go back from a particular lemma to all of its token occurrences in the corpus, with detailed information about spelling variation in particular source texts or emendational readings.

In *The Winter's Tale*, for example, two items in the passage TLN 2862-4 may be used to illustrate the links from text token to lemma and back again, namely *will* and *love*:

> Women will loue her, that she is a Woman
> More worth then any Man: Men, that she is
> The rarest of all Women.

The First Folio serves as textual source for the edited text, which in turn is parsed and disambiguated to yield actual wordforms which can then be assigned to their respective lemmata. Both items in the example are ultimately linked to the verb forms, and are thus kept distinct from the respective noun homographs. The data-structure can be visualized with entity-relationship diagrams such as the one given as Figure 1.

6. Each database entity has a set of further information categories. Chronological information, for example, is part of the lemma record.

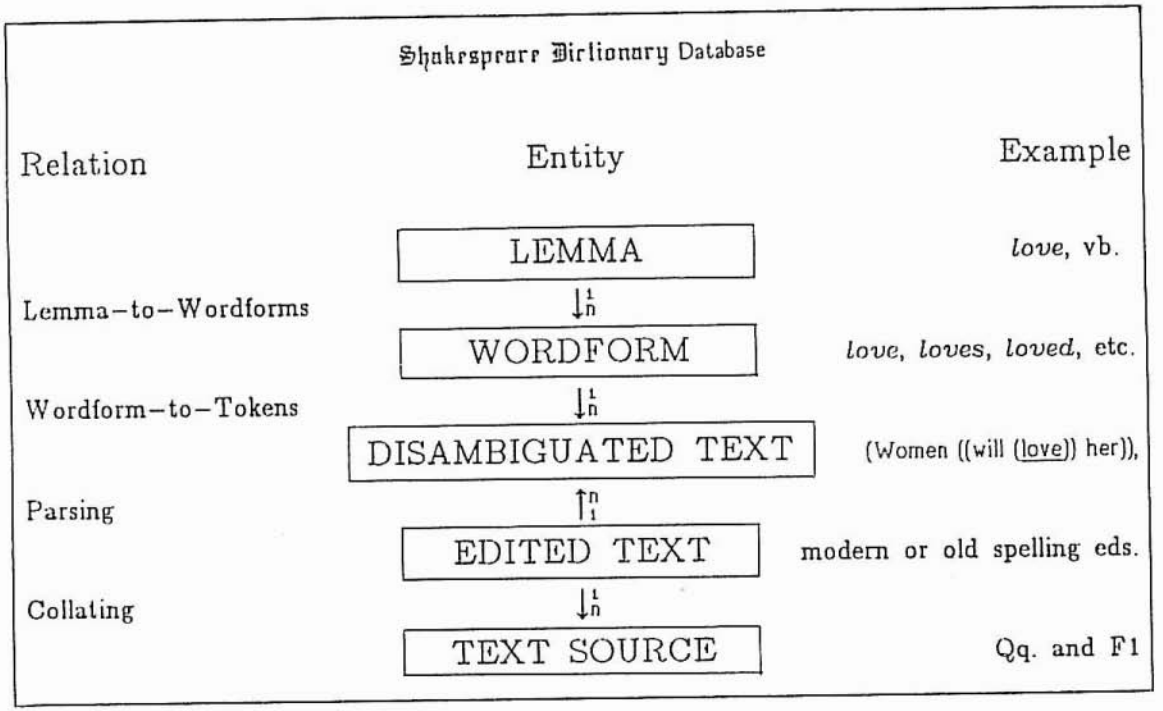


Figure 1 Database design: from text to lemma

It can be accessed to provide lists of new words for specific plays such as *The Winter's Tale*, and it can be used to study the time depth of Shakespeare's vocabulary. Only ca. 20% of Shakespeare's vocabulary is contemporary vocabulary not attested prior to 1588. More than 60% of Shakespeare's vocabulary can be traced back to Old or Middle English (cf. Figure 2). The etymological data include word histories and loan relations, again supplemented by chronological data. Etymological information uses a more systematic grouping than in, for example, the *Oxford English Dictionary* and its etymological derivatives. There are, in fact, more than thirty types of lexical information coded in the lemma-record database area.

The vocabulary of the Shakespeare corpus has two subsets which would normally not be admitted to monolingual dictionaries: foreign words and proper names. The current database version contains 551 foreign lemmata, mainly French and Latin, which occur in foreign contexts. There are 1,673 proper names, not counting variants as distinct names, or counting individual persons or places with the same name as distinct.

7. Whereas the lemma is normally taken as the lexical unit par excellence, wordforms traditionally belong to grammar. The database links both levels systematically, but it may still be helpful to understand the wordform level as a complete inflectional morphology, a "grammatical window" to the database. Single inflectional categories can be studied quantitatively. The 4,217 inflected noun plural forms make up the largest class of inflected forms. The noun plural *loves* is one of these. *Love* is used twice as a proper name. There are 7 contractions: *love's*. Substantive textual variants can be listed with their act-scene-line reference, down to the level of quarto and First Folio orthography, e.g. *Ioue* 1H6 5.5.82 (F1) and *Lord* HAM 3.2.169 (Q2). Wordform entries make further variational phenomena accessible, such as the option between different inflectional allomorphs in the verb paradigm. In the present tense there are alternative forms for the second person (*loves* versus *loveth*, and there are both inflected and uninflected past tense forms (*lovedst/lov'dst* versus *loved*) (cf. Neuhaus 1981). This information can, of course, be systematized for a structural description of Shakespeare's inflectional usage as part of a Shakespeare grammar.

8. The database allows for further analyses on the vocabulary level. There is a complete morphology for all lemmata which gives detailed structural descriptions of derivations, compounds, and other combina-

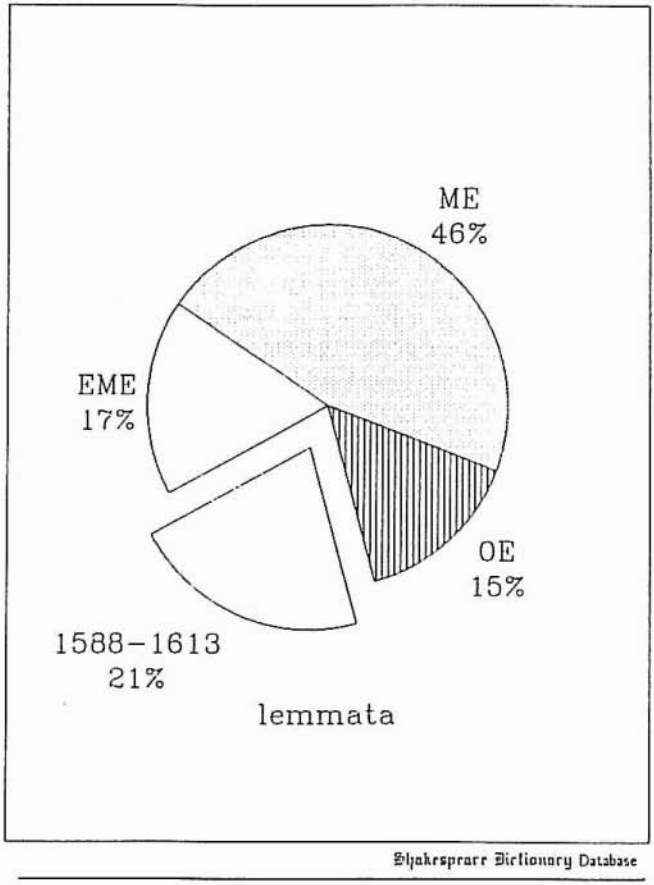


Figure 2 Chronological segments of Shakespeare's vocabulary

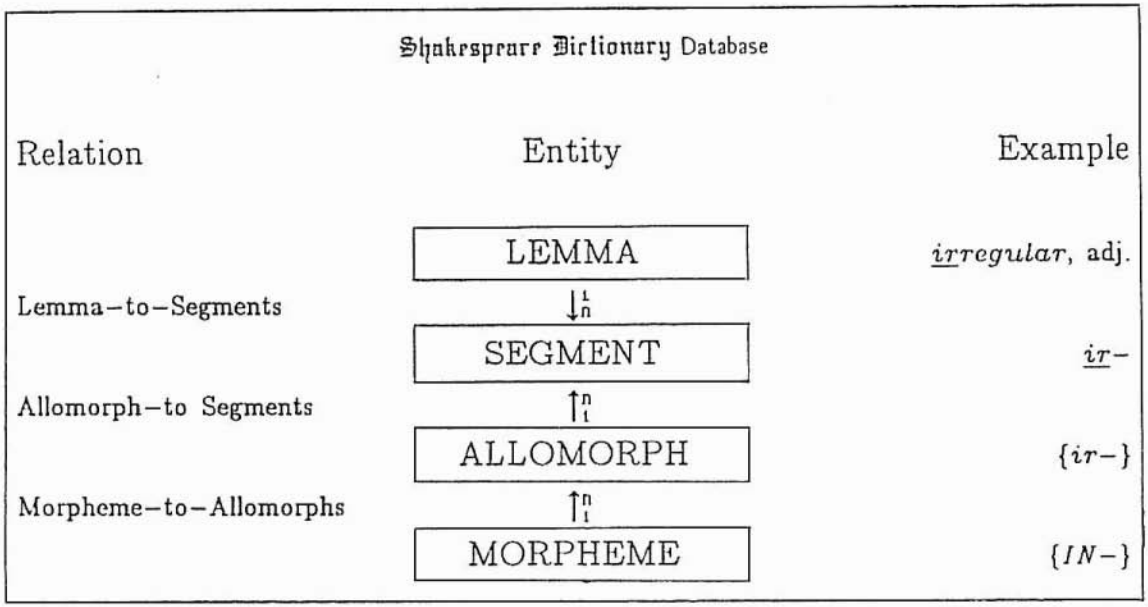


Figure 3 Database design: from lemma to morpheme

tions, as well as all inflected wordforms, as they occur in the text. In order to understand the database structure with its systematic links between information entities it may be profitable to look at the architecture of the morphological area of the database and compare the underlying database schema with some actual sample queries taken from the morphology database. There are four base object-classes (entities): lemmata, segments, allomorphs, and morphemes having cardinality values between 4,500 and 40,000 records. Queries allow for a direct retrieval on three levels: the conventional level of the lemma, the level of allomorphs, and the morphemic level. This is achieved by a virtual record, defined as a subschema. In this way the database design mirrors a structural morphological analysis directly. The concept of a morphological family, defined as a set of lemmata which have at least one morpheme in common, is thus immediately accessible for database queries.

An example may further illustrate the concept of a morphological family. The ultimately Latin prefix morpheme {IN-} has database links to allomorphs such as {im-} which in turn occurs as a segment in the lemma *impure*. The allomorph {il-} occurs as a segment in the lemma *illegitimate*, and the allomorph {ir-} as a segment in the lemma *irregular* (cf. Figure 3). In Shakespeare's vocabulary there are almost 200 lemmata which belong to this {IN-} family. Of these, 25 lemmata have their first recorded citation in Shakespeare, such as *immediacy*, n. in *King Lear*, or *disinsanity*, n. in *Two Noble Kinsmen*, if the chronology of the Riverside edition is used for dating the plays. This morphological material as well as other data structures such as inflection, etymology, and others are currently being prepared for the first volume of specific printed selections from the Shakespeare Dictionary Database.

References

Bailey, Richard W. et al. 1975. *Michigan Early Modern English Materials*. Ann Arbor: Xerox University Microfilms.

Evans, Gwynne Blakemore. 1974. *The Riverside Shakespeare*. Boston: Houghton Mifflin.

Finkenstaedt, Thomas, Ernst Leisi and Dieter Wolff. 1970. *A Chronological English Dictionary. Listing 80,000 Words in Order of their Earliest Known Occurrence*. Heidelberg: Winter.

Hinman, Charlton. 1968. *The First Folio of Shakespeare*. New York: Norton.

Howard-Hill, Trevor Howard. 1969ff. *Oxford Shakespeare Concordances to the Text of the First Folio*. Oxford: Clarendon.

Neuhaus, Heinz Joachim. 1978. "Author Vocabularies Compared with Chronological Dictionaries", *Bulletin of the Association for Literary and Linguistic Computing* 6, 15-20.

Neuhaus, Heinz Joachim. 1981. "Wortbildungsstrukturen bei Shakespeare," *Shakespeare Jahrbuch West* 117, 26-31.

Neuhaus, Heinz Joachim. 1985a. "Die Münsteraner Shakespeare Wortdatenbank," *Prime Time* 4, 17-19.

Neuhaus, Heinz Joachim. 1985b. "Design Options for a Lexical Database of Old English". In Alfred Bammesberger, ed., *Problems of Old English Lexicography. Studies in Memory of Angus Cameron*. Regensburg: Pustet. 197-209.

Neuhaus, Heinz Joachim. 1986. "Lexical Database Design: The Shakespeare Dictionary Model". In Winfried Lenders et al., eds., *Proceedings of the 11th International Conference on Computational Linguistics*. Bonn: IPK. 441-444.

Neuhaus, Heinz Joachim. 1989 (forthcoming). "Designing Lexical Databases", *Lexicographica* 4, Tübingen.

Onions, Charles Talbot and Robert Eagleson. 1986. *A Shakespeare Glossary*. 3rd ed. Oxford: Clarendon.

Schmidt, Alexander and Gregor Sarrazin. 1902. *Shakespeare-Lexikon. Vollständiger englischer Sprachschatz mit allen Wörtern, Wendungen und Satzbildungen in den Werken des Dichters*. 3rd ed. Berlin: Reimer.

Spevack, Marvin. 1968-1980. *A Complete and Systematic Concordance to the Works of Shakespeare*. 9 volumes. Hildesheim: Olms.

Spevack, Marvin, Thomas Finkenstaedt and Heinz Joachim Neuhaus. 1974. "SHAD: A Shakespeare Dictionary". In John Mitchell, ed., *Computers in the Humanities*. Edinburgh: Univ. Press. 111-123.

Widmann, Ruth. 1971. "The Computer in Historical Collation: Use of the IBM 360/75 in Collating Multiple Editions of A Midsummer Night's Dream". In Roy Albert Wisbey, ed., *The Computer in Literary and Linguistic Research*. Cambridge: Univ. Press. 57-63.

Progress report on the diachronic part of the Helsinki Corpus[1]

Merja Kytö
University of Helsinki

The texts selected for the basic part of the *Helsinki Corpus*, intended to cover the Old, Middle and early Modern British English periods, have now been keyed in with textual parameter codes on floppy disks and transferred to the mainframe for preliminary use. At the moment the basic part totals 1.5 million words (the figure excludes passages in foreign languages, and our own and the editor's comments);[2] the supplementary part, Scots (at present 300,000 words) and early American English (at present 155,000 words), will be added to the basic corpus in due course:

BASIC PART

	Words
Old English	395,600
Middle English	596,900
EModE, British	529,400
	1,521,900

The following table will give a rough idea about the distribution of the words among the different sub-periods:

OLD ENGLISH

		Words	
I	- 850	2,000	0.5 %
II	850- 950	86,500	21.9
III	950-1050	243,400	61.5
IV	1050-1150	63,700	16.1
		395,600	100.0

MIDDLE ENGLISH

		Words	
I	1150-1250	108,000	18.1 %
II	1250-1350	96,400	16.2
III	1350-1420	181,400	30.4
IV	1420-1500	211,100	35.3
		596,900	100.0

EModE, BRITISH

		Words	
I	1500-1570	179,300	33.9 %
II	1570-1640	189,200	35.7
III	1640-1710	161,000	30.4
		529,500	100.0

As regards the programs and applications, our main tools have been the OCP[3] and WordCruncher[4] program packages. For our purposes the possibility of modifying the standard character file is necessary: with WordCruncher we use a special Helsinki Corpus character file, which defines our double character letters (*t and *T for lower and upper case thorn, *d and *D for eth etc.) and special characters used in coding conventions (equal sign for superscript as in *ou=r=*, tilde for abbreviations as in *co~stantly* for *constantly* etc.). Similarly, it is possible to define these special characters in the OCP command file. One of our code lines, which follow the OCP Cocoa format, gives condensed information on the main textual parameters of a text and identifies an example retrieved from the data. This code line (as well as any of the other code lines) can easily be converted into a WordCruncher reference code with the help of an editor or a simple conversion program.

The basic part of the Corpus (about 10 MB) was indexed with WordCruncher (with no Stopword file) and is now available as one Computer Book in the BookShelf. Similarly, the material for the main periods (Old, Middle and early Modern British English) and the sub-periods within each main period are available in single books. The division of the Books according to the (sub-)periods is an advantage for a researcher who wants speedy access to the material. Needless to say, what really makes the use of WordCruncher worthwhile is the opportunity it offers to have direct access to the wider context in the running text. The advantage of the OCP, on the other hand, is that the

SELECT command makes it possible to benefit fully from the textual parameters and retrieve a set of examples which represent a combination of parameter values tailored by the researcher according to his aims and needs.

Our experience with the WordCruncher files made us quite uneasy about how firmly we can rely on the accuracy of the text of our text files. We found several typing errors in our coding conventions, even when only a number of the reference categories had been entered. It seems that the human eye, despite the best intentions, is not alert enough when having to go through large amounts of text requiring this type of mechanical accuracy. Originally, the transfer from CP/M system to the MS-DOS system created problems as regards the differences between the versions of the WordStar word processing program we had been using. A filter program (and later, with the WordPerfect program, the Ctrl+F5 option saving a file as a DOS-file) was needed to make sure that the files to be transferred to the mainframe were pure Standard-ASCII character files.

We also used a special debugger program, which checked, as far as was automatically possible, that our coding conventions followed the right format (i.e., that double brackets opening a sentence coded as italics, for instance, are also closed accordingly). Errors of this kind, of course, may prove fatal when running concordance programs based on string identification. The program was devised by Mr M. Jankowski of the Adam Mickiewicz University in Poznan, Poland.

