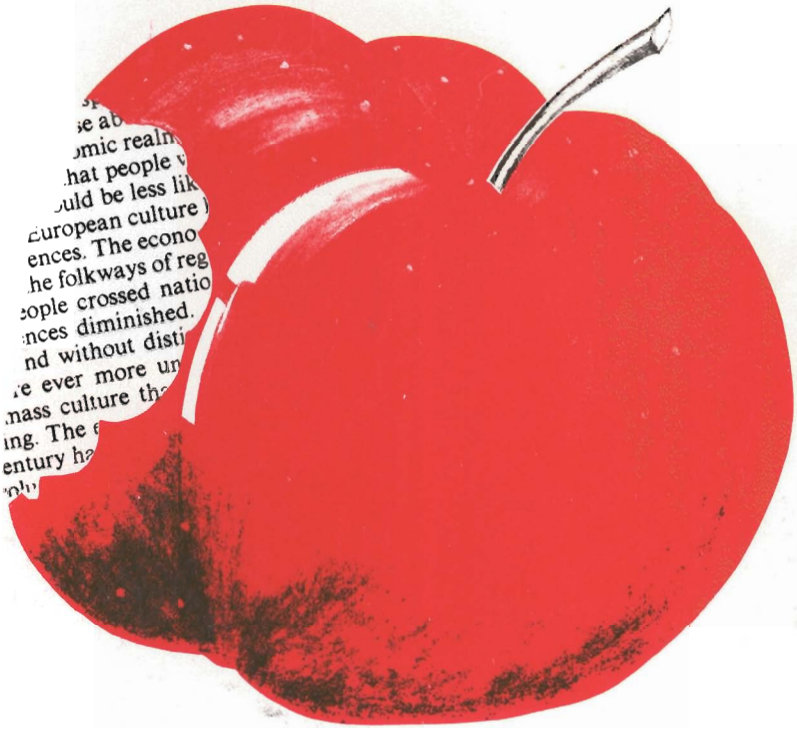
ICAME Journal

International Computer Archive of Modern English

No. 14

April 1990



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, University of Oslo

The Child Language Data Exchange System

Brian MacWhinney
Carnegie Mellon University

Catherine Snow
Harvard University

The CHILDES (Child Language Data Exchange System) project provides an international system for exchanging and analyzing child language transcript data. This system has developed three major tools for child language research: (1) the CHILDES database of transcripts, (2) the CHAT system for transcribing and coding data, and (3) the CLAN programs for analyzing CHAT files. Here we sketch out the current shape of these three major tools and the organizational form of the CHILDES system. A forthcoming book (MacWhinney, in press) documents these tools in detail.[1]

Child language research thrives on naturalistic data – data collected from spontaneous interactions in naturally occurring situations. However, the process of collecting, transcribing, and analyzing naturalistic data is extremely time-consuming and often quite unreliable. To improve this process, the Child Language Data Exchange System (CHILDES) has developed tools that facilitate the sharing of transcript data, increase the reliability of transcription, and automate the process of data analysis. These new tools are bringing about such significant changes in the way in which research is conducted in the child language

field that researchers who deal with naturalistic data will want to understand their nature.

Background

The dream of establishing an archive of child language transcript data has a long history, and there were several individual efforts along such lines early on. For example, Roger Brown's original Adam, Eve, and Sarah transcripts were typed onto stencils and mimeographed in multiple copies. The extra copies have been lent to and analyzed by a wide variety of researchers – some of them attempting to disprove the conclusions drawn from those data by Brown himself! In addition, of course, to the copies lent out or given away for use by other researchers, a master copy – never lent and in principle never marked on – has been retained in Roger Brown's files as the ultimate historical archive.

Such storing and lending of hard copies of transcripts formed an historical precedent for the establishment of a true, comprehensive, international, crosslinguistic child language data archive, but a revolution in the basic conception of such an archive was made possible by the emergence of computers as tools for storage, analysis, and communication. In the traditional model, everyone took his copy of the transcript home, developed his/her own coding scheme, applied it (usually by making pencil markings directly on the transcript), wrote a paper about the results and, if very polite, sent a copy to Roger. The original database remained untouched. The nature of each individual's coding scheme and the relationship among any set of different coding schemes could never be fully plumbed.

The dissemination of transcript data allowed us to see more clearly the limitations involved in our analytic techniques. As we began to compare hand-written and typewritten transcripts, problems in transcription methodology, coding schemes, and cross-investigator reliability became more apparent. But, just as these new problems arose, a major technological opportunity also emerged. As microcomputer word-processing systems became increasingly available, researchers started to enter transcript

data into computer files which could then be easily duplicated, edited, and analyzed by standard data-processing techniques. Computer storage and exchange allow us not only to transcend the limitations of non-computerized analyses, but also to change the basic conception of an "archive." Rather than primarily serving as an historical record, or as a means of short-circuiting the painful, time-consuming process of transcribing for some researchers, a computer archive can become a constantly growing dataset, enriched by every user, since anyone who borrows from the system undertakes at the same time to contribute to the system.