Mr Jankowski also helped us to set up the ChiWriter multifont word processor program:[5] this program allows us to reconvert quite easily our double characters back into original Old and Middle English characters, both on the screen and in print-outs.

We hope to do the first proof-reading of the corpus in spring 1989 in order to polish up the files by the end of the year. We have also sketched out the list of contents for the manual, and hope to have the guide ready to accompany the tape and diskettes. We shall shortly contact the publishers about copyright settlements. In the meantime, the material is available for researchers in Helsinki, and we welcome those interested to join our team.

Most of the pilot studies carried out on the corpus material are mentioned in our contribution to the *Proceedings* of last year's ICAME conference.[6]

Notes

1. This is a revised version of a talk given at ICAME 9TH, Birmingham, May 1988.
2. The Corpus totals 1.6 million words with passages in foreign languages, and our own and the editor's comments included. The figures are based on the dates of the manuscripts of the texts. The counts, completed in November 1988, were obtained with OCP (version 2.3 for mainframe use, see *Oxford Concordance Program, Users' Manual, Version 2*, comps. S. Hockey and J. Martin, Oxford University Computing Service, Oxford, 1988).
3. See Note 2; *Micro-OCP,* Oxford University Computing Service, Oxford: Oxford University Press 1988.
4. See *WordCruncher. Text Indexing and Retrieval Software. Version 4.1*. Provo, Utah: Brigham Young University and Electronic Text Corporation, 1987.
5. *ChiWriter. The Scientific/Multifont Word Processor for the IBM-P.C. (and Compatibles).* Horstmann Software Design Corporation, San Jose, CA 95150, 1987.
6. See "The Helsinki Corpus of English Texts: Classifying and Coding the Diachronic Part", *Corpus Linguistics, Hard and Soft. Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora,* ed. Merja Kytö, Ossi Ihalainen and Matti Rissanen, pp. 169-179, Amsterdam: Rodopi, 1988.

Three problems connected with the use of diacronic corpora[1]

Matti Rissanen
University of Helsinki

In compiling and testing the diachronic part of the Helsinki Corpus of English Texts, our project group has come across three problems which arise from the use of computer corpora in studies of syntax and vocabulary. While these problems are mainly associated with work on diachronic corpora, they may be universal enough to deserve somewhat more general consideration. They could be called "The philologist's dilemma", "God's truth fallacy", and "The mystery of vanishing reliability". The first could be described as pedagogical, the second methodological and the third pragmatic.

"The philologist's dilemma" pertains to the very essence of the use of text corpora in linguistic or philological research. Particularly in the historical study of language, there is a risk that corpus work and computer-supported quantitative research methods will discourage the student from getting acquainted with original texts, from being on really intimate terms with his material and thus acquiring a profound knowledge of the language form he is studying. In the extreme case, this might mean the wane of philologically oriented language studies and result in a great impoverishment in the field of the historical research of language. We would soon be missing the scholars who have a solid, semi-intuitive knowledge of Old and Middle English, based on an extensive reading of original texts. Unquestionably, scholars of this type are the best guarantee of the continuous advancement of our knowledge of the earliest stages of English.

The best way to avoid this risk of impoverishment is constantly to remind ourselves and our students of the importance of reading the texts which form the corpus. Students ought to be trained to see wider textual and extralinguistic contexts, to get a glimpse of the author and society behind the text. It should be our duty to emphasize that, first and foremost, the computer only stores sets of data and organizes and lists them rapidly and efficiently. In the analysis, synthesis and conclusions, the machine does not replace the human brain. We will be able to ask the right questions, draw inferences and explain the phenomena

revealed by our data only if we develop a good overall mastery of the ancient language form we are studying.

In teaching our students to use computer corpora, either individually or in class, we should give sufficient attention to the description of the texts on which the corpus is based. Ideally, a set of these texts should be available in or near the location of the computer facilities, in paper copy and, if possible, in original editions. Some information on the source texts, their character and availability should be included in the corpus manual.

"God's truth fallacy" is, in fact, closely related to the problem discussed above, because it, too, pertains to the student's attitude to his corpus as a research tool. An authoritative corpus may easily create the erroneous impression that it gives an accurate reflection of the entire reality of the language it is intended to represent. This risk is particularly acute with a historical corpus as we are not intuitively aware of its limitations in the same way we are with corpora containing present-day language. If a corpus is intended for one research purpose only, the ill effects of this fallacy are not remarkable, but if it is intended to offer a basis for a variety of studies over an extended period of time – as most corpora are – we ought to be aware of this problem.

One way to avoid the "God's truth fallacy" is to keep the corpus open-ended – to structure it in a way that makes improvement and supplementation easy and uncomplicated. If this can be effected in a way that constantly reminds the user of the unfinished and unclosed state of the corpus, so much the better. Once again, a careful description of the texts, in the manual or in other appropriate contexts, may help to remind the user of the scope and necessary limitations of the corpus: what kind of genres and levels of language he may find in it and, even more significantly, what types of language are not included.

Inevitably, there are problems in keeping a corpus open-ended. The most obvious of these is that the results based on earlier and later versions of the same corpus are not directly comparable. But I regard this as a lesser evil in comparison to the idea of a (necessarily) limited and one-sided corpus giving skewed results and fettering research for decades. In this time of easy communication and ever-improving computer facilities with on-line services, updating the old version and distributing the new one is a simple task. Revised corpus versions would not, of course, be introduced every year; five-year intervals might be appropriate and realistic.

"The mystery of vanishing reliability" is connected with the detailed textual coding attached to, e.g., the Helsinki Corpus. Perhaps paradoxically, this fine-meshed coding, which we have considered an important aim in our corpus project, may also become a problem. The number of parameter values is, of course, inversely proportional to the amount of evidence in each information area sampled. For this reason, particularly in a corpus divided both according to chronology and text type, it may be difficult to maintain the reliability of the quantitative analysis of less frequent syntactic and lexical variants. The problem becomes even more obvious if attention is paid to sociolinguistic parameters.

This problem is discussed at some length by Merja Kytö and myself in an earlier report,[2] and there is no need to repeat the details of that discussion in the present context. We point out that the best way to cope with this problem would be to compile very large corpora (cf. the success of the gigantic Birmingham University International Language Database), but the restrictions of the hardware and software available for linguists set certain limits to the size of the corpus.

Another solution we offer to this problem, applicable to a text-type-sensitive diachronic corpus, is to classify and code the texts according to text categories containing more than one type of text. These larger categories aim at diachronic representativeness and are still highly experimental. In our report, we enumerate nine "diachronic text prototypes". After further study and experiments, we have reduced the number of the categories to five; two of these are divided into two subcategories.[3] To give our diachronic prototypes some theoretical coherence, we are now using an application of Egon Werlich's text type division as the basis of our grouping.[4]

At the moment, our prototypical text categories are the following:

directive (laws, documents)
instructive:

- secular (handbooks, recipes, etc.)
- religious (homilies, sermons, rules, etc.)

argumentative (trials, etc.)
narrative:

- non-imaginative (chronicles, diaries, biographies, etc.)
- imaginative (fiction, romances, etc.)

expository (scientific treatises, philosophy, etc.)

All categories except the argumentative include texts dating from Old, Middle and early Modern English. There are, of course, interesting text types which have not been grouped under our prototypical categories: the most important of these are private and official correspondence and drama texts. Our diachronic corpus also contains samples of Bible translations from Old English to the Authorized Version and translations of Boethius' *De Consolatione Philosophiae* from King Alfred through Chaucer to Queen Elizabeth.

This categorization has, for better or worse, been included in the coding scheme of our corpus. We still regard it as preliminary and liable to further changes. We hope to find out, through pilot studies, whether it is useful synchronically and diachronically – in other words, whether the texts grouped under one and the same category label share relevant linguistic and textual features.

Notes

1. This is a revised version of a talk given at ICAME 9TH, Birmingham, May 1988.
2. "The Helsinki Corpus of English Texts: Classifying and Coding the Diachronic Part", *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora*, eds. M. Kytö, O. Ihalainen and M. Rissanen, 169-179, Amsterdam: Rodopi, 1988.
3. The group working particularly on this problem includes Merja Kytö, Anneli Meurman-Solin, Terttu Nevalainen, Helena Raumolin-Brunberg and Matti Rissanen.
4. E. Werlich, *A Text Grammar of English*. Heidelberg: Quelle and Meyer, 1983 (1982).

The COMMUNAL Project: Extracting a grammar from the Polytechnic of Wales Corpus

Clive Souter
University of Leeds

1. Introduction

The following is a brief survey of current work being conducted at Leeds for the COMMUNAL[1] project by Eric Atwell, Timmy O'Donoghue and myself. The project aims to produce a natural language interface for intelligent knowledge based systems (IKBSs), with written English statements, directives and queries as input, and written/spoken system responses. The scope of English input is intended to include the sort of 'semi-grammatical' constructs common to informal language. The Leeds role in the project is in the interpretation of input, incorporating a morphological-analyser, context-free grammar, parser and semantic translator. A team at University of Wales College at Cardiff are conducting the corresponding work in generation, and are responsible for knowledge representation. This paper presents the work done so far on the development of the grammar and the parser, both of which, it is hoped, can be derived either wholly or partially from a corpus of naturally-occurring English collected at the Polytechnic of Wales by Robin Fawcett.

2. The Polytechnic of Wales Corpus

2.1. Background

The Polytechnic of Wales Corpus (henceforth, "the corpus") has been chosen as a source for the grammar for the following three reasons:

[1] Use of informal spoken English.
[2] Use of a version of Systemic-Functional grammar developed by Robin Fawcett in both the corpus and the COMMUNAL project.
[3] A corpus which has not been previously exploited for computational linguistic purposes.

Because of point [3] above, and because the corpus is quite rare in being hand-parsed and transcribed from a spoken text, I will take time to describe its origins and format. The corpus was originally collected

for a child language development project to study the use of various syntactico-semantic constructs between the ages of six and twelve. A sample of approximately 120 children in this age range from the Pontypridd area was selected, and divided into four cohorts of 30, each within three months of the ages 6, 8, 10, and 12. These cohorts were subdivided by sex (B,G) and socio-economic class (A,B,C,D). The latter was achieved using details of

i) 'highest' occupation of either of the parents of the child,
ii) educational level of the parents.

The children were selected in order to minimise any Welsh or other second language influence. The above subdivision resulted in small homogeneous cells of three children. Recordings were made of a play session with a Lego building task for each cell, and of an individual interview with the same adult for each child, in which the child's favourite games or TV programmes were discussed.

2.2. Transcription

The first 10 minutes of each play session commencing at a point where normal peer group interaction began (the microphone was ignored) were transcribed by 15 trained transcribers. Likewise for the interviews. Intonation contours were added by a phonetician, and the resulting transcripts published in four volumes (Fawcett and Perkins, 1980). A short report on the project was also published (Fawcett, 1980).

2.3. Syntactic/semantic analysis

Again ten trained analysts were employed to manually parse the transcribed texts, using an informal Hallidayan systemic-functional grammar which handles phenomena such as raising, dummy subject clauses and ellipsis. Despite thorough checking, some inconsistencies remain in the text owing to several people working on different parts of the corpus.

2.4. Availability

The resulting parsed corpus consists of 11,396 (sometimes very long) lines, each containing a parse tree. The corpus of parse trees fills approximately 1 Mb and is available from the Polytechnic-of-Wales on magnetic tape in TAR or VMS Backup format. There are 184 files, each with a reference header which identifies the age, sex and social class of the child, and whether the text is from a play session or an in-

terview. The corpus is also available in wrap-round form with a maximum line length of 80 characters, where one parse tree may take up several lines.

2.5. *Systemic-Functional Grammar categories*

The grammatical theory on which the manual parsing is based is Robin Fawcett's development of a Hallidayan Systemic-Functional Grammar, described in Fawcett (1981). Elements of structure, such as sentence (Z), subject (S), complement (C), and adjunct (A) may be filled by formal categories: clauses (cl), groups (cf phrases in TG or GPSG) such as nominal group (ngp), prepositional group (pgp) and quantity-quality group (qqgp), or clusters such as genitive cluster (gc). The top-level symbol is Z (sigma) and is invariably filled by one or more clauses. Trees tend to be fairly flat immediately below the clause level, which can dominate for example, S OX M C A+ , and this has a direct effect on the size of the formal grammar which we have extracted. Some areas have a very elaborate description, eg: adjuncts, modifiers, determiners, auxiliaries, while others are relatively simple, eg: main-verb (M), and head (H). This fact has consequences for the easy use of the grammar in parsing programs: if the grammar is ultimately intended to cover a wide range of structures, then a correspondingly large lexicon will probably be required, such as the machine-readable version of LDOCE. However, the 'theory-neutral' syntactic classification of LDOCE may not necessarily be straightforwardly mapped onto the sometimes elaborate corpus syntax.

2.6. Notation

The tree notation employs numbers rather than the more traditional bracketed form to define tree structure, in order to capture discontinuous units. The number directly preceding a group of symbols refers to their mother. The mother is itself found immediately preceding the first occurrence of that number in the tree. In the example section of a corpus file given below, the first tree shows a sentence (Z) consisting of two daughter clauses (CL), as each clause is preceded by the number one. The long lines have been folded manually for ease of reading. The first number in each tree is a sentence reference.

Example 1: A sample section of a PoW Corpus file

**** 58 1 1 1 0 59

6ABICJ (filename)

1) [FS:Y...] Z 1 CL F YEAH 1 CL 2 S NGP 3 DD THAT 3 HP ONE 2 OM 'S 2 C NGP 4 DQ A 4 H RACING-CAR
2) Z CL 1 S NGP 2 DD THAT 2 HP ONE 1 OM 'S 1 C NGP 3 DQ A 3 MO QQGP AX LITTLE 3 H TRUCK
3) [HZ:WELL] Z 1 CL 2 S NGP HP I [RP:I] 2 AI JUST 2 HAD 2 C NGP 3 DQ A 3 MO QQGP AX LITTLE 3 H THINK 1 CL 4 & THEN 4 S NGP HP I 4 M THOUGHT 4 C CL 5 BM OF 5 M MAKING 5 C NGP 6 DD THIS 6 HP ONE
4) Z 1 CL 2 S NGP HP I 2 AI JUST 2 M FINISHED 2 C NGP 3 DD THAT 3 HP ONE 1 CL 4 & AND 4 S NGP HN FRANCIS 4 M HAD 4 C NGP 5 DD THE 5 H IDEA 5 Q CL 6 BM OF 6 M MAKING 6 C NGP 7 DQ A 7 RACING-CAR
5) [FS:THEN-I] Z CL 1 & THO 1 S NGP HP I 1 M MADE 1 C NGP DD THIS
6) Z CL 1 & THEN 1 S NGP HP FRANCIS 1 OX WAS 1 AI JUST 1 X GOING-TO 1 M MAKE 1 C NGP HP ONE 1 A CL 2 B WHEN 2 S NGP H YOU 2 M CAME 2 CM QQGP AX BACK 2 CM QQGP AX IN
7) [NV:MM] Z 1 CL F NO [FS:FRAN...] 1 CL 2 S NGP HP WE 2 M HAD 2 C NGP 3 DQ AN 3 H IDEA 3 Q CL 4 BM OF 4 M MAKING 4 C NGP 5 DQ FOUR 5 H THINGS
8) Z 1 CL F YEAH 1 CL 2 S NGP HP I 2 M PLAYED 2 C PGP 3 P WITH 3 CV NGP HP IT 2 A PGP 4 P AT 4 CV NGP H HOME
9) Z CL F YEAH

Occasionally when the correct analysis for a structure is uncertain, the one given is followed by a question mark. Likewise for cases where unclear recordings have made word identification difficult. Apart from the grammatical categories and the words themselves, the only other symbols in the tree are three types of bracketing:

a) square [NV...], [UN...], [RP...] for non-verbal, repetition, etc,
b) angle <...> for ellipsis in rapid speech,
c) round (...) for items recoverable from previous text.

3. Extracting grammar rules from the corpus

In order to create a formal grammar for use in parsing, simple context-free phrase structure rules have been extracted from the corpus. (The systemic-functional grammar used for the hand-parsing of the corpus was not computationally formalised.) The trees are first transformed into bracketed format using an algorithm which relates the numbers be-

tween categories in the original numerical form. We have yet to find a way of capturing discontinuities in bracketed trees and in simple phrase-structure rules so these are currently filtered out of the extraction process. So far we have used two alternative methods for rule extraction which differ regarding their treatment of filling and componence. Basically, filling is a relationship between two categories within a node, such as "S_ngp", whereas componence is a relationship between categories at different nodes in the tree, such as "ngp → dd h" (see Fawcett 1981: 6-7). The first maintains Fawcett's distinction between filling and componence relationships and results in just over 8,500 distinct rules, the second treats filling and componence as a single relationship, which still yields over 4,500 distinct rules. The second strategy is both more economical and intuitively more sensible, since the internal structure of a nominal group is the same whether it fills a subject or a complement, and filling must allow for branching when nominal groups are co-ordinated.

Of the total of 8,500 unique rules extracted from the whole corpus using the first method, some occur very frequently and some only once. Inevitably, some rules will have been extracted from malformed or inconsistent input, but hopefully these will be relatively rare, and allow a core grammar to be defined using only the fairly frequently occurring rules. The frequency threshold by which a rule is accepted or not will have to be carefully established, but initial examination of the rare rules is not encouraging. More than 6,000 rules occur only once out of the 8,500 total. Two-thirds of the 4,647 unique rules yielded by the second method of extraction are also singletons. Many of these singletons are not produced by errant input, but are genuine examples of rare structures in (spoken) English (Actually some would be considered to be not so rare, if we trusted our native-speaker intuition!). Other rare rules result from the elaborate category-labelling for a particular construct, or the flat shape of the tree immediately below the clause level. Once the genuinely false rules have been manually weeded out, a clearer picture will emerge. However, it already seems clear that the rule-frequency curve will be very similar to that identified with word-frequency, where only a few items occur very frequently, and very many occur only once or twice (Zipf 1936).