The origin of the CHILDES system can be traced back to the summer of 1981 when Dan Slobin, Willem Levelt, Susan Ervin-Tripp, and Brian MacWhinney discussed the possibility of creating an archive for typed, hand-written, and computerized transcripts to be located at the Max-Planck Institut für Psycholinguistik in Nijmegen. In 1983, the MacArthur Foundation funded meetings of developmental researchers in which Elizabeth Bates, Brian MacWhinney, Catherine Snow and other child language researchers discussed the possibility of soliciting MacArthur funds to support a data exchange system. In January of 1984, the MacArthur Foundation awarded a two-year grant to Carnegie Mellon University for the establishment of the Child Language Data Exchange System with Brian MacWhinney and Catherine Snow as Principal Investigators. These funds provided for the entry of data into the system and for the convening of a meeting of an Advisory Board for the System. The early status of the system is described in MacWhinney and Snow (1985).

The reasons for developing a computerized exchange system for language data are immediately obvious to anyone who has produced or analyzed transcripts. With such a system, we can (1) widen our database, (2) exercise greater scientific precision in coding and transcription, and (3) automate the analysis of large amounts of conversational text. The CHILDES system has addressed each of these possibilities by developing three separate, but integrated, tools. The first tool is the Database itself, the second tool is the CHAT transcription and coding format, and

the third tool is the CLAN package of analysis programs. Let us look at the current status of each of these three tools.

The Database

The first major tool in the CHILDES workbench is the database itself. The importance of the database can perhaps best be understood by considering the dilemma facing a researcher who wishes to test a detailed theoretical prediction on naturalistic samples. Perhaps the researcher wants to examine the interaction between language type and pronoun omission in order to evaluate the claims of parameter-setting models. Gathering new data that are ideal for the testing of a hypothesis may require months or even years of work. However, conducting the analysis on a small and unrepresentative sample may lead to incorrect conclusions. Because childlanguage data are so time-consuming to collect and to process, it is simply not feasible to undertake certain kinds of studies of great potential theoretical interest. For example, studies of individual differences in the process of language acquisition require both an intensive longitudinal analysis and large numbers of subjects – a combination which is practically impossible for a single researcher or a small research team. As a result, conclusions about differences in child language have been based on analysis of as few as two children, and rarely on groups larger than 25. A similar problem arises when linguistic or psycholinguistic theory makes predictions regarding the occurrence and distribution of rare events such as dative passives or certain types of NP-movement. Because of the rarity of such events, large amounts of data must be examined to find out exactly how often they occur in the input and in the child's speech.

In these and other cases, researchers who are trying to focus on theoretical analyses are faced with the dilemma of having to commit their time to basic empirical work. However, there is now a realistic solution to this dilemma. Using the CHILDES database, a researcher can access data from a number of research projects that can be used to test a variety of hypotheses. The CHILDES database includes a wide variety of language samples

from a wide range of ages and situations. Although more than half of the data come from English speakers, there is also a significant component of non-English data. Many of the corpora have been formatted into the CHAT standard and we are now in the process of checking that formatting for syntactic accuracy. The total size of the database is now approximately 140 million characters (140 MB). The corpora are divided into six major directories: English, non-English, narratives, books, language impairments, and second language acquisition.

English Data

The directory of transcripts from normal English-speaking children constitutes about half of the total CHILDES database. The subdirectories are named for the contributors of the data. Except where noted, the data are from American children and are transcribed in CHAT format.

Bates: This subdirectory contains data collected by Elizabeth Bates from videotape recordings of play sessions with a group of 20 children first at 20 months and then at 28 months.

Bernstein-Ratner: These data were collected by Nan Bernstein-Ratner from nine children aged 1-1 to 1-11. There are three samples from each child at three time points, all transcribed from high-quality reel-to-reel audiotapes in UNIBET notation.

Bloom: This subdirectory contains the appendix to Bloom (1970) "One Word at a Time" with language samples from Lois Bloom's daughter Allison between ages 1-4 and 2-10. The subdirectory also contains a large corpus of longitudinal data from Bloom's subject Peter between ages 1-9 and 3-1.

Bohannon: This subdirectory contains transcripts collected by Neil Bohannon from one child aged 2-8 interacting with 17 different adults.

Brown: This subdirectory contains three large longitudinal corpora from Adam, Eve, and Sarah collected by Roger Brown and his students. Adam was studied from 2-3 to 4-10; Eve from 1-6 to 2-3; and Sarah from 2-3 to 5-1.

Clark: This subdirectory contains data from a longitudinal study of a child between age 2-2 and 3-2 by Eve Clark. The transcripts pay close attention to repetitions, hesitations, and retracings.

Evans: This subdirectory contains transcripts contributed by Mary Evans from 16 dyads of first graders at play.

Fawcett: This subdirectory contains data collected by Robin Fawcett from 96 British children aged 6 to 12. The data are accompanied by a full syntactic coding, but are not yet in CHAT format.

Fletcher: This subdirectory contains transcripts from 72 British children ages 3, 5, and 7. The data were collected by Paul Fletcher and are not yet in CHAT format.

Garvey: This subdirectory contains 48 files of dialogues between two children with no experimenter present. Each dyad is taken from a larger triad, so that there are files with A and B, B and C, and C and A from each triad. There are 16 triads in all. The triads range in age from 3-0 to 5-7. The transcriptions are exceptionally rich in situational commentary.

Gathercole: This subdirectory contains cross-sectional data from a total of 16 children divided into four age groups in the period between 2 and 6 years. The children were observed at school while eating lunch with an experimenter present. There is detailed description of actions and situational changes.

Gleason: This subdirectory contains data collected by Jean Berko-Gleason from 24 subjects aged 2-1 to 5-2. The children are recorded in interactions with (1) their mother, (2) their father, and (3) at the dinner table.

Hall: This subdirectory contains extensive data collected by Bill Hall from 38 four-year-olds in a variety of situations. The target children were from four groups: White working class, Black working class, White professional, and Black professional.

Higginson: This subdirectory contains data from 17 hours of early language interactions recorded by Roy Higginson. The children are aged 1-10 to 2-11, 0-11 to 0-11, and 1-3 to 1-9.

Howe: This subdirectory contains data from 16 Scottish mother-child pairs in their homes in Glasgow collected by Christine Howe. The ages of the children are around 1-17.

Korman: This subdirectory contains the speech of British mothers to infants during the first year. The data were contributed by Myron Korman and are not yet formatted in CHAT.

Kuczaj: This subdirectory contains data from a large longitudinal study of Stan Kuczaj's son Abe from 2-4 to 5-0.

MacWhinney: This subdirectory contains data from a longitudinal study of Brian MacWhinney's sons Ross and Mark from 1-2 to 5-0. Data were collected from 5-0 to 9-0, but they are not yet transcribed.

Sachs: This subdirectory contains a longitudinal study of Jacqueline Sachs' daughter Naomi from 1-2 to 4-9. Only partial data are available from 1-2 to 1-8.

Snow: This subdirectory contains a longitudinal study of Catherine Snow's son Nathaniel from 2-5 to 3-9.

Suppes: This subdirectory contains a longitudinal study Patrick Suppes' subject Nina from age 1-11 to 3-3.

VanHouten: This subdirectory contains data from Lori VanHouten comparing adolescent and older mothers and their children at ages 2 and 3.

VanKleeck: This subdirectory contains data from 37 children age 3 to 4 in a laboratory setting contributed by Anne VanKleeck.

Warren-Leubecker: This subdirectory contains data from 20 children interacting either with their mothers or their fathers. One group of children is aged 1-6 to 3-1 and the other group is aged 4-6 to 6-2. The data were contribued by Amye Warren-Leubecker.

Wells: This extensive corpus from Gordon Wells contains 299 files from 32 British children aged 1-6 to 5-0. The samples were recorded by taperecorders that turned on for 90 second intervals and then automatically turned off. The data are not yet in CHAT format.

Non-English Data

With the exception of the data from Afrikaans, Polish, and Tamil, the various non-English data sets have no English glosses or morphemic codings. Therefore, they are currently most useful to researchers who are familiar with the languages involved.

Afrikaans: Jan Vorster of the South African Human Sciences Research Council contributed a large syntactically-coded corpus of data from children between 18 and 42 months learning Afrikaans. The data do not have English glosses, but they are in CHAT format, and, given the extensive syntactic coding, they are well suited for syntactic analysis.

Danish: Kim Plunkett of the University of Aarhus contributed longitudinal data from four children learning Danish. The data are in CHAT format without English glosses.

Dutch: This subdirectory contains a longitudinal study of a single child from Steven Gillis of the University of Antwerp and another longitudinal study by Loekie Elbers and Frank Wijnen of the University of Utrecht. Both corpora are in CHAT.

French: This subdirectory contains a longitudinal study of a single child by Christian Champaud of the CNRS in Paris and another longitudinal study of a single child by Madeleine Leveillé of the CNRS in Paris.