As a bench mark against which to evaluate these rulesets, a formal grammar based on the structures given in Fawcett 1981 was written using PS rules which allowed repeated and optional daughters. When this grammar was expanded into simple PS rules (with a limit of three

put on possible co-ordinations, which is by no means generous), over 18,000 distinct rules were produced. This confirmed our suspicions that there were actually gaps in the rules extracted from the corpus. The observation that a comprehensive grammar for English could be as openended as its vocabulary, might be disturbing from the grammarian's point of view, but all is not necessarily lost. It is to be hoped that some of the rarer constructs described by these rules are not likely to be used by a potential user of an English interface to an IKBS for a specific domain, where it is reasonable to suppose that a subset of the grammar will be adequate. Future work will hopefully establish the extent of such a subset, by a frequency threshold and other means.

4. Parsing strategies

Several quite different parsing strategies are currently under investigation, both deterministic and non-deterministic.

The grammar rules have been extracted from the corpus in Prolog form to be supplied to the parser-generator built into Prolog. Unfortunately system limitations only allow about 1,000 rules to be included before the memory is exhausted. The method used and problems encountered in this and a corresponding exercise for the LOB treebank are discussed more fully in Atwell and Souter (1988). A similar parser is built into POP11, which has yet to be tested using the large ruleset.

The Alvey NL Toolkit contains a deterministic chart parser developed by John Phillips and Henry Thompson at Edinburgh (see Phillips 1986, Phillips and Thompson 1987). The parser is available in Common and Franz LISP, and is based on a GPSG type grammatical theory, where syntactic categories are represented as sets of features. It seems likely that a more straightforward category-based technique without feature unification will be sufficient for the systemic-functional grammar (SFG) to be used in the early parsers for COMMUNAL. However, deterministic parsers of this ilk generally suffer from ungraceful failure. That is, if the grammar cannot describe a particular structure, the parser fails, with no attempt at even a partial analysis.

An alternative strategy is to use non-deterministic probabilistic parsing, which has no absolute distinction between grammatical and ungrammatical structures, but instead uses probabilities extracted from a manually-parsed corpus to evaluate possible trees for a given string of words. In this framework, 'grammaticality', or the lack of it, is treated as a sliding scale between high and low frequencies. The technique of simulated annealing, which has been applied in APRIL, a related pro-

ject being conducted by Geoff Sampson, Eric Atwell and Robin Haigh at Leeds, involves random moves in the search space of all possible trees for a string of words, with evaluation of each move to decide whether it should be accepted (see eg Sampson 1986, 1987a, 1987b). Using a variant of simulated annealing which uses educated guesses for possible moves on the basis of the frequencies extracted form the Polytechnic of Wales corpus would at least permit a best SFG analysis for all possible input sentences, including those which might be considered semi-grammatical.

Note

1. COnvivial Man-Machine Understanding through NAtural Language, funded by RSRE, ICL and Longman.

References

Atwell, Eric S. and D. Clive Souter. 1988. Experiments with a very large corpus-based grammar. To appear in *Proceedings of the 15th International Conference on Literary and Linguistic Computing*, (ALLC '88). Jerusalem, June 5-9, 1988.

Fawcett, Robin P. 1980. Language development in children 6-12: Interim report. *Linguistics* 18: 953-958.

Fawcett, Robin P. and Michael R. Perkins. 1980. *Child language transcripts 6-12*. With a preface, in 4 volumes. Department of Behavioural and Communication Studies, Polytechnic of Wales.

Fawcett, Robin P. 1981. *Some proposals for systemic syntax*. Department of Behavioural and Communication Studies, Polytechnic of Wales.

Garside, Roger G., G.N. Leech and G.R. Sampson (eds). 1987. *The computational analysis of English*. London: Longman.

Phillips, John D. 1986. A simple, efficient parser for phrase-structure grammars. *AISB Quarterly* 59: 14-18.

Phillips, John D. and Henry S. Thompson. 1987. A parsing tool for the natural language theme. Software Paper no 5, Department of Artificial Intelligence, Edinburgh University.

Sampson, Geoffrey R. 1986. A stochastic approach to parsing. In *Proceedings of the 11th International Conference on Computational Linguistics* (COLING '86). 151-155.

Sampson, Geoffrey R. 1987a. The grammatical database and parsing scheme. In Garside et al (1987). 82-96.

Sampson, Geoffrey R. 1987b. Probabilistic models of analysis. In Garside et al (1987). 16-29.
Zipf, George K. 1936. *The psycho-biology of language: An introduction to dynamic philology.* London: George Routledge.

A quantitative look at the Guangzhou Petroleum English Corpus

Zhu Qi-bo
Guangzhou Training College of the Chinese Petroleum University

1. Introduction

In 1987, the Guangzhou Petroleum English Corpus (shortened as GPEC) was set up in Guangzhou by the present author under the guidance of Prof. Gui Shi-chun. The purpose of building the corpus is threefold. First, to get to know more about the features of Petroleum English (English used in Petroleum Industry, shortened as PE). Second, to provide teachers and learners of Petroleum English with a series of vocabulary lists and some other firsthand information. Finally, to gain some empirical knowledge in developing a model for processing a medium-sized corpus on a microcomputer. A total of 700 texts about 500-600 words long were drawn randomly from Petroleum English writings. The materials were then processed on an IBM PC/XT in the Guangzhou Foreign Language Training Center of the Petroleum Ministry with the usual software environment of DOS 2.0, BASIC, and dBASE III. Among other things, the computational analysis showed that in GPEC (411,612 running words) there are 24,506 different word types, and 11,259 word types are *hapax legomena.*

The present study is an attempt to examine some of the computational results of GPEC in a quantitative way. In the study, data sets from GPEC are compared with those from other corpora, to test the validity of the sampling techniques of GPEC and to find out some similarities and dissimilarities between Petroleum English and General English (ordinary everyday English, shortened as GE). At the same time, some tentative explanations are made to account for the common features and differences.[1]

2. A quantitative look at the Guangzhou Petroleum English Corpus

2.1. Frequency distribution

Tables 2.1-A and 2.1-B present a comparison of the frequency distribution of GPEC and the BROWN Corpus. The cross-corpus view reveals

Table 2.1-A
Frequency distribution table (GPEC)

Rank	Cum.Pro[2]	Rank	Cum.Pro
5	19.450	1997	80.341
8	24.206	3017	85.343
50	39.529	3173	85.912
100	45.078	5002	90.550
160	49.480	7988	94.262
502	62.136		
797	68.337		
1004	71.234		
1568	77.200		

Table 2.1-B
Frequency distribution table (the BROWN Corpus; cf. Kučera and Francis 1967: 300-310)

Rank	Cum.Pro	Rank	Cum.Pro
5	18.203	2006	76.256
8	22.347	3005	80.663
50	40.658	3244	81.487
100	47.430	4985	86.204
160	51.663	7920	90.516
500	61.927		
800	66.590		
1000	68.860		
1600	73.911		

a common feature: it takes a small number of high-frequency words to cover a large proportion of the texts and a larger number of words to get an extra percentage point. Generally speaking, there is close similarity between the distribution of GPEC and that of BROWN. This is confirmed by the test for the goodness of fit between theoretical percentage and observed percentage (chi-square value: 0.5174, df=13, $p<0.001$). For the distribution of the first 500 words, the two corpora are more or less the same with the percentage deviating only by about 0.029. As the figures go further down, the divergence of subject matter coverage in the two corpora becomes greater. The discrepancy may be attributed to the difference of subject matter coverage in the two corpora. While the BROWN Corpus is a large-scale corpus covering a wide range of subject matter, GPEC is a subject-specific corpus, cover-

ing only five sub-fields of one subject. On the one hand, some common words will invariably occur in the language for all subjects; on the other hand, it takes many more words to discuss a variety of topics. Alford (1971:82) stated: 'If there is a spread of subject-matter, the text coverage declines for greater numbers of different words. Coverage can be up to 10% lower for the higher figures'. The pattern of frequency distribution change as shown in Tables 2.1-A and 2.1-B testifies to this statement.

2.2. The 100 most over-represented and 100 most under-represented words in GPEC

Words with F (frequency) ≥ 10 and r (range) ≥ 5 in both BROWN and LOB[3] were chosen to make up a list called the LOB-BROWN LIST. Words with F ≥ 4 and r ≥ 3 in GPEC were selected to form a GPEC LIST. When these two lists are compared, a rank list of the distinctiveness coefficient[4] values is obtained (for example, *oil* is the first one in the table with the DIS.C. of -0.9650; it is the most over-represented word; the most under-represented word is *my* with a DIS.C. of 0.9754). When the grammatical classes of the 100 most over-represented words and the 100 most under-represented words are compared, it is found that the grammatical classes of words vary with distinctiveness coefficients. As the distinctiveness coefficients draw close to -1, we find more nouns; as the distinctiveness coefficients go towards 1 there are more verbs. Table 2.2-A shows a comparison of the noun-verb ratio in the two lists. The contrasts between them are self-explanatory: GPEC reflects a more nominal character and LOB-BROWN a more verbal character. If 'a complex, heavy style in English is usually the result of a failure to use meaningful verbs' (Fries 1940:62), the above comparison will serve to prove the complex and formal style of PE.

Table 2.2-A
The noun-verb ratio in the GPEC and LOB-BROWN lists

Word class	GPEC	LOB-BROWN
noun	78	31
verb	9	32
adjective	11	11
others	2	26
noun-verb ratio	1:0.12	1:1.03

2.3. The 50 most frequent words in GPEC

Table 2.3-A lists the 50 most frequent word types in GPEC, as compared with the same word types in three other corpora: JDEST,[5] LOB, and BROWN. Where the rank in these sources is lower than 50, this is marked by "50". The first thing to be noticed is that 35 out of the 50 words rank among the top 50 in all of the 4 corpora. A cursory look at the 35 words reveals that the bulk of these words are function words, which seem to be the mainstay of writings of all fields. There are also some words that are unique to GPEC in Table 2.3-A. They are:

oil, gas, water, pressure, well, should

Of these 6 words 5 are closely related to the subject matter. These words are, in effect, well in line with the author's desire for topicality. Since GPEC is a special-purpose corpus concentrating on Petroleum English, it is no wonder that the most frequent words include *oil*, *gas*, *water*, *pressure* and *well*. These words indicate that the sampling principle of homogeneity has been adhered to.

Table 2.3-A
The ranks of the 50 most frequent words in GPEC and their ranks in three other corpora[6]

Word	GPEC	JDEST	LOB	BROWN
the	1	1	1	1
of	2	2	2	2
and	3	3	3	3
to	4	4	4	4
in	5	5	6	6
a	6	6	5	5
is	7	7	8	8
be	8	8	15	17
for	9	9	11	11
are	10	10	27	24
by	11	12	20	19
with	12	14	14	13
as	13	13	13	14
that	14	11	7	7
or	15	19	31	27
this	16	15	22	21
on	17	16	16	16
from	18	22	25	26

at	19	18	19	18
it	20	17	10	12
oil	21	50	50	50
which	22	20	28	31
was	23	35	9	9
an	24	21	34	29
gas	25	50	50	50
can	26	23	50	50
not	27	24	23	23
have	28	25	26	28
water	29	50	50	50
pressure	30	50	50	50
has	31	26	42	44
used	32	30	50	50
may	33	36	50	50
will	34	29	48	47
these	35	27	50	50
been	36	31	37	43
well	37	50	50	50
than	38	37	50	50
more	39	33	50	48
if	40	40	45	50
other	41	38	50	50
when	42	41	44	45
were	43	50	35	34
should	44	50	50	50
one	45	28	38	32
all	46	42	39	36
but	47	32	24	25
such	48	34	50	50
into	49	47	50	50
also	50	39	50	50

Here, a useful coefficient, Spearman's rank correlation coefficient (Gui Shi-chun 1986:145), is calculated to measure the similarity or dissimilarity of GPEC ranks to those of the other three corpora. We use the rank order of GPEC (GR) to subtract the corresponding rank order of the same word in another corpus (JR, for example), and obtain the difference (DGJ) and square of the difference (DDGJ). The calculation is based on the formula (Gui Shi-chun 1986:145):

$$rho = \frac{6 \cdot sum\ ((X-Y) \cdot (X-Y))}{N \cdot (N \cdot N-1)}$$

In our case, N=50, X=GR, Y=JR (or LR,BR). With the help of some dBASE commands, a calculation table is worked out. The following results are obtained showing the relationship of GPEC to the other three corpora:

rho for GPEC and JDEST = .81 (relationship: closely related)
rho for GPEC and LOB = .68 (relationship: related to a certain degree)
rho for GPEC and BROWN = .68 (relationship: related to a certain degree)

(The interpretation of the relationship here is based on the correlation coefficient table in Gui Shi-chun 1986: 143).

According to the above analysis, closer correspondence in rank is found between JDEST and GPEC than between LOB, BROWN and GPEC. This is because the most frequent words reflect the stylistic character of the texts. Despite discrepancies in subject matter, GPEC and JDEST belong to the same major text category – scientific English texts.

2.4. Numerals

Table 2.4-A is a list of numerals (word types) *one* through *nine* and Arabic symbols 1 through 9 with frequencies from GPEC and two other corpora. It is notable that there is an inverse relationship between the numerals and frequency (only with a slight deviation at 8 in one corpus). This indicates more or less the same fashion of using numerals in different types of texts. Subject matter and the size of the corpus appear to exert little influence on rank order. Benford (as quoted in Carroll et al. 1971:xliv) states that in some lists of apparently random numbers lower first digits occur more frequently than higher. If his law holds true, the general agreement of Benford's statement and the GPEC corpus data may suggest that the present author has succeeded in achieving the principle of "randomness" in choosing his texts for the corpus.

Table 2.4-A
The frequencies of numerals *1/one* to *10/ten*[7]

Numeral	LF	BF	GF	DIST.C	X	G
1	785	358	393	-0.6722	1065.118	a
2	583	287	314	-0.6512	752.533	a
3	420	246	232	-0.6218	503.913	a
4	307	172	157	-0.6399	385.461	a
5	250	146	129	-0.6321	307.886	a
6	189	122	106	-0.6036	216.314	a
7	119	83	89	-0.5430	114.955	a
8	129	78	100	-0.5591	130.638	a
9	91	66	62	-0.5525	90.403	a
one	3088	3297	741	0.277	357.089	a
two	1549	1412	548	0.051	4.792	c
three	697	610	248	0.038	1.043	
four	369	359	102	0.188	19.461	a
five	261	286	54	0.350	46.881	a
six	231	220	43	0.365	41.711	a
seven	118	114	21	0.387	23.845	a
eight	92	104	17	0.405	21.827	a
nine	79	81	10	0.533	28.550	a

Another interesting finding is the consistent over-representation in GPEC of Arabic symbols, while numerals spelled with letters are over-represented in LOB and BROWN. This is due to a large extent to the larger number of figures, graphs and tables in GPEC.

2.5. Personal pronouns

Table 2.5-A shows that personal pronouns are consistently under-represented in GPEC as compared with LOB and BROWN. PE writers are usually scientists and engineers who are interested in things and processes, in properties and characteristics. So they usually prefer to use impersonal expressions. The same finding was made at the University of Nottingham when a group of researchers investigated the construction of German chemical texts. One of the results of their investigation was the daring exclusion of some pronouns from their basic ESP course (Hockey 1980:79).

Table 2.5-A
The frequencies of some personal pronouns[7]

Word	LF	BF	GF	DIS.C	X	G
I	6696	5156	154	0.881	4782.280	a
we	2926	2653	426	0.457	766.544	a
you	3590	3286	88	0.882	2782.732	a
he	8885	9667	172	0.913	7920.376	a
she	3912	2859	0	1.000	3320.799	!
they	3579	3619	520	0.479	1071.049	a
us	657	672	64	0.619	305.176	a
me	1554	1181	0	1.000	1341.365	!
him	2258	2619	27	0.947	2202.449	a
her	4030	3037	0	1.000	3465.971	!
them	1699	1789	131	0.690	955.915	a
my	1813	1319	8	0.975	1478.846	a
your	853	923	40	0.802	618.758	a
their	2808	2670	371	0.503	887.776	a

(where GF < 4, "0" is used)

2.6. Particles

According to Table 2.6-A some particles are over-represented in GPEC while others are under-represented. Many of the words in the latter group frequently form part of phrasal verbs. We may assume that PE prefers more formal and more precise verbal expressions, e.g. *extract* for *draw out*.

Table 2.6-A
The frequencies of some particles[7]

Word	LF	BF	GF	DIS.C	X	G
along	250	355	160	-0.127	12.913	a
throughout	122	141	66	-0.101	3.580	
within	344	359	246	-0.261	69.904	a
between	867	730	560	-0.262	160.147	a
during	497	585	428	-0.317	169.673	a
above	296	296	251	-0.348	116.428	a
via	20	48	39	-0.474	30.156	a
below	150	145	219	-0.567	227.038	a
near	207	198	132	-0.228	29.618	a

inside	130	174	72	-0.072	2.199	
with	7197	7290	2931	0.006	0.154	
at	6043	5377	2168	0.038	8.962	b
to	26760	26149	9554	0.063	142.469	a
on	7027	6742	2320	0.098	96.927	a
after	1119	1070	359	0.111	20.206	a
over	1264	1237	355	0.182	62.560	a
across	264	282	73	0.210	18.168	a
about	1895	1815	435	0.272	200.929	a
up	1860	1896	358	0.365	347.606	a
down	885	895	162	0.385	181.014	a
out	2035	2096	354	0.410	470.147	a
off	551	640	155	0.223	44.397	a
back	934	967	113	0.550	358.454	a

3. Conclusions

The quantitative analysis above gives evidence of the validity of the sampling principles used in compiling the Petroleum English Corpus. A comparison with other corpora has also revealed some characteristic features of this type of text. Petroleum English is more formal and more impersonal than General English. It includes subject-specific terms, more Arabic numerals, and more precise one-word equivalents of phrasal verbs. Nevertheless, we should not isolate Petroleum English from General English, because the basic words and patterns of General English are also an essential part of Petroleum English. It is the author's hope that the corpus study will make a contribution towards the teaching of Petroleum English in such areas as the planning of the curriculum, the compiling of textbooks, and the writing of exercises and tests.

Notes

1. My sincere thanks go to Professor Gui Shi-chun, President of Guangzhou Foreign Language Institute, for his comments and suggestions and to Professor Wang Zongyan of Zhongshan University, Professor Wu Qianguang, and Professor Li Xiaoju of Guangzhou Foreign Language Institute, who jointly verified the data sets of GPEC used here.
2. CUM.PRO = cumulative proportion

3. LOB: refers to the Lancaster-Oslo/Bergen Corpus, a British equivalent of the Brown Corpus; see Hofland and Johansson (1982).
4. Distinctiveness coefficient (DIS.C.): a measure to estimate the over/under-representation of words. It is calculated by the following formula (cf. Hofland and Johansson 1982:14):

$$\text{DISC.C.} = \frac{x1-x2}{x1+x2}$$

where x1= the adjusted frequency in the combined word list of LOB-BROWN LIST, x2= the adjusted frequency in GPEC. The calculations involved in the adjustment of frequencies are as follows (take the word *I*, for example):

x1 = [6696 (i.e. F in LOB) + 5156 (i.e. F in BROWN)]/2 = 5926
x2 = 154 (i.e. F in GPEC) • 2.44 = 375.76

Thus the DIS.C. is:

$$\text{DIS.C.} = \frac{x1-x2}{x1+x2} = \frac{5926-375.76}{5926+375.76} = 0.881$$

Usually DIS.C. takes the value of -1 to 1. A positive value like 0.881 for the word *I* indicates under-representation in GPEC and over-representation in the BROWN-LOB LIST. A value of "0" means that the frequencies in BROWN-LOB LIST and in GPEC LIST are the same. A value of "1" or "-1" shows that the frequency in one of the two groups is too low (marked "0") to be compared. Since the value "0.881" is quite near "1", we can say that, in LOB and BROWN, we can find many more occurrences of the word *I* than we can in GPEC.
5. JDEST: Jiaoda Corpus of English for Science and Technology. It is a million-word corpus designed to meet the needs in the study of EST. See Yang (1985).
6. Ranks for LOB and BROWN are sorted according to frequencies of the words as presented in Hofland and Johansson (1982:471-544).
7. LF = frequency of the word in LOB; BF = frequency of the word in BROWN; GF = frequency of the word in GPEC. For LF and BF, see Hofland and Johansson (1982: 471-544).
   X: refers to chi-square-value. G: refers to significance level (a: 0.01, b: 0.05, c: 0.001, !: the theoretical frequency is too low in one of the groups; a space means the difference is not significant).

References

Alford, M. H. T. 1971. Computer assistance in language learning and in authorship identification. In *The computer in literary and linguistic research* (ed. R.A. Wisbey). Cambridge: University Press. 77-86.

Carroll, J. B., Peter Davies, and Barry Richman. 1971. *The American Heritage word frequency book.* New York: American Heritage Publishing Co.

Fries, Charles C. 1940. *English word lists.* Washington.

Gui Shi-chun. 1986. *Standardized test: Its theory, principles and methodology.* Guangzhou: Guangdong Publishing House of Higher Education.

Hockey, Susan. 1980. *A guide to computer applications in the humanities.* London: Duckworth.

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

Kučera, Henry. 1969 Computers in language analysis and in lexicography. *The American Heritage dictionary of the English language* (ed. William Morris). New York: American Heritage and Houghton Mifflin.

Kučera, Henry and W. Nelson Francis. 1967. *Computational analysis of present-day American English.* Providence, R.I.: Brown University Press.

Yang, Huizhong. 1985. The JDEST Computer Corpus of Texts in English for Science and Technology. *ICAME News* 9: 24-25.

ICAME 9TH

The 9th International Conference on English Language Research on Computerized Corpora, Birmingham University, 18th - 22nd May, 1988.

The conference was organised by members of the research team responsible for the innovative, corpus-based *Collins COBUILD English Language Dictionary* (1987). It was attended by more than sixty participants from eleven countries (Belgium, Canada, England, the Federal Republic of Germany, Finland, Holland, Japan, Norway, Sweden, Switzerland, the United States). Some thirty papers were read. As at previous conferences, there were papers reflecting the various stages of computer corpus work, from the compilation of corpora and the development of analytical tools to the use of machine-readable texts for studies of particular aspects of the English language. Abstracts of most of the papers are given below. Fuller versions of three of the papers (by Kytö, Rissanen and Souter) are included in the article section of this issue. Selected papers will be published in a volume edited by Jan Aarts and Willem Meijs (Amsterdam: Rodopi).

Apart from the papers, the programme included a text-processing software workshop and a visit to the COBUILD Project. On the social side, there was a visit to Stratford-on-Avon, including a performance of "Much Ado About Nothing", and a conference supper with Professor David Lodge as a special invited guest reading from his new novel. The participants are indebted to the organisers (Antoinette Renouf, John Sinclair, Jeremy Clear) for a successful and well-organised conference. The next ICAME conference will take place in Bergen in June 1989.

List of papers

Birmingham University:

Antoinette Renouf, Work in progress at the English Language Research and Development Unit – see abstract

John Sinclair and Jeremy Clear, Steps in automatic lexicography – see Renouf's abstract

Brigham Young University:
Charles D. Bush, An inventory of search capabilities for language corpora – see abstract
Randall Jones, The creation of a spoken American English corpus using TV interviews

Helsinki group:
Ossi Ihalainen, Dialectal syntax: In search of data – see abstract
Merja Kytö, Progress report on the diachronic part of the Helsinki Corpus – see article
Matti Rissanen, Three problems connected with the use of diachronic computer corpora – see article

Lancaster group:
Andrew Beale, Retrieving collocations and constituents from tagged corpora – see abstract
Nick Campbell, Measuring speech-rate in the Lancaster Spoken English Corpus – see abstract
Gerry Knowles, Prosodic transcription by rule
Anne Wichman, Stylistic variation in intonation

Leeds group:
Eric Atwell, CCALAS: A new centre for ICAME-related research
Geoffrey Sampson, Optimisation parsing
Clive Souter, The COMMUNAL Project: Extracting a grammar from the Polytechnic of Wales Corpus – see article

Lund group:
Karin Aijmer, Report from ongoing work on conversational phrases in English – see abstract
Bengt Altenberg, Phraseology in the London-Lund Corpus – see abstract
Mats Eeg-Olofsson, The computer processing of collocations in the London-Lund Corpus – see abstract
Anna-Brita Stenström, What is the role of discourse signals in sentence grammar? – see abstract

Nijmegen group:
Theo van den Heuvel, Form and interpretation in automatic morphological analysis – see abstract
Pieter de Haan, Structure frequency counts of modern English: Progress report – see abstract

Nijmegen group (CELEX):
Hans Kerkman, Applying lexical databases in research and technology

Françoise Keulen, Building a lexical database of English – see abstract

Survey of English Usage:

Ewa Jaworska, A grammatical database for the Survey of English Usage – see abstract

Geoffrey Kaye, A grammatical database and query system for the Survey of English Usage – see abstract

Other papers:

Nancy Belmore (Concordia University), Working with Brown and LOB on a microcomputer – see abstract

Edward Finegan (for Edward Finegan and Douglas Biber, University of Southern California), Problems in the automatic grammatical analysis of seventeenth-century English texts – see abstract

Göran Kjellmer (Gothenburg), Patterns of collocability – see abstract

Henry Kučera (Brown University), Text research and spelling checkers/grammar checkers

Willem Meijs (Amsterdam), Progress report on the Amsterdam projects – see abstract

Jacques Noël (for A. Moulin and Jacques Noël, Liège), CD-ROM corpora – see abstract

Dieter Mindt (Berlin), Prepositions in LOB and BROWN – see abstract

Joseph Schmied (Bayreuth), Compiling a corpus of East African English – see abstract

Kay Wikberg (Oslo), Using WordCruncher as a means of studying lexical patterning and thematic progression – see abstract

Abstracts

Report from ongoing work on conversational phrases in English

Karin Aijmer
Lund University

Conversational phrases can be grouped according to the specific functions they are used to perform. One group consists of social formulae. These are used when we perform speech acts of a social or polite nature such as greeting, thanking, apologizing, requesting, offering.

Social formulae are of different kinds. In the request 'can you open the window please', *please* shows explicitly that the utterance is intended as a request. What is of interest is that also the introductory *can you* identifies the utterance as a request. *Can you* is however different from other phrases because it cannot stand alone as a request. Instead it is a part of a syntactic formula which includes an unspecified verb phrase.

A formula may be more or less fixed in form and occur with a high or low degree of patterning. The degree of patterning is one way in which speech acts can differ from each other as can be illustrated with examples from thanking and requesting.

Acts of thanking were typically expressed by formulae and showed considerable patterning in the large discourse. As to requests their formulaic character was often more hidden, and it was difficult to know if a particular utterance (or part of an utterance) should be regarded as formulaic.

In all 300 examples of gratitude phrases were investigated in the London-Lund corpus of spoken English. Only a small number of patterns was represented in the material (formed on the basis of *thank you* and *thanks*).

It is typical of formulae that they have to be performed on particular social occasions and at specific points in the larger conversational interaction. In order to analyse thanks, the situations were classified according to the antecedent event causing gratitude and according to situation. Acts of thanking were for instance analysed as the second part of compliments, well-wishes, blessings, congratulations, greetings.

Thanking was especially frequent after offers (48 exx). In 33% of the examples *thank you* was used indirectly at the end of the larger conversation as a closing signal either alone or as a part of a larger formula (*thank you for that most interesting contribution*).

Requests turned out to be more difficult than acts of gratitude to describe systematically. One reason was that they were often ambiguous between a more literal interpretation and the interpretation as a request for something. Thus out of 647 examples of requests 33% were expressed indirectly. Both questions and declarative sentences could be analysed as request patterns. 93% of the questions contained a formula with a modal auxiliary (most frequently *could* or *would*).

Phraseology in spoken English

Bengt Altenberg
Lund University

In July 1987 a new project called 'Phraseology in Spoken English' was started at the Survey of Spoken English, Lund University. The aim of this project is to make a detailed investigation of recurrent word combinations in the London-Lund Corpus with special emphasis on collocations and prefabricated expressions that reflect the speech process and the organization of discourse. The work will be carried out in three main phases:

Phase 1: Collection of data

- retrieval of recurrent word combinations in the corpus;
- elimination of irrelevant (ie. phraseologically uninteresting) combinations on the basis of structural criteria;
- production of concordances and frequency lists;

Phase 2: Lexico-grammatical analysis

- grammatical classification of the resulting combinations in terms of word class, phrase category and clause function with the aid of automatic tagging and parsing programs;
- estimation of the collocability and idiomaticity of the combinations;

Phase 3: Prosodic and functional analysis

- analysis of the positional and prosodic characteristics of the combinations, their discourse functions, and their distribution in different speech types.

As this plan indicates, the aim is to produce a rather comprehensive description of the phraseology of spoken English (statistical, lexical, grammatical, prosodic and functional). A corpus-based description of this kind is of obvious interest in a number of areas such as lexicography, language teaching, discourse analysis, computational linguistics, text-to-speech conversion, etc. In this respect the project will serve as an important supplement to similar investigations based wholly or largely on written material, such as the Brown Corpus (Kjellmer), the LOB Corpus (Johansson and Hofland, Beale) and the Birmingham Corpus (Sinclair et al).

However, since a major part of the London-Lund Corpus consists of spontaneous speech – a variety which more than any other reflects language as a mental process – the results are also likely to be of considerable psycholinguistic and theoretical interest. By focusing on an important borderline area between lexis and grammar the project will provide valuable information about the speech process and the nature of linguistic competence (as distinct from communicative competence). Hence, a secondary aim of the project will be to explore questions like the following: what expressions are retrieved as prefabricated chunks straight from memory (the mental lexicon) and what expressions are generated by productive grammatical rules? How clear-cut is the division between lexis and grammar? Does the evidence suggest that the lexicon has a larger part to play in linguistic competence than has previously been assumed, or does there seem to be overlap (and hence redundancy) between the lexical and the grammatical component? Should we postulate a 'janus-like' phraseological component at the intersection of the lexicon and the generative rules of grammar?

The computational aspects of the project are described in Mats Eeg-Olofsson's abstract (see below).

Retrieving collocations and constituents from tagged corpora

Andrew Beale
University of Lancaster

Two areas of research are discussed in this paper:

1) the Distributional Lexicon and
2) a system for retrieving constituents from parsed text samples.

Early work on the Distributional Lexicon (DL) is documented in Garside, Leech and Sampson (1987). The word tagged version of the LOB Corpus (Johansson and Hofland, 1987) was used as the input text to the DL programs. In broad outline, the DL programs may be thought of as a combination of two suites: one suite produces hierarchised lists showing the frequencies and distribution of alphabetically sorted grammatical words, and their right or left neighbours; the other suite carries out lemmatisation.

The lemmatisation routines arrange inflexional or morpho-syntactic variants against an alphabetically sorted list of lexemes or headwords. The first procedure in this suite finds the lexeme and general grammatical tag associated with each input grammatical word in the input running text by scanning through ordered compound lists of tags and words or tags and suffixes.

The lemmatisation procedures were integrated with the procedures for producing the hierarchised lexicon and the whole of the LOB Corpus was processed in six batches. The complete DL, including over 80 per cent hapax legomena is on five tapes occupying about 92 Mbytes.

Filtering programs can be used to select subsets of data from the DL. The basic filtering program copies all records of items with a frequency above a specified threshold. If this threshold is set to one, the data set is reduced to less than 18 per cent of the original DL. More sophisticated filtering programs can output combinations of grammatical words with a combined frequency above (or below) a threshold proportionate to the frequency of the keyword member. This provides an automatic measure of the 'strength of bonding' between two neighbouring grammatical words but it does not show collocability of discontinuous words.

A manually parsed sample of about 60,000 words of American English press reports was compiled as part of the IBM/UCREL project in language modelling for speech recognition. The parsed text samples are in the form of strings of labelled bracketing and grammatically tagged words. Another parsed corpus of texts is currently being compiled using a skeletal system of context-free rules to guide the manual parsers.

Students type in parsed structures interactively at the terminal, via a constituent inputting program. The program checks that labels are legal constituent tags, and that labelled brackets balance. It returns appropriate error messages and allows re-editing of any ill-formed structures after the structure for each sentence has been keyed in.

The constituent retrieval program is, in fact, a modified version of the inputting program. Instead of inputting structures, the retrieval program extracts occurrences of specified constituents from the treebank, sorts them and sums together tokens of the same type. The constituent retrieval program is used as a tool for checking the treebank and for development of the associated grammar.

References

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

Garside, R. G., G. N. Leech and G. R. Sampson. 1987. *The computational analysis of English: A corpus-based approach*. London and New York: Longman.

Johansson, S. and K. Hofland. 1987. 'The tagged LOB Corpus: Description and analyses', in *Corpus linguistics and beyond*, edited by Willem Meijs. Amsterdam: Rodopi. 1-20.

Working with Brown and LOB on a microcomputer

Nancy Belmore
Concordia University

Some years ago Gert van der Steen observed that a search through a 600,000 word Dutch corpus for all instances of a single word, using the fastest available system search program on a Cyber 173 mainframe,

took 30 times more turn-around time, during evening hours, than the same search on a Data General Eclipse minicomputer. By now, working with a microcomputer at your own desktop is an even more attractive alternative to a mainframe than a minicomputer, with numerous advantages apart from turn-around time. This paper describes the use of an Apple microcomputer, a Macintosh II, to set up a system for working with the Brown and LOB one-million word corpora, both tagged and untagged. The Macintosh family of microcomputers has numerous advantages which other microcomputer manufacturers have begun to incorporate in their own machines. Some of the most important derive from the superior Macintosh graphics.

As test data, one file from the Brown tagged corpus and one file from the LOB tagged corpus were down-loaded from a Vax computer. It took about 14 minutes to download the 106K Brown file and about 25 minutes to download the 156K LOB file using 9600 baud modem and communications software called Red Ryder. The data from the Brown file were then exported to a word processor called MindWrite which can import and export data in standard text file format. It is a combined outliner and word processor with sophisticated search, modification and sorting capabilities.

Two new files were generated from the MindWrite text file, a word file and a sentence file. The word file was generated by re-formatting and editing the text file. The sentence file was generated by re-formatting the re-formatted and edited word file. The two files were then ready for export to a Fourth Dimension (4D) database. 4D is sophisticated relational database management software with extensive export-import facilities for data exchange with micros, minis and mainframes, its own PASCAL-like programming language and the ability to execute procedures written in any computer language which will compile on any of the Motorola 68,000 family of chips.

To export the MindWrite word and sentence files, they were first saved as text files. Two 4D input files were then created, one to receive the MindWrite word file; the other, to receive the MindWrite sentence file. To do this, it was only necessary to respond to 4D's request for a description of the fields in each record and then to designate one or more display formats (what 4D calls a layout) for the data in each file. This can be a default layout or a layout you design yourself. Importing a file consists of a few simple steps, all of which are a simple matter of pointing and clicking.

Sorting in 4D is also simple. Words within the word file were first sorted by word and within word, by tag and then by tag and within tag, by word.