German: This subdirectory contains four corpora. The first is a non-CHAT set of diary notes by Clara and Wilhelm Stern on the development of their three children. The second is a set of transcripts from 13 children between ages 1 and 14 from Klaus Wagner of the University of Dortmund. The third is a set of protocols taken from older children by Jürgen Weissenborn of the Max-Planck Institut in the context of experimental ellicitations of route descriptions. The fourth are transcripts of non-continuous interactions collected by Henning Wode of the University of Kiel from his children in German during a period when they are also learning English. None of the German data are yet in CHAT format.

Hebrew: Ruth Berman of Tel-Aviv University has contributed one longitudinal study of a Hebrew-learning child and cross-sectional transcripts for children from ages 1 to 6. All the data are in CHAT format.

Hungarian: Brian MacWhinney has entered transcripts of four Hungarian children studied for a 10 month period.

Italian: Elena Pizzuto of the CNR in Rome has contributed data in CHAT from a longitudinal study of a single child.

Polish: Richard Weist of SUNY Fredonia has contributed data in CHAT from four children learning Polish. The data are coded morphemically in a way that is very useful for comparative analysis.

Slobin: Dan Slobin of the University of California at Berkeley has contributed data from a comparative study of clausal semantic structures in English, Italian, Serbo-Croatian, and Turkish. Reformatting of the data into CHAT is not yet complete.

Spanish: Jose Linaza of the University of Madrid has contributed data from a longitudinal case study of a child between ages 2 and 4. The data are not yet in CHAT.

Tamil: R. Narasimhan and R. Vaidyanathan of the Tata Institute in Bombay have contributed a longitudinal study of a Tamil child between ages 9 and 33 months.

Narrative Data

The data in this directory are narratives, currently mostly derived from retellings of stories in books and movies.

Gopnik: The files in this directory were contributed by Myrna Gopnik. They are stories elicited by teachers from children between the ages of 2 and 5.

Hicks: The data in this subdirectory were contributed by Deborah Hicks. They were elicited by showing the silent film "The Red Balloon" to children in grades K through 2 and asking them to then tell the story in each of three different genres.

The data are transcribed in CHAT and coded for a variety of anaphoric devices.

Sulzby: The data in this subdirectory were contributed by Elizabeth Sulzby. They contain discussions with children aged 3 and 4 about their favorite books.

Books

The database also includes the complete text of several books and articles. We have obtained permissions from the publishers to include these books in the database. There is also an extensive computerized bibliography of research in child language development.

Carterette and Jones: This subdirectory contains the complete text of "Informal Speech" by Edward Carterette and Margaret Jones. Conversations with first, third, and fifth grade California school children and adults are transcribed both orthographically and in the CHILDES UNIBET phonemic notation. The files were entered from the original computer tape that was used to prepare the book; they are not reformatted into CHAT, but will be in the future.

CHILDES/Bib: With support from CHILDES, Roy Higginson of Iowa State University used a variety of existing resources to compile a rich computerized bibliography of research in child language development that can be searched with the CLAN program called BIBFIND. The status of this independent CHILDES tool is discussed in detail in the accompanying article in this issue by Higginson.

Haggerty: This subdirectory contains the text of an article from 1929 that reports the exact conversation carried on in the length of one day by the author's 31-month-old daughter. The file is not reformatted into CHAT, but will be eventually.

Isaacs: This subdirectory contains the complete text of "Intellectual Growth in Young Children" by Susan Isaacs (1930) and "Social Development in Young Children" by Isaacs (1933). The author records interesting interactions with upper-middle class British children, often in nearly verbatim form.

Weir: This subdirectory contains the phonetic transcriptions from the appendix to "Language in the Crib" by Weir (1970).

Language Impairments

In the next few years we plan to substantially increase the amount of data in the system on language disorders and impairments. Currently, these corpora are available:

CAP: This subdirectory contains transcripts gathered from 60 English, German, and Hungarian aphasics in the Comparative Aphasia Project directed by Elizabeth Bates. The transcripts are in CHAT format and large segments have full morphemic coding and error coding.

Bliss: This subdirectory contains a set of interviews with 7 language-impaired children and their matched normal controls collected by Lynn Bliss at Wayne State University and formatted in CHAT.

Conti-Ramsden: This subdirectory contains transcripts of five British specifically language-impaired preschool children interacting separately with their mothers, their fathers, and a normally developing MLU-matched younger sibling. The data are in CHAT and were contributed by Gina Conti-Ramsden of the University of Manchester. Control transcripts from the sibling interacting with the mother and the father are also included.

Feldman: This subdirectory contains a set of CHAT files collected by Heidi Feldman at Children's Hospital in Pittsburgh from 14 children suffering from various forms of early brain damage. The data are part of an ongoing project entitled "Foundation of Language Assessment" directed