4D provides two types or searches: by criteria or by formula. In a search by criteria, the user is presented with a dialog box and can specify searches of virtually any degree of complexity or length. If the user is satisfied with the search criteria, they can be saved for later use with other data. Several searches by criteria were executed. One was suggested by the first sort, which showed that although *as* had only occurred four times, it had been assigned three different tags. All sentences in which the word *as* had been tagged *cs* (subordinating conjunction), *in* (preposition) or *ql* (qualifier) were extracted. In another search, all instances in the sample in which a word ending in *-ed* and classified VBN (past participle) were located. In a search by formula, the user writes his own formula rather than making selections from a dialog box and he can then use almost all the commands of the 4D programming language.

Most of us have had experience with screen and printed displays which are hard to read and thus make progress slow. Both 4D and MindWrite simplify generating versions of a data set (or a part of one) which resolves this problem. For temporary display purposes, e.g., MindWrite was used to change abbreviated to unabbreviated forms, to automatically number the items displayed, and to highlight the key items found in a search. Without the highlighting features of MindWrite and 4D, it would be frustrating to use the power of the computer to perform complicated searches because the results would be hard to locate on the screen or printed page.

Current plans are to download more files from the two tagged corpora, following the procedures for importing, editing, re-formatting and export to 4D described in this paper. Before, or in some cases, after export to 4D, some additional editing is desirable, including data compression.

A valuable addition to the corpora available for processing on a microcomputer would be the Brown Corpus tagged with the LOB tagging suite. This would permit an exact determination of the precise differences between the Brown and LOB tagging systems and would be a useful tool in refining word classification systems for specific or even general purposes. Work on the required editing of the untagged Brown Corpus so that it will meet the expectations of the LOB tagging programs is underway.

While the system described here is by no means perfect, like the work of other colleagues, it does show a number of the advantages of working with large corpora on a microcomputer. The present configuration includes a Mac II with 1 megabyte of RAM and a 40 Mbyte internal hard disk. For a fully-powered research tool, it will be important to upgrade to the PMMU Motorola chip, to expand to a full eight megabyte of RAM and to have removable mass storage.

For their pioneering and continuing work on large corpora, we are all indebted to Francis, Kučera, Leech, Johansson and their colleagues. With the revolution in computers which has occurred within the past 10 years and still continues, the invaluable research tools they pioneered in providing are available to us all at our own desktops. It is an exciting future.

Problems in the automatic grammatical analysis of seventeenth-century English texts

Douglas Biber and Edward Finegan
University of Southern California

This paper explores the application of the tagging programs developed for twentieth century English texts to texts written in earlier centuries going back to the seventeenth century. Programs developed for automated grammatical tagging of the Brown, LOB, and London-Lund corpora were applied to prose texts of various genres written in the seventeenth century.

Aside from the obvious lexical lacunae not tagged because they are absent from the modern lexicon, other kinds of failure were identified. These comprised matters of spelling, morphology, and syntax. Many of these lexical lacunae are overcome by the tagging procedures' ability to tag from grammatical structures.

The paper thus concludes that most of the lacunae can be accommodated by relatively few and relatively minor adaptations in the tagging routines accompanied by some modification of the lexicon. (This conclusion is illustrated through discussion of several specific examples.)

An inventory of search capabilities for language corpora

Charles D. Bush
Brigham Young University

In the spring of 1988, the HUMANIST electronic mail discussion group saw a flurry of comments on what one participant called the "desiderata of retrieval software – the features we'd like to see in a text retrieval program." One underlying purpose of the discussion was to establish criteria for evaluating text retrieval programs (TRPs) and, by extension, to suggest enhancements to existing TRPs. Several HUMANISTs contributed their ideas until John J. Hughes summarized the discussion with a "Blue Sky Wish List" of 72 items. The participants in that discussion were all literature scholars and so, understandably, the wish list reflected the desired capabilities for literary studies.

A language corpus is not the same as a literary textbase. Briefly, in a literary text the focus is on the author, with the text as that author's artistic creation. The text is studied as a means to study the author and how he plied his craft. The discussion is of method, technique, composition, structure, balance, symbolism, metaphor, creative innovation. In a language corpus, the focus is on the language itself, not on the author – it is English we are studying, not Shakespeare. We study a language corpus as a cross section of language, a representative sample of all authors or speakers. The discussion is of broad patterns, usage, grammar, syntax, semantics – the generalities of language.

This paper is an inventory of search capabilities in the same spirit as the HUMANIST discussion, but with emphasis on language corpora rather than literary texts – these are the features *we* would like to see in a text retrieval program.

Measuring speech-rate in the Lancaster Spoken English Corpus

Nick Campbell
University of Lancaster

A part of the Lancaster Spoken English Corpus, consisting of approximately five thousand syllables, was transcribed phonemically and measured for duration at the syllable level.

To factor out the considerable variation in syllable size and structure, the transcription was used as input to a set of rules for the prediction of duration, then, by differencing the output of these rules and the measured observations, a factor quantifying the closeness of fit was obtained. Since the rules contained no rate information, this factor can be used as a guide to speech rate variation within the text.

Syntactic and structural features are known to have durational correlates, but semantic discoursal aspects of the text were also found to account for much of the variance.

The computer processing of collocations in the London-Lund Corpus

Mats Eeg-Olofsson
University of Lund

As described in greater detail in Bengt Altenberg's contribution (see above), the work in the project Phraseology in Spoken English will be carried out in three phases:

Phase 1: Collection of data and statistical analysis
Phase 2: Lexical-grammatical analysis
Phase 3: Functional and prosodic analysis

This report will concentrate on the first phase. The chief tasks are to find all recurrent word combinations in the London-Lund Corpus, eliminate all irrelevant combinations, and produce frequency lists.

The end products of the computer processing in this phase will be of two kinds: preliminary material, consisting of a concordance of "maximal" recurrent word combinations, and working material, consisting of a concordance of those maximal recurrent word combinations that are linguistically relevant.

Recurrent word combinations, which are the point of departure for this project, have certain properties that complicate the data processing: they are of variable length and they frequently overlap in the text, especially by inclusion.

The approach to the data processing problems in the project has been strongly inspired by the solution adopted for the production of the third part (Collocations) of the *Frequency dictionary of present-day Swedish* (NFO3, see Allén *et al.* 1975). The main idea is to produce a KWIC concordance of the entire corpus, sorted in zigzag order. This particular sorting order (key word, first word right of key word, first word left of key word, nearest word but one right of key word, nearest word but one left of key word etc) makes it easy to retrieve all recurrent combinations that a certain word enters into and to eliminate instances of combinations that are included in other instances, thus not "maximal".

The above-mentioned NFO3 solution had to be slightly modified, in order to reduce as much as possible the size of the files produced and make it possible to use personal computers for interactive processing.

The processing will take place in the following steps:

1) Production of a preliminary inventory of word pair types in the corpus, stored as a frequency list in a direct access file
2) Production of concordances of recurrent word pairs, one for each corpus text
3) Manual elimination of irrelevant instances from the textual concordances
4) Merging the textual concordances into a single corpus concordance, sorted in zigzag order
5) Retrieval and storage of maximal word combinations from the corpus concordance
6) Elimination of included instances of word combinations from the corpus concordance, now sorted in textual order

The resulting file will constitute the preliminary material, a concordance of maximal recurrent word combinations. The working material

will be produced by manual elimination of linguistically uninteresting instances from the preliminary material.

The first of the above steps (preliminary inventory of word pair types) has now been completed, resulting in a cumulated frequency list containing about 140,000 different word pairs, out of which approximately 40,000 are recurrent in the half-million-word corpus.

The main technical problems in establishing the pair frequency list are the identification of instances of word pairs differing by orthographic variation (capitalization, hyphenation, phonetic transcription), and the reconstruction of the turns in the original conversations from the machine-readable version of the corpus. Apparently these problems cannot be solved fully automatically. Instead, approximate solutions have been worked out, which over-generate slightly in identifying pairs and referring words to the same turn, respectively. The cases where these decisions have to be corrected manually appear to be quite rare.

Reference

Allén, S. *et al.* 1975. *Nusvensk frekvensordbok* 3. (Frequency Dictionary of Present-Day Swedish 3.) Stockholm: Almqvist & Wiksell.

Structure frequency counts of modern English: Progress report

Pieter de Haan
University of Nijmegen

The goal of this project is to present a survey of the distribution of the most common syntactic structures in modern English, based on an investigation of the Nijmegen corpus currently in the LDB. The presentation will be in the form of a book, which will contain:

1. A description of the syntactic categories used.
2. A survey of the way in which the various categories function in the corpus texts, and the way in which the various functions are realised in the corpus texts.
3. A survey of the most common clause patterns found in the corpus, related to the text variety. In this section there will be a cer-

tain emphasis on the basic patterns and their relationship to modified patterns (e.g. because of occurrences of inversion, object preposing, etc.).

The project has been undertaken because it is felt that, although the LDB has been distributed to a great number of places, it would be desirable to have some sort of survey in print, which can be consulted by anyone interested in this sort of general information, without having to dig into the LDB itself.

Secondly, anyone interested can decide, on the basis of the information in this book, at which point he wants to start further exploration of the LDB or indeed any other corpus for comparative studies: most of the basic work will already have been done and possible further research activities may present themselves.

We are now trying to determine what sort of classifications with respect to functions and/or categories can be usefully employed in the presentation of this material. We are thinking of creating classes of complexity, where we would look at the distribution of e.g. NPs with a certain measure of complexity by the functions they have in the various corpus texts (similar to what was shown in my Amsterdam paper). It appears that a classification like the following might be very useful:

I	(det)	Head		
II	(det)	Prem+	Head	
III	(det)	Head	Pom+	
IV	(det)	Prem+	Head	Pom+

where the + means: one or more. In this way we would be able to distinguish between NPs with premodification, those with postmodification, and those with both kinds of modification. Making allowances for the presence or absence of the determiner function in all the classes might make this classification useful for other types besides NPs.

We still have to work out an appropriate classification for PPs and clauses, in terms of complexity. We also have to take a definitive decision as to the exact representation of the figures in the book. The latter is linked up with the way we represent the text varieties in the corpus.

Although in my Amsterdam paper I worked with two categories, Fiction and Non-fiction, we have now found that this sort of division does not lead to homogeneous groups in all cases. An analysis of the most common functions by the most common realisations in relation to the

text categories has shown us that one of the Non-fiction texts behaved quite differently in certain respects from the other two Non-fiction texts. It is obvious that taking these texts as one large group will mean loss of information.

References

Aarts, J. 1984. The LDB: A Linguistic Data Base, *ICAME News* 8: 25-30.

Haan, P. de. 1987. Exploring the Linguistic Database: Noun phrase complexity and language variation. In Meijs, W. (ed.) *Corpus linguistics and beyond.* Amsterdam: Rodopi. 151-165.

Halteren, H. van. 1984. User interface for a Linguistic Data Base. *ICAME News* 8: 31-40.

Form and interpretation in automatic morphological analysis

Theo van den Heuvel
University of Nijmegen

The language-independent Nijmegen approach to syntactic corpus analysis (the TOSCA method) involves a formal description that encompasses the linguistic aspects of many levels. Not only syntactic but also morphological and lexical rules are included, as well as rules that govern punctuation and orthographic conventions (van den Heuvel, 1988). The division was inspired by the need for disciplined intervention (van den Heuvel, 1987). The grammar consists of logical groups of rules.

The role of the grammar as an autonomous linguistic result demands that it should not contain heuristics to speed up the analysis process, such as computational shortcuts or probabilistic information. In this paper we discuss the morphological component for English.

Furthermore the relations between morphology on the one hand and the syntax and the lexicon on the other are considered. A prominent feature of the morphology is the distinction between form and interpretation.

This distinction is useful in the case of English, because many morphemes have more than one interpretation, for example, verb present

tense or noun plural. The ideas behind this setup are illustrated with examples taken from a corpus.

References

van den Heuvel, Th. 1987. Interaction in syntactic corpus analysis. In Meijs (1987): 235-252.

van den Heuvel, Th. (1988). TOSCA: An aid for building syntactic databases. *Literary and Linguistic Computing* 3: 147-151.

Meijs, W. (ed). 1987. *Corpus linguistics and beyond.* Amsterdam: Rodopi.

Dialectal syntax: In search of data

Ossi Ihalainen
Helsinki University

In his book *Dialectology: An introduction* (p. 41) Nelson Francis points out that there has been very little research on dialectal syntax simply because researchers do not have enough texts at their disposal.

The Helsinki Corpus of Dialectal British English is an attempt to provide researchers with large machine-readable texts and thus remedy the situation at least to some extent. In my paper I shall try to show how such a corpus could be used for preliminary basic research leading to a more detailed linguistic description.

The three samples studied (Somerset, Cambridgeshire, Yorkshire) reveal, perhaps somewhat surprisingly, striking differences in the frequencies of some common linguistic items. For example, there turn out to be statistically significant differences in the frequencies of the pronoun IT in the three texts studied. Closer analysis of the situation shows that in Cambridgeshire English THAT replaces IT in many of its standard English functions. Thus, one will regularly find THAT as a formal subject in sentences like "That was easy to get a job" and "That's cold today". As far as the distribution of THAT is concerned, Yorkshire English turns out to be closest to Standard English. Somerset English, while significantly different in its pronominal use from both Cambridgeshire English and Yorkshire English, shares some features with Cambridgeshire English. Thus, THAT may occasionally (but far less frequently

than in Cambridgeshire English) be used as a formal subject IT as in "That's nice up there". On the other hand, unlike in Cambridgeshire English (or Yorkshire English), the personal pronoun HE/HIM is regularly used in Somerset English for the Standard English neuter pronoun IT, as in "He do start at Brue", where HE refers to the river Brue.

Large machine-readable texts also make it very easy to check the accuracy of statements about the distribution of dialectal forms made by earlier scholars. For example, it has been believed that shadow pronouns occur only in Irish and Scottish English, and even in these mainly in possessive constructions ("the man 'at his wife died" 'the man whose wife died'). Machine-readable texts are easy to search for this kind of feature. Such a search will reveal in seconds that shadow pronouns are in no way restricted to Scottish or Irish English. Nor are they restricted to possessives but occur in other functions as well ("This is the certificate that the schoolmaster signed it"), although one must say that possessives are particularly susceptible to pronoun retention.

Other features discussed in this paper are the comitative/instrumental distinction ("Harry walked with his mother" vs "Harry walked with a stick"), relative clauses modifying the NP subject of an existential sentence ("There was an alleyway go through like that"), comparative clauses ("They are cleverer than what we was") and peculiar progressives ("She was sat in that chair").

The most important thing about these texts is, however, that they will reveal a great number of essential syntactic structures that questionnaire-based collections of dialectal material like the SED leave out of account. For example, although ellipsis is a characteristic feature of any spoken variety of English and there is no reason to assume that it works in regional varieties precisely the way it works in Standard English, one looks in vain for information about this syntactic point in the SED.

This is not meant as a critique of the SED. It is as good a collection of data as collections of this type go. But its data-gathering methods were inherently incapable of eliciting the kind of data that are needed to describe grammatical systems.

Although a lot of syntactic information can be gleaned from regular running texts, tagging and parsing these would greatly enhance their value as sources of linguistic data.

Reference

Francis, W. N. 1983. *Dialectology: An Introduction*. London and New York: Longman.

A grammatical database for the Survey of English Usage

Ewa Jaworska
University College London

In its linguistic dimension, the project on the grammatical database aims at the development and the implementation of a consistent and relatively 'theory-neutral' system of concepts and categories with which to describe the grammar of sentences in written and spoken texts. The project is aimed at a variety of researchers concerned with the grammar and use of English. Through a purpose-designed query system, researchers will be able to gain quick access to data on specific grammatical constructions occurring in the corpus instead of consulting a somewhat unreliable catalogue of the grammar of the corpus stored on slips or reading the texts and identifying the constructions themselves. The general terms of the description of each phrase are its structure, function, and position in the minimal containing constituent. The objectives of consistency, efficiency and accessibility of the coding system provide the guidelines for the choice of analyses of specific constructions. The analysis of the Extended Verb Phrase (EVP; a phrase consisting of any modals and auxiliaries and the main verb, and verb complements) will serve to illustrate. In its simple form (e.g., *He might have been talking to them in Coptic*) the EVP is analysed very much in the spirit of the 'verbal complex' tradition while, at the same time, verb complements are its immediate constituents. Other constructions involving the EVP (e.g., *He was very angry and talking to them in Coptic*) lend themselves to an analysis which is more adequate from the empirical point of view.

A grammatical database and query system for the Survey of English Usage

Geoffrey Kaye
IBM UK Scientific Centre, Winchester

The SEU grammatical database is primarily a research tool. It is expected to be in use over many years, and consequently must be flexible and extendible to accommodate enhancements. The criteria for designing such a database have already been presented (Kaye 1988). This design must be complemented by two facilities that are equally flexible and extendible. The first is the data collection system for extracting the relevant grammatical constructions from the texts of the SEU and coding each one with a description of its grammatical properties and function in the context in which it occurs. The second is a query system by which scholars can access the data using familiar conventional grammatical terminology with the minimum learning period. The criterion of extensibility means that the definitions of the grammatical categories must be stored in one place within the system, and be available to all three components; the database and its associated loader, the data collection system, and the query system.

A prototype has been built and tested for a subset of the 'grammar' used in the coding. This has demonstrated satisfactorily the validity of the design criteria established for the project, and has provided the necessary information to enable a fully operational system to be built.

Speed and accuracy of data collection are essential. This has been obtained by removing the need for any data entry via the keyboard and by controlling the grammatical coding from a series of contextually related menus. Which menu is displayed is determined by the selection made from the previous menu. The selection of the text of the grammatical construction, and the actual grammatical coding are all done using a mouse.

A similar overall strategy is followed in the query system. As far as possible the user is not required to type anything on the keyboard unless the search is to be restricted by a specific lexical item, e.g. a particular lexical verb. The same set of menus for describing the grammatical structure and function are used in both the coding and in the query specification. This has many advantages over using a free format English specification such as "Find all prepositional phrases functioning as

an adverbial with a semantic role of space position where the preposition is complemented by a finite clause". Here the user has a significant amount of typing to formulate the query, with the attendant possibility of typing errors, and the query system has to be able to handle the wide variability in specifying the question and the terminological idiolect of the user. A menu system provides in essence a constrained dialogue using a constrained terminology. The user must select items from menus that have been written using a chosen set of grammatical terms. As each selection is made the system presents the user with a new set of descriptive terms that allow the search to be made more restrictive. The options presented are determined by the elections made so far. The user is allowed to specify the search with whatever degree of detail is felt to be desirable. It is not essential to make a selection from every menu. This approach gives the user the comfort of seeing the grammatical items described in familiar grammatical terminology, relieves him of excessive typing, and removes from the query interpreter the need for a detailed grammatical and semantic analysis of the query, which would be needed if a free format natural language interface were used.

The implementation of the system is described and exemplified using sample prepositional phrases.

References

Kaye, G. 1988. The design of the database for the Survey of English Usage. In M. Kytö, O. Ihalainen, M. Rissanen (eds.). *Corpus linguistics, hard and soft.* Amsterdam: Rodopì. 145-168.

Building a lexical database of English

Françoise Keulen
CELEX – CENTRE FOR LEXICAL INFORMATION
University of Nijmegen

CELEX is a joint project involving five Dutch research institutions. During the first stage of the project (January 1986 – December 1988), two multifunctional lexical databases were developed, containing detailed lexical information on Dutch and English. Market research

clearly shows a growing demand for on-line accessible lexicon information both in the academic and the commercial/industrial world. Natural language-like interfaces in man-machine communication, new approaches to (semi-)automatic translation, and the interest in getting more detailed insights into the structure of lexicons call for electronic databases with various types of lexical information.

CELEX provides expertise in accessing and using computerized lexical information with help of the CELEX user interface FLEX, which provides the facilities to consult the databases in a flexible and user-friendly manner. On-line help facilities and a detailed user manual help users specify, create and manipulate their individual lexicons that can be used in fundamental and applied linguistic research as well as in language and speech technology.

The CELEX databases hold information on orthography, phonology, morphology, syntax and frequency information for present-day Dutch and English. The following types of information are available:

orthography	various graphemic representations syllabification (word division) spelling variation
phonology	phonetic transcriptions syllable structure primary and secondary stress patterns
morphology	decomposition into free and bound morphemes affix lists full inflectional paradigms
syntax	word class subclassification
frequency	per headword/wordform/morpheme etc., based on large corpora of written text

The two main sources of lexical information used for the development of the English database are the *Oxford Advanced Learner's Dictionary of Current English* (OALD) and the *Longman Dictionary of Contemporary English* (LDOCE). For the latter, we used the ASCOT version developed by the University of Amsterdam, the Netherlands, which is a restructured and updated version of the original input dictionary. Both sources contain detailed information on stems, (regular and irregular) inflections, phonetic representations, word class codes,

subclassification, language labels (British English, American English) and some minor types of data. Together with additional information such as syllabification (word division positions), inflectional affixation and frequency information, all the data were thoroughly examined, checked, reformatted and, where necessary, corrected and updated. Next, all relevant material was selected, merged and subsequently stored in separate tables using the ORACLE relational database management system (DBMS), which allows for the best possible manipulation and integration of several sources into a well-structured multi-table database.

Numerous conversions and corrections concerned the phonology, syntax, inflectional affixation codes, language labels etc. of stems and wordforms. With respect to phonology, for example, a lot of time was spent on combining and enhancing information from four different sources. Syntactic codes taken from OALD and ASCOT were compared, combined and rewritten into one standard coding system. The inflectional paradigms of stems and their corresponding inflections were thoroughly examined and completed so as to provide all possible regular and irregular inflectional patterns. Some information on British English or American English stems has been included, covering British English spelling variants (*jail – goal*), American English spelling variants for British wordforms (*behaviour – behavior*), typically British English wordforms (*abattoir*), and typically American English wordforms (*sidewalk*).

All operations mentioned above ultimately resulted in the first prototype of the English database (Version 1.0), which contains approximately 30,000 stems and their corresponding inflections (over 50,000), the phonetic transcriptions of these stems, syntactic information (word class and subclassification), spelling variation, frequency counts and other types of lexical information, all of which has been included in the English master database.

Future extensions have been scheduled with respect to, for example, a more elaborate representation of phonetic information (including RP variation), derivational morphology and disambiguated frequency data. Furthermore, the prototype of this database will be enriched with large numbers of new entries such as lemmas from the COBUILD type list that do not occur in any of the dictionary sources but which are clearly a reflection of current English, 'multi-words' from OALD and ASCOT, abbreviations, proper and geographical names. This will ultimately result in a lexical database – comprising the five basic types of information – which is as fully as possible a reflection of present-day English.

Patterns of collocability

Göran Kjellmer
University of Göteborg

A study was made of the tendencies different types of words have to occur in clusters in modern (American) English. Two corpora of text were used as a basis for the study, the Brown Corpus and a sub-corpus containing all the collocations (recurring, grammatically well-formed sequences) contained in it. The results of a comparison between the two corpora were unequivocal. Words differ very markedly in their tendency to cluster. Singular nouns and base forms of verbs are highly collocational while adjectives and adverbs are not. Singular proper nouns are ambiguous in this respect. What decides whether a type of words shows this tendency to cluster is very largely some functional or contextual restriction of the type. There is a continuum in English words (including names), from those whose contextual company is entirely predictable (*Angeles, Fidel*) to those whose contextual company is entirely unpredictable (*therefore*), but the evidence indicates that most of the words are to be found towards the *Angeles* end of the scale.

Progress report on the Amsterdam projects

Willem Meijs
University of Amsterdam

ASCOT

After a six-month extension subsidized by the Amsterdam University Arts Faculty the ASCOT project (financed by the Dutch Research Organization ZWO) was finished 31 August 1987. Detailed descriptions of various aspects of the work in the second stage of the project can be found in Akkerman (1988) and Akkerman et al. (1987, 1988a, 1988b). The resulting software – a scanning system which automatically attaches detailed syntactic information to the words in an input text (including derived and/or inflected ones, as well as "multi-word" combinations) on a context-free basis – is in principle available for

bona fide academic researchers, subject to agreement by Longman Publishing Company. Alternatively one can send in pre-edited texts to be processed by the ASCOT system.

MORPHGRAM

Work on an improved morphological component to replace the one now contained in the ASCOT package will start towards the end of 1988 under the heading MORPHGRAM. As the name suggests this will be a morphological analyzer in the form of a grammar, to be written in the PARSPAT formalism (van der Steen 1987). MORPHGRAM is expected to be considerably faster and yet more comprehensive than the present morphological analyzer, and will be more in line with recent psycholinguistic left-to-right models of lexical access to morphologically complex words – cf. Bergman (1988), Meijs (forthcoming).

LINKS in the Lexicon

In 1987 the research-work in the LINKS project involved: investigation of an extensive sampling of definitions with "Human" and "Abstract" kernels with a view to the choice between a syntactic or a more semantically-oriented approach; further refinement of the inventory of definition-types, the development of a typology of prominent definition-structures and the development of a parser-grammar for noun-definitions.

After the parser-grammar for noun-definitions similar grammars will be developed for verb and adjective definitions. The results of the analyses are being stored in an LDB-"shell" as developed at Nijmegen University. This will enhance the possibilities for systematic inspection, sampling etc. On the basis of the syntactic-semantic typology (cf. Vossen 1988, Vossen et al.1988a, 1988b), the various types of syntactic definition-structures are then fitted out with various semantic parameters. The analyzed and semantically enriched definitions will then finally be stored in a database structure which should allow easy and systematic cross-referencing. Due to a number of external factors the LINKS-project is running somewhat behind schedule. We therefore intend to file a request with ZWO for additional funding for a period of half a year in 1989. Extra support for some of the activities planned in the LINKS-project may also derive from a new project called LEXALIZA, which began in February 1988.

LEXALIZA

None of the existing systems for (semi-)automatic analysis and text-processing (can) make use of such a rich set of syntactic, semantic and stylistic data as can be provided by the ASCOT and LINKS packages. Rather than use existing programs, therefore, we felt we should develop a processing-system that is specially geared to the ASCOT/LINKS data. This led to plans for a new project, LEXALIZA. The acronym stands for "LEXicaal-gestuurde anALYSE" ("lexically-guided analysis") and the aim of the project is thus the development of a language-processing system which capitalizes as much as possible on the kinds of syntactic and semantic information provided by ASCOT and LINKS and yielding a proportionally rich description/analysis of input texts. The plan envisages a four-year research project, starting 1 February 1988. Funding for the project (to be carried out by one full-time research assistant) was sought and obtained from the Amsterdam University Arts Faculty.

Some of the basic ideas underlying the project are presented in Meijs (1988a and 1988b), who discusses experimental psycho-linguistic findings in connection with inferencing, priming and spreading semantic activation which are compatible with neuro-physiological evidence as well as with Functional Grammar views that a great deal of what is normally called "world-knowledge" is verbally coded in much the same way in which such knowledge is represented in dictionary definitions. A central assumption in LEXALIZA is that computational discourse-processing should resemble human discourse-processing as much as possible. What the experimental evidence shows is that human text-processing is basically parallel processing: the incoming data are constantly being checked for consistency against the data contained in long-term memory and against the results of what has been processed before, and this happens simultaneously on all levels and in an interactive fashion, down from elementary signal level, via the morphological, syntactic and semantic levels, through to higher pragmatic, organizational levels.

References

Akkerman, E., W.J. Meijs and H.J. Voogt-van Zutphen. 1987. Grammatical tagging in ASCOT. In W.J. Meijs (ed) *Corpus linguistics and beyond.* Amsterdam: Rodopi. 181-193.

Akkerman, E., W.J. Meijs and H.J. Voogt-van Zutphen. 1988a. ASCOT: A computerized lexicon with an associated scanning system. In M.

Kytö, O. Ihalainen and M. Rissanen (eds) *Corpus linguistics, hard and soft.* Amsterdam: Rodopi. 35-43.

Akkerman, E., H.J. Voogt-van Zutphen and W.J. Meijs. 1988b. *A computerized lexicon for word-level tagging.* ASCOT Report No 2. Amsterdam: Rodopi.

Akkerman, E. 1988. An independent analysis of the LDOCE grammar coding system. In *Computational lexicography for natural language processing,* eds. B. Boguraev and T. Briscoe. London: Longman.

Bergman, M.W. 1988. *The visual recognition of word-structure: Left-to-right processing of derivational morphology.* Doct. Diss. Nijmegen University. Enschede: Sneldruk.

Meijs, W.J. 1988a. Knowledge-activation in a large lexical data-base: Problems and prospects in the LINKS-project. In *Amsterdam Papers in English* (APE) No I. English Department, Amsterdam University.

Meijs, W.J. 1988b. Spreading the word: Knowledge-activation in a functional perspective. To appear in J. Connolly and S. Dik (eds), *Functional grammar and the computer.* Dordrecht: Foris.

Meijs, W.J., forthcoming. Morphology and word-formation in a machine-readable dictionary: Problems and possibilities. To appear in *Morphologica* (1988).

van der Steen, G.J. 1987. *A program generator for recognition, parsing and transduction with syntactic patterns.* Doctoral Diss., Amsterdam University.

Vossen, P. 1988. The meaning descriptions in the lexicon provided by the LINKS-project. To appear in J. Connolly and S. Dik (eds), *Functional grammar and the computer.* Dordrecht: Foris.

Vossen, P., M. den Broeder and W.J. Meijs. 1988a. The LINKS project: Building a semantic database for linguistic applications. In M. Kytö, O. Ihalainen and M. Rissanen (eds) *Corpus linguistics, hard and soft.* Amsterdam: Rodopi. 279-293.

Vossen, P., W.J. Meijs, and M. den Broeder. 1988b. Meaning and structure in dictionary definitions. In *Computational lexicography for natural language processing,* eds. B. Boguraev and T. Briscoe. London: Longman.

Prepositions in LOB and BROWN

Dieter Mindt
Freie Universität Berlin

Prepositions in LOB

In order to find out the frequency and distribution of the prepositions in LOB we made a frequency count of the items bearing the tag IN in LOB.[1] There are 134 items bearing the tag IN. Of these, 37 items could not safely be identified as prepositions. For this reason we assigned them to a separate list named "unclear cases". In this list there are items such as *wid* or *as though* which require closer inspection before they can be added to the list of clearly identified prepositions.

The list of unclear cases contains 37 items, whose frequency of occurrence in LOB is 616. The total number of occurrences bearing the tag IN is 124,369. The 616 occurrences of unclear cases make up .47% of all the occurrences of items bearing the tag IN. We therefore felt justified to exclude the unclear cases from the first steps of the analysis of English prepositions. This leaves 97 prepositions with a total of 123,753 occurrences, which means that about every eighth word in the LOB Corpus and probably in English texts in general is a preposition. On this basis, we compiled a ranklist of the prepositions in LOB. Table 1 gives the figures for prepositions with more than 1,000 occurrences.

Table 1: A ranklist of prepositions in LOB

Preposition	Frequency of occurrence in LOB
of	35,287
in	20,250
to	10,876
for	8,738
with	7,170
on	6,251
by	5,724
at	5,473
from	4,672
as	2,804

into	1,658
about	1,282
than	1,020

The ranklist presents the central prepositions with a frequency of 1,000 and more occurrences which can be divided into two groups. The complete ranklist of the 97 prepositions makes it possible to distinguish between central and marginal prepositions. There are four groups of prepositions in LOB:

Table 2: Groups of prepositions in LOB

Group	Range of frequency	Number of prepositions	Absolute frequency	Relative frequency
Group 1	over 20,000	2	55,537	44,87%
Group 2	1,001 - 11,000	11	55,668	44,98%
Group 3	101 - 1,000	32	11,164	9,02%
Group 4	1 - 100	52	1,384	1,11%

It is obvious that the core of English prepositions is made up of groups 1 and 2. The prepositions in group 1 and group 2 cover 89.85% of all occurrences of clearly identified prepositions in English. This means that about 90% of English prepositional usage can be explained by 13 prepositions. Within this central area the two most frequent prepositions, *of* and *in*, account for about every second occurrence. Groups 3 and 4 contain marginal prepositions. These groups contain 84 prepositions altogether. These 84 prepositions account for only about 10% of prepositional usage in English.

Prepositions in BROWN

It is interesting to compare the distribution of the central prepositions in British English with the distribution of the same prepositions in the BROWN Corpus of American English. The data published by W. Nelson Francis and Henry Kučera, *Frequency Analysis of English Usage* (Boston 1982) can serve as a basis of comparison. For the prepositions with a frequency of 1,000 occurrences or more (groups 1 and 2) the results in comparison with the figures from LOB are given in table 3.

Table 3: Prepositions in LOB and BROWN

Preposition	Frequency of occurrence in LOB	Frequency of occurrence in BROWN
of	35,287	36,432
in	20,250	20,870
to	10,876	11,165
for	8,738	8,996
with	7,170	7,286
on	6,251	6,183
by	5,724	5,246
at	5,473	5,377
from	4,672	4,371
as	2,804	121
into	1,658	1,790
about	1,282	1,242
than	1,020	497

There are the following distributional similarities:

1. Both in LOB and BROWN there are clearly two identical groups of prepositions, with *of* and *in* belonging to the first group in both corpora.
2. The two most frequent prepositions (*of* and *in*) have the same rank order in British and American English.
3. The rank order of the four following prepositions (*to, for, with, on*) is also identical.
4. The application of the chi-square test leads to the result that there are no significant differences in the distribution of the two most frequent prepositions in LOB and BROWN (*of* and *in*).
   Chi-square: .14 df:1 0.05: 3.84[2]
5. For the six most frequently used prepositions in LOB and BROWN there is also no significant difference in distribution.
   Chi-square: 5.45 df:5 0.05: 11.07

The six most frequent prepositions in LOB have an occurrence of 88,572, which means that they cover 71.57% of the occurrences of the clearly identified prepositions in LOB. The same six most frequent prepositions in BROWN have an occurrence of 90,932. The total frequency of the elements with the tag IN in BROWN is 122,601 (Fran-

cis and Kučera 1982:538). The six most frequent prepositions in BROWN cover 74.17% of all occurrences of prepositions in BROWN.

Except for the prepositions given in table 3 there is no preposition in BROWN with a frequency of occurrence exceeding 1,000 so that in this respect there is also a very close correspondence between British and American English. Striking differences between LOB and BROWN have to be stated for the prepositions *as* and *than*, which occur in tenth and thirteenth position in LOB. Their frequency in BROWN is so low that they could be assigned to neither group 1 nor to group 2. Whether this is due to a difference in prepositional use between British and American English or rather to a difference in the tagging of the LOB and BROWN corpora can only be decided after a closer inspection of these prepositions.[3]

Conclusion

There is a close correspondence between British and American English with regard to the distribution of prepositions. The functional core of the English language which is represented by the prepositions seems to be largely identical irrespective of other differences between British and American English.

Notes

1. The analysis of the prepositions in LOB was made in collaboration with Christel Weber.
2. We use the following notational conventions: "Chi-square" stands for the empirical chi-square value; "df" stands for the number of degrees of freedom; "0.05" gives the critical value on the significance level alpha = 0.05.
3. The total frequency of occurrences of *as* in LOB is 7,337, and 7,254 in BROWN. For *than* there are 1,646 occurrences in LOB and 1,794 in BROWN. This supports the hypothesis that the difference is due to a discrepancy in tag assignment. [Editor's comment: confirmed during the discussion of the paper]

CD-ROM corpora

A. Moulin and Jacques Noël
Université de Liège

With no more than £2,000 worth of equipment (a PC with a hard disk, and a CD-ROM reader) we can now interrogate more text corpora and more machine-readable dictionaries than could be searched on our university mainframe only just a year ago. The promising recent developments are also in the field of software, and in particular of various types of retrieval and concordancing software packages which can now be run on our micro- or minicomputers: BYU Concordance, which we use to interrogate our bilingual dictionary and text corpora (work by J. Jansen); dBase III, which has so far been used in our work on LDOCE, the *Longman Dictionary of Contemporary English* (work by A. Michiels and J. Jansen).

Our most novel pieces of equipment, however, are the CD-ROM reader and two encyclopedias on compact disk (GROLIER and McGRAW-HILL). We have no figures available on McGRAW-HILL, but there is no doubt that it is a very large corpus of science English. As to GROLIER, it is not restricted to science and represents some 60 megabytes (some 9 million words) of text. The problem which we want to address is how useful GROLIER and McGRAW-HILL are as language corpora, in English lexicography as well as in ELT (English Language Teaching) at university and other advanced levels. In order to reduce our problem to somewhat more manageable proportions, we focused on adverbs in *-ly*, with special attention paid to examples of verb-adverb collocations. For convenience, we decided to use a checklist generated from adverb entries and from run-ons of adjective entries in our LDOCE database. For the rest, our study of adverbs in GROLIER and McGRAW-HILL relies most heavily on the COLLINS-ROBERT bilingual dictionary, under BYU Concordance, and, to a lesser extent, on BYU searches carried out on the Brown Corpus.

Let us first emphasize a few limitations inherent to this paper. We have limited ourselves to one-word adverbs in *-ly* and did not attempt to deal with multiword adverbial phrases, despite the fact that multiword searches are easy in both CD-ROM files, thanks to the excellent retrieval software provided with our encyclopedias. This limitation does not allow us to do justice to what is probably the most useful feature

of monolingual dictionaries like COBUILD and bilingual ones like COLLINS-ROBERT: the attention they pay to formulaic and idiomatic aspects of everyday language in a single work of reference. Bearing these limitations in mind, we can say that our study confirms most of our expectations. If you are interested in adverbial usages and collocations that belong exclusively to spoken or conversational language, an encyclopedia is obviously not the place to go even if GROLIER does offer the occasional mention of a popsong or the like: "I honestly love you" (Olivia Newton-John). Exemplification of "signals" and other aspects of spoken language which is undoubtedly one of the strengths of the new COBUILD dictionary (see entries like *actually*, *honestly*) is simply not to be found in McGRAW-HILL or GROLIER. For the rest, however, as an example database, GROLIER is quantitatively and qualitatively richer than BROWN and LOB. It is the equivalent of the best dictionaries, monolingual and bilingual (COLLINS-ROBERT) with respect to not only the range of one-word adverbs (from *abjectly* to *zoologically*, with all the functions recognized by Quirk et al 1985) but also the range of collocations attested. Only the most speaker-oriented disjunct functions and positions are so to speak not attested; for example, "*Honestly*, I..." (unambiguously disjunct) versus "I honestly love you", which seems to some extent ambiguous between the predominantly subjunct function and the disjunct reading. Our study also confirms our hunch and first impressions that McGRAW-HILL's limitation to science and technology restricts its range of adverbial usages to whatever belongs conceptually to science discourse. Our encyclopedias are of course unique as sources of examples of adverbial usage in science English (example: *broadly*).

Let us now turn to the beginning of our checklist (not reproduced here) of adverbs in *-ly* (generated from LDOCE) together with the number of occurrences of these adverbs in GROLIER and McGRAW-HILL. Quite arbitrarily, our list goes from *abjectly* to *alertly*. We have not tabulated the frequencies in BROWN and LOB, which can be found in *L.E.T. vocabulary list* (Engels et al 1981). The following figures for a selection of items from the list – for GROLIER, McGRAW-HILL, and BROWN, respectively – show that in all cases the number of occurrences in BROWN is smaller, even than that in McGRAW-HILL: *abjectly* (1, 0, 1), *ably* (23, 0, 2), *abnormally* (36, 18, 1), *actually* (566, 131, 128). BROWN's scores are at best comparable to McGRAW-HILL's, even if the latter's range of adverbs seems definitely more restricted than that attested in BROWN. For the rest, our table shows

how rich the range of adverbs in GROLIER is as compared also to the range attested in the other encyclopedia.

Lastly, let us consider the treatment of just one item – *bitterly* – in COBUILD as a representative example of how a good monolingual dictionary captures collocation. COBUILD enters only one definition of *bitterly* in the (subjunct) sense of "resentfully" in the entry for *bitter*: "... he said bitterly". Unfortunately, COBUILD fails to cross-refer to the entry for *bitterly* in Definition 3 (A bitter argument, war, struggle...: ...bitter fighting,... campaign) and Definition 4 (A bitter wind or bitter weather: ...bitter cold), as these are obviously related to the two definitions of *bitterly*; these seem to characterize two (disjunct) uses which we take to involve collocations, because the choice of a synonym – including the one given in the COBUILD column (=desperately) – would be odd:

1. Bitterly means strongly and intensely; used to refer to strong emotions such as anger, hatred, shame, or misery. EG No man could have hated the old order more bitterly... He was bitterly ashamed... I was bitterly disappointed... She sat in her room and wept bitterly.
2. If the weather is bitterly cold or if you are bitterly cold, it is extremely cold or you are extremely cold. EG On a bitterly cold New Year's Day.

To conclude, let us list the combinations of *bitterly* with adjectives and verbs that we found in GROLIER: a total of 55 occurrences with some items like *cold* and *opposed* occurring more than once:

Adjectival: antagonistic, cold, compassionate, disappointed, dystopian, intolerant, ironic, satiric, satirical

Verbal: attacked, condemned, complained, contested, criticized, divided, fought, opposed, quarrelled, resisted

Our conclusion is that, irrespective of whether or not these cooccurrences can rightly be viewed as collocations (cp the problem of ambiguity between adjunct and disjunct), they cannot be generated from the dictionary entries and examples: if we are correct, GROLIER is an indispensable source of examples, as are all major corpora. The new technologies (CD-ROM, WORM, etc) will, it is to be hoped, increasingly make authentic text and examples more and more generally available more and more inexpensively. This would be particularly helpful for non-native learners and teachers.

References

Dictionaries and other reference works

COBUILD *Collins COBUILD English language dictionary.* London: Collins. 1987.

GROLIER *Grolier, The electronic encyclopedia. A 20-volume encyclopedia on CD-ROM.* Danbury, CT. 1986.

LDOCE *Longman dictionary of contemporary English.* London: Longman. 1978.

McGRAW-HILL *McGraw-Hill science and technical reference set.* 1987.

ROBERT-COLLINS Le Robert – Collins. *Dictionnaire francais-anglais, English-French.* 1985.

Other references

Engels, L. K. et al. 1981. *L. E. T. vocabulary list.* Leuven: Acco.

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

Sinclair, J. McH. (ed.). 1987. *Looking up: An account of the COBUILD project in lexical computing.* London: Collins.

Work in progress at the English Language Research and Development Unit

Antoinette Renouf
University of Birmingham

In the area of corpus development, we are continuing to amass data. The 1 million word SHAPE corpus of speech and writing was completed in March 1988, and the UCLES corpus of examination papers is almost ready for statistical analysis. In response to the 1985 ICAME request, we are assembling a new spoken corpus, aiming at 1 million words, and trying to capture at least some data that is not easily accessible to researchers, such as the commentary that accompanies joint activity; and incidental chat. At the time of speaking, this corpus stood at around 200,000 words, 100,000 of which had been transcribed.

Recent research activity has focussed on the design of a number of projects, including one for the development of a standardised word-list/lexicon, and another for a system of automatic text summarisation. A series of grammar seminars in the Department have been taking place to establish a framework for a range of grammars.

Software development has centred on the tagger, which has been refined to produce a more accurate representation. It runs on concordances and the output can in turn be concordanced, on any combination of word and tag. Some improvements have also been made to the corpus access routines. Enhanced on-line facilities exist for the 1 million word sub-corpus on the PDP, and it is now possible to make overnight requests for concordances from the full 18 million word corpus.

John Sinclair and Jeremy Clear demonstrated some new software that interactively produces concordances for a given word and sorts them in order of collocational 'typicality' according to a significance algorithm.

Compiling a corpus of East African English

Josef Schmied
University of Bayreuth

The proposed project for a corpus of (contemporary educated) East African English (abbreviated CEAE) will comprise texts written by (Black) Kenyans and Tanzanians between 1986 and 1990. This first presentation of the project explains the rationale it is based on, gives some guiding principles and examples, and describes some of the problems encountered or anticipated.

A corpus of East African English can not only serve as a useful text basis for the quantitative study of a language variety, but also be applied more specifically in the discussions of some current problems with English in East Africa. This is considered particularly important as the preparation of a corpus should not be seen in isolation from its intended uses. It can, for instance, be used as a guideline in deciding whether certain unusual expressions in East African English are individual *ad hoc* formations, jargon with a specific limited range of currency, or general usage. In view of the common complaints about falling standards of English in many African countries a corpus will pro-

vide a point of reference to determine common educated usage. Finally, the CEAE can also serve as a basis for comparative variety studies, not only of first-language varieties, as in the Brown and LOB corpora and the proposed Australian corpus, but also of second-language varieties, such as the Kolhapur Corpus from India. These purposes may at times be contradictory. The aim to keep the corpus comparable to previously compiled corpora may be difficult to achieve in the sociolinguistic context of a completely different region of the world. Compiling corpora for second-language varieties implies the need to define an English-speaker, in the sense that only the well-educated or those who use the language in their daily activities can be considered reasonably proficient English speakers. Although it might be interesting to include different levels of proficiency in order to define intra-variety stages in the teaching norm, this would spoil the results for inter-variety comparisons.

As the CEAE will also be used for sociolinguistic analysis, the information collected about the authors of texts is particularly important. Normally information about the author's sex, age, education, region and first language will be included. This, of course, constitutes a problem if the texts are taken from printed material; the general guideline here will be the author's name. As there is still a close correlation between birth place and mother tongue in East Africa the name usually gives a good indication of this social background. The selection of texts within the categories will therefore not be random, but according to principles of mother tongue and origin.

As most comparable corpora the CEAE will comprise at least one million words. The basic text categories familiar from Brown and LOB will be maintained, but other text types will be added in a monitor corpus for comparative and specific analyses, as in the Australian corpus project. The general length of texts will be 5,000 words, but longer texts will be included in complete form in a larger monitor corpus. Although, for comparative reasons, all more recent corpus projects try to maintain the same general text categories as the older corpora, they also have to reflect national differences in environment, publication processes or reading habits. In addition to the text categories in Brown and LOB a systematic selection will be made of specifically produced texts written by Africans from a closed set of sociolinguistic variables, i.e. an equal number of members of both sexes (male – female), different socioeconomic and age groups (form 4, form 6 and university students, etc.), and first-language background.

The general coding conventions will be taken from the LOB Corpus. In addition to the general coding, a specific coding for East Africanisms (constructions unusual in British Standard English) is planned as a first stage of analysis.

Some sampling problems deserve attention. When English speakers have learnt English as a second language, it is, for instance, often difficult to decide whether they have acquired a stable knowledge of the language or whether they are still learning it. For the purpose of this corpus, speakers of East African English must at least have secondary education. Another difficulty in connection with longer texts may occur in representing each text category appropriately (e.g. very few novels are published in Tanzania at the moment because of economic constraints). A problem concerning short texts (e.g. letters-to-the-editor) is that several have to be combined to provide a text unit of 5,000 words. In this case texts by authors from the same regional background (as judged by their names) will constitute one unit of text category. Another problem of text selection is that it will obviously be impossible to select the corpus by stratified random sampling, because there are no annual national bibliographies in East Africa. Due to the nature of the East African publishing scene many books not published by Longman or Heinemann are sometimes difficult to trace even just shortly after publication. But the advantage of a corpus is, of course, that it provides access not only to texts of one category, but to many texts of various categories.

The discussion concentrated on the questions as to what extent the Brown and LOB categories have to be maintained in text sampling, whether texts by White and Indian authors should be included and how it was possible to find a norm as a basis for later analysis.

What is the role of discourse signals in sentence grammar?

Anna-Brita Stenström
Lund University

In this paper I suggest a combined model for analyzing 'discourse signals', a model that also indicates the syntactic functions (in relevant

cases) of items that are generally used for pragmatic purposes in the conversational interaction. Discourse signals can be realized by single words (*yes, right, now*), two-word combinations (*I mean, you know*) and longer strings (*it's allright, I'm sure that's right*). We use them in order to:

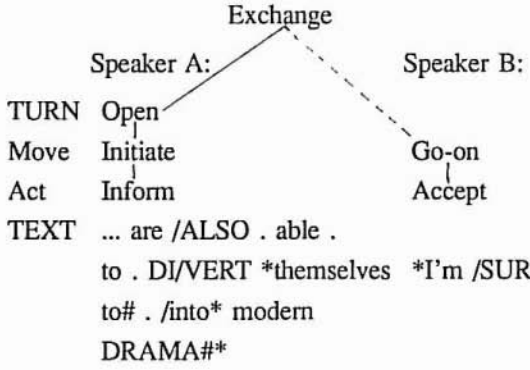
* take, keep, and yield the turn
* empathize with the listener and invite feedback
* structure the message

The starting-point for determining what a particular item does in a conversation is its vertical position in the exchange and its horizontal position in the turn. Is it used as a purely interactional move/act, or should it rather be described syntactically as a clause constituent, or maybe both? Naturally, position is not the only decisive factor. Exactly what the item does is ultimately the combined effect of its position in the exchange/turn and its literal meaning.

Most of the items that we use as discourse signals can be provided with two or more word-class labels pointing to their potential functions in a sentence. *Okay*, for instance, can be labelled both adjective ('it's okay') and verb ('I'll okay this'), and *right* can be labelled adverb ('right through Paris'), adjective ('a right angle'), noun ('the right to vote'), and verb ('to right a wrong'). Exceptions are *mhm*, *yes*, (*yeah*), *ah*, *aha*, *ooh* and *oh*, which are purely interactional and can only be assigned discourse tags.

Used as discourse signals, single words can only be analyzed at the level of interaction. As such, they are not integrated in any clause structure. By contrast, discourse signals consisting of two or more words can always be analyzed both at the interactional level as moves and/or acts and at the grammatical level as phrase and/or clause constituents.

If the model I suggest is applied to an example from the corpus (the London-Lund Corpus of Spoken English) we get the following analysis (the broken line indicates 'no speakershift'):



CLAUSE	S	V	C	S complement	V	C
Phrase	NPH	VPH	JPH	NPH	VPH	JPH
Word	RA	VB	JA	RC	VB	JA

which shows the two-level analysis of 'I'm sure that's right'. This string of words serves as a Go-on move at the level of interaction and as a complex sentence at the level of grammar.

A model of this kind facilitates the analysis of discourse functions and emphasizes the multiple use of a number of speech-specific items, here referred to as discourse signals. It has obvious pedagogical applications and will be even more useful when, as I hope, the analysis can be achieved directly on the computer screen.

Using WordCruncher as a means of studying lexical patterning and thematic progression

Kay Wikberg
University of Oslo

The aim of this paper is to show how the study of text structure can be made more efficient by using WordCruncher, software developed for lexical analysis by Brigham Young University. For this pilot study, three texts (here referred to as F, G and H), 1,000-1,700 words long,

all dealing with the subject of light, were taken from photographic manuals in English.

The concept of 'lexical pattern' was defined as "thematically linked clusters of lexical items which, through the combination of recurrence and information dynamics, contribute to the development of the discourse topic".

The most time-consuming part is writing the texts on to DOS files. However, preindexing the files, which basically involves putting in reference codes for paragraphs and pages and removing possible word processing codes, is much faster since you can use macros and rely on automatic numbering.

The first step in the actual analysis involved spotting the most common central lexemes in the texts, the frequencies of which were provided by WordCruncher. The initial paragraph of each text was crucial for both lexical cohesion throughout the texts and for global coherence. The Combine function turned out to be particularly useful since it allows you to handle a base and its inflected forms as a lexical item and to relate this to a list of collocates. The occurrence of combinations of forms from each list can be examined at different spans from each other.

The second step in the analysis involved identifying themes and rhemes, which was done manually, and relating the distribution of the central lexemes to information structure. The concordance produced by WordCruncher helped to show how central lexemes (word types) in thematic (or, less frequently, rhematic) position developed the topic by providing more specific information such as type of light (e.g. natural light, ultraviolet light), properties of light/sunlight/colour, etc.

It was possible to find differences in the three writers' style. Thus the authors of text G put the focus on the photographer whereas the author of text F, particularly, focused on sun light. Frequent use of *when*-clauses in text H showed that the author was very interested in lighting conditions.

Although the speed and many-sidedness of WordCruncher is best seen when dealing with big corpora, its application to this rather small corpus provided a wealth of precise lexical information to support the manual analysis. The only problem that arose was when the text once disappeared from the screen when searching the Brown Corpus. The error was simply remedied by reinstalling the programme.

Reviews

Dieter Mindt. *Sprache, Grammatik, Unterrichtsgrammatik: Futurischer Zeitbezug im Englischen I & Darstellungen*. Frankfurt am Main: Diesterweg, 1987. (Schule und Forschung. Schriftenreihe für Studium und Praxis). ISBN 3-425-04443-5. 150 pp. + 71 pp. Reviewed by **Herman Wekker**, University of Groningen.

The purpose of Dieter Mindt's book is to demonstrate a new way of compiling grammars for the teaching of English as a foreign language, and the example used for this purpose is that of future time reference in English. Mindt examines what a corpus of present-day English and two widely used English grammars have to say about future reference, and he offers a comparison of the results of these analyses. Given the range and complexity of the data to be considered, the grammatical analyses could only be carried out with the help of a large computer and of certain sophisticated statistical techniques (in particular cluster analysis). The author is right in claiming (p. 9) that the results of his study throw new light not only on hitherto neglected aspects in the description of English, but also on the way in which the textbooks have been put together.

The book opens with a short chapter on the role of linguistics and grammar in English language teaching, and a second chapter on recent developments in language acquisition research. The chapters that follow are concerned with an introduction to Mindt's new concept of pedagogical grammars (chapters 4 and 5), a description of the body of texts and the relevant expressions of futurity (chapters 6 and 7), an analysis of the data (chapter 8), a survey of the distribution of futurity expressions and of the types of co-occurrence (chapters 9 and 10), and finally a summary of the results, with suggestions for further research (chapter 11). The book is accompanied by a separate 71-page appendix, entitled *Darstellungen*, which contains the tables and the diagrams that go with chapters 9 and 10. Both the book and the appendix are written in German.

Mindt's basic assumption is that pedagogical grammars should be constructed according to principles that are independent of pure linguistic grammars. He argues that materials deriving from linguistic theories are not the best ones suited for language teaching. What he sets out to do is to show that it is possible to develop an independent theory for the construction of grammars for foreign language teaching. The

author's approach differs from that of other foreign language acquisition researchers in that he strongly emphasizes the value of an analysis of foreign language data for the construction of teaching materials; the latter, in his view, should come as close as possible to the authentic use of the language by native speakers.

Mindt distinguishes carefully between what he calls didactic grammars and pedagogical grammars. Didactic grammars are defined as linguistic descriptions of corpora compiled for specific language teaching purposes, the existing English language corpora being felt to be unsuitable. The compilation and the analysis of such specific corpora should be related to the needs of various target groups of foreign language learners. Frequency of occurrence and patterns of co-occurrence are among the most relevant parts of the analysis. Every specific corpus will yield a specific didactic grammar. Pedagogical grammars, on the other hand, are seen as derivatives of didactic grammars, in the sense that they make selective use of the linguistic information contained in the didactic grammars. However, pedagogical grammars should not be based exclusively on the structural and distributional analyses of didactic grammars. They should also take into account the demands that are made by language teaching methodology with regard to the selection, grading and presentation of teaching materials. Other, e.g. psychological and socio-cultural, factors are seen to have an equally important role to play. Unlike didactic grammars, which are intended to be purely descriptive, pedagogical grammars will be prescriptive. One didactic grammar may serve as the source of one or more pedagogical grammars for different aims or groups of learners.

The present study is a first attempt to develop a didactic grammar of future time reference in modern British English, involving the following eight expressions of futurity: (1) *will* + infinitive, (2) *shall* + infinitive, (3) *going to* + infinitive, (4) the present progressive, (5) the simple present, (6) *will* + progressive infinitive, (7) *shall* + progressive infinitive, and (8) *going to* + progressive infinitive. Mindt's corpus, which he wanted to be representative of colloquial spoken English, consists of two parts: (a) the 34 fragments of the *Corpus of English Conversation* (Survey of English Usage), totalling 170.000 words, and (b) 12 British plays, totalling about 184.000 words. The Survey fragments, labelled CONV, were recorded between 1953 and 1976, and of the drama texts, labelled PLAYS, one appeared in 1963 and the others between 1971 and 1980. In order to be able to compare his corpus findings with the contents of current teaching materials, Mindt, in addition,

examined two courses that are widely used by 10 to 15 year-olds in schools in the Federal Republic. The works are: (a) *English-H*, labelled H, consisting of 5 volumes, totalling about 128.600 words, and (b) *Learning English Modern Course*, labelled LE, also consisting of 5 volumes, totalling about 152.500 words. The complete body of materials studied adds up to about 635.100 words.

The main conclusion of Mindt's investigation is that the reference grammars of English provide insufficient and often misleading information on the use of expressions of futurity. Teaching materials not based on an explicit plan or on information taken from grammars turn out to be closest to authentic English usage, as it appears from the CONV and PLAYS texts. Other interesting results of Mindt's corpus analysis are the high degree of homogeneity of the CONV and PLAYS texts, as far as the distribution and the co-occurrence of the eight futurity expressions are concerned, the high overall frequency of *will* in comparison to *going to*, the unexpected importance of *shall*, the striking infrequency of the remaining expressions, the over-emphasis on the use of *going to* in relation to *will* in both H and LE and the complete absence of *shall*.

This study has a great deal to offer to textbook writers, teachers and teaching methodologists. It can serve as an excellent example of the kind of corpus-based linguistic research that is immediately relevant to the practical needs of second and foreign language teachers. Further similar research in other areas of English grammar is highly to be recommended.

Meyer, Ch. F. *A Linguistic Study of American Punctuation*. New York: Peter Lang, 1987. Reviewed by **Anna-Brita Stenström**, Telemark Regional College.

'How do we punctuate?' and 'Why do we punctuate the way we do?' These are the questions that Meyer sets out to answer in this study.

American punctuation is known to have become highly conventionalized, the main reason being the strong influence exerted by style manuals and usage books. These, Meyer argues, have a common defect: they conceal the systematic nature of punctuation by viewing it as consisting of individual marks governed only by syntactic considerations and deal with punctuation merely in terms of rules disregarding the underlying principles.

In order to give an accurate description of current American usage Meyer studied the punctuation norms followed in a subcorpus of the Brown Corpus, representing three contrasting styles: journalism, fiction and learned writing. He concentrated on the 'structural' markers: period, question mark, exclamation mark, comma, dash, semicolon, colon, and parenthesis.

Meyer describes punctuation as an integrated hierarchical system based on linguistic principles, sometimes affected by stylistic and pragmatic considerations.

From a syntactic point of view, punctuation marks form a four-level hierarchy, with the period, the question mark and the exclamation mark at one end and the comma at the other. Level 1 marks set off sentences, while the syntactic units set off by marks at levels 2, 3 and 4 are progressively less complex. Due to their hierarchical nature, the marks reveal the syntactic structure of a text by indicating whether constituents are syntactically superordinate or subordinate. Marks used 'out of place', on the other hand, have a stylistic effect, eg by causing constructions to become more closely or more distantly related or by highlighting a particular construction.

From a semantic point of view, punctuation marks can either be used to create a semantic effect, eg by distinguishing restrictive from non-restrictive modifiers and by differentiating homonyms, or reinforce a semantic effect, eg by indicating the degree of semantic integration. Meyer found that punctuation marks were less effective in the first capacity than in the second, in other words that there were few cases in the corpus where punctuation alone conveyed the meaning of the sentence. One example is non-restrictive modification.

The correlation between prosody and punctuation was found to be weak and unsystematic. Meyer states that 'there is only a tendency for prosody to correspond to syntax and hence to punctuation'. The opposite tendency emerged in writing where 'Most structural punctuation designates junctures in speech' (69). For instance, writers generally avoid punctuation that would disrupt the prose rhythm if the text were to be read aloud, and adverbials are usually punctuated in the same way as they would be prosodically separated in speech. By and large, Meyer concludes, punctuation is better at reinforcing positions in writing that would be separated by prosody in speech than at creating structure.

The prescriptions of style manuals were generally but not always followed in the Brown Corpus. Among the violations that occurred the

most frequently and with a fairly equal distribution across all the styles, he mentions the omission of punctuation with non-restrictive modifiers. The majority of the rule violations occurred in the fictional style.

Different styles in the corpus were found to have a different distribution of punctuation marks. Learned prose, with long and complex sentences, had for instance comparatively few periods but contained the majority of the colons, semicolons and parentheses, while question marks and exclamation marks occurred most often in fiction.

I read this book with a growing interest, not realizing at first that punctuation could be such a fascinating subject. Not only does Meyer give an excellent overview of punctuation as an integrated system, he also provides useful general rules specifying where punctuation marks would be appropriate in a particular context. Furthermore, his book reflects current usage by presenting the punctuation norms followed in a genuine American corpus.

However, I would like to make two comments. One concerns the correlation between prosody and punctuation, the other the three integrating linguistic areas.

First, the discussion of the correlation between prosody and punctuation is not entirely convincing. Most importantly, no direct comparison is possible, since there is no spoken version to match the written corpus. Meyer gets the prosodic data from other people's studies of British English and his own intuition. Moreover, after some fairly vague statements concerning the general tendencies, he only deals with adverbials.

Second, I miss a discussion of the correlation between syntactic-semantic-prosodic units on the one hand and punctuation on the other. To what extent do such units exist and how do they correlate with punctuation?

Douglas Biber. *Variation across Speech and Writing.* Cambridge: Cambridge University Press, 1988. Reviewed by **Bill Grabe**, Northern Arizona University.

Much has been written in recent years about the relationships among varieties of spoken and written language. For the most part, this research has been plagued by analyses which have used a limited number of texts, a limited number of linguistic features, and/or an unsophisticated research design. As a result, for example, some research has argued that there are more adverbs in writing; some research has argued that there are more adverbs in speech; and other research has argued

that there is little difference in adverb frequencies across genres. Similar puzzles are evident for a number of linguistic features. Biber's book, *Variation across Speech and Writing*, overcomes these sorts of limitations and accounts for earlier contradictory findings by examining textual variation based on a multidimensional model of text structure. In the process of exploring variation across a wide range of text types, he also presents an overall theoretical framework for analyzing textual variation, a framework termed a "multi-feature/multi-dimensional" approach.

The book itself has a four part organization. Part one introduces background concepts and issues. After examining the difficulties encountered in prior research, Biber here provides the rationale for his primary study, the research design used (automated linguistic feature counts and factor analysis), and the theoretical support for his interpretations of the factor analysis dimensions. Part two describes the methodology employed in some detail. He explains how texts and linguistic features were chosen, how scores were created for each linguistic feature, and how factor scores were created for each text genre along each textual dimension. He also provides a succinct summary of the factor analytic procedures used in the research. Part three presents the results of his research, giving a detailed description of the textual dimensions defining text genre variation. It also discusses relations among the various text genres according to each dimension created by the analysis. Part four consists of four appendices. For computational researchers, these appendices may be even more intriguing than the findings of the research. In these appendices Biber describes the algorithms and functions he programmed for the automatic grammatical analysis of the texts, the mean frequency counts of *all* features in each genre, and the Pearson correlation coefficients of all linguistic features. These appendices will allow any knowledgeable reader to explore for him- or herself the results of Biber's research approach as well as examine numerous specific linguistic relations among texts.

Biber's research study examined 481 different texts of 23 different types (17 written genres and 6 spoken genres) including academic prose, general fiction, biographies, press reportage, popular lore, professional letters, face-to-face conversation, broadcasts, and planned speeches. The texts, comprizing approximately 1 million words in all, were taken primarily from the Lancaster-Oslo/Bergen (LOB) Corpus of British English and the London-Lund Corpus of Spoken English. These texts were analyzed by computer to identify and count the relevant linguistic fea-

tures (67 in all). Following the automatic grammatical analysis, the texts were hand-edited for errors. The counts generated by the linguistic feature counting program were then used in a factor analysis (Principal Factor Analysis, promax rotation) to create a seven factor solution. The first six factors are given interpretations as textual dimensions which can define relationships among the various text genres. That is to say, each dimension is given a functional communicative interpretation; the interpretation is based on the strong co-occurrence patterning of linguistic features.

A simplified explanation of Biber's first dimension will illustrate how he analyzes text variation. Biber's first dimension is interpreted as "Involved versus Informational Production". He describes the linguistic features which co-occur on this dimension with positive values as "marking a high level of interaction and personal affect" as well as "a generalized and fragmented presentation of content," and features with negative values as being "highly informational [and showing] almost no concern for interpersonal involvement or affective content." Positive oriented linguistic features include first and second person pronouns, private verbs, emphatics, and WH-questions for involvement; hedges, discourse particles, contractions, non-phrasal *and*, and the pronoun *it* for fragmented content. Negative linguistic features include high frequency of nouns, high frequency of prepositions, few verbs, long words, and a high type/token ratio for non-involvement. Discussions in the research literature provide the motivation for interpreting the co-occurrence of these features in terms of the dimension label given. This approach is repeated for interpreting each of the six dimensions of the factor analysis.

The six textual dimensions are then used by Biber to explore the various ways in which texts will be different from or similar to other texts. Genres are compared along each dimension to create an elaborate multidimensional analysis of textual variation. Considering the basic issue motivating Biber's research, the relationships among spoken and written texts, it was found that no dimension of the six interpreted provided an absolute distinction between oral and written genres. That is, no dimension clearly distinguished all written texts from all spoken texts. Thus, the traditional dichotomy between oral and written texts which is so often assumed does not appear to be a valid dimension of textual variation in any strict interpretation of the results. Biber further argues that any realistic description of textual variation among genres will require comparison along each of the six dimensions that he pro-

poses. This study demonstrates that relationships among texts are complex, reflecting the multidimensional communicative structure of discourse, but they are interpretable.

Biber closes his discussion with two final chapters in part three. The first extends his analysis of variation to exploring differences and similarities among texts *within* genres. In particular, he notes that academic prose as a general genre includes considerable internal variation along all six dimensions, suggesting that each sub-genre (e.g., natural science, medicine, social science) is quite different from the other sub-genres. The second extension of Biber's research approach describes further applications of the multi-feature/multi-dimensional framework. In this last chapter, he explores its use for studies of dialect variation; discourse, stylistic, and historical comparisons; comparisons of stance types; the study of cross-linguistic textual variation, and the development of a theory of text typology in general.

There are many strong features of this book. It gives a nearly exhaustive account of important previous research on spoken and written variation, and it reconciles how researchers could come to very different conclusions about relations among spoken and written genres while seemingly examining the same sorts of texts. Biber's book also offers a strong rationale for the theoretical interpretations of each dimension. It illustrates with numerous examples how abstract quantitative results can be verified by the careful and detailed micro-analysis of specific texts in the corpus; it presents extensive data which other researchers should find useful for their own studies on linguistic aspects of discourse structure; finally, it suggests a way to develop a theoretical framework for an overall theory of text typology.

In summary, this book has much to recommend, and it should be required reading for all researchers interested in linguistic analyses of text corpora. Biber's discussion in the last chapter suggests that interested researchers need not be limited only to linguists but may also include rhetoricians, literary critics, composition researchers, and literacy theorists. Whether all details of Biber's results hold up equally well under further careful scrutiny is a matter for future empirical research. For the moment, he has provided discourse analysts with a powerful new approach to the study of textual variation and, perhaps also, the beginnings of a general theory of text typology.

Shorter notices

Towards an international corpus of English

Sidney Greenbaum, University College London, has recently suggested that we 'should now be thinking of extending the scope for computerized comparative studies in three ways: (1) to sample standard varieties from other countries [than Britain and the United States] where English is the first language, for example Canada and Australia; (2) to sample national varieties from countries where English is an official additional language, for example India and Nigeria; and (3) to include spoken and manuscript English as well as printed English. To facilitate comparative studies, it is essential that such corpora be assembled along parallel lines and in the same period (preferably one calendar year) and also that they be analysed in similar ways.' (*World Englishes* 7, 1988, p. 315.)

As the readers of this journal will know, some work of this kind has already been carried out. Other work is on the way. The Survey of English Usage at University College London is willing to initiate discussions with those interested in the proposal for an international corpus of English.

Survey of machine-readable text corpora

Lita Taylor and Geoffrey Leech, University of Lancaster, have recently carried out a survey of machine-readable text corpora. A descriptive list is available on the ICAME file server (as file SURVEY CORPORA) and can also be ordered from: The Norwegian Computing Centre for the Humanities, P.O. Box 53, Universitetet, N-5027 Bergen, Norway.

The ICAME network server

A network server has been set up at the EARN/BITNET node in Bergen (coordinator: Knut Hofland). The server can be reached from any network that has a gateway to EARN/BITNET like Uninett, Janet, Arpa, Csnet, etc. The server holds information about the material available, some text samples, an ICAME bibliography, programs and documentation, and network addresses. See further *ICAME Journal* 12 (1988), pp. 81-83.

Server

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

Coordinator

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

Material available through ICAME

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

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

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

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

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

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

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

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

LOB Corpus, tagged version, vertical format (available on tape only): Each word is on a separate line, together with its tag, a refer-

ence number, and some additional information (indicating whether the word is part of a heading, a naming expression, a quotation, etc).

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

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

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

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

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

Kolhapur Corpus (tape or diskette): A million-word corpus of printed Indian English texts. See the article by S.V. Shastri, *ICAME Journal* 12, pp. 15-26.

Lancaster Spoken English Corpus (tape or diskette): A corpus of approximately 52,000 words of contemporary spoken British English. The material is available in orthographic and prosodic transcription and in two versions with grammatical tagging (like those for the LOB Corpus). There is an accompanying manual. See further *ICAME Journal* 12, pp. 76-77.

Most of the material has been described in greater detail in previous issues of our journal. Prices and technical specifications are given on

the order forms which accompany the journal. *Note that tagged versions of the Brown Corpus cannot be obtained through ICAME. The same applies to audio tapes for the London-Lund Corpus and the Lancaster Spoken English Corpus.*

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

A manual for the Kolhapur Corpus can be ordered from: S.V. Shastri, Department of English, Shivaji University, Vidyanagar, Kolhapur-416006, India. The price of this manual is US $15 (including airmail charges). Payment should be sent along with the order by cheque or international postal order drawn in favour of The Registrar, Shivaji University, Kolhapur.

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

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

This centre also assists in distributing material. All order forms are sent to Bergen.

Conditions on the use of ICAME corpus material

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

a) collecting and distributing information on (i) English language material available for computer processing; and (ii) linguistic research completed or in progress on this material;
b) compiling an archive of corpora to be located at the University of Bergen, from where copies of the material can be obtained at cost.

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

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

Editorial note

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

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

ISSN 0801-5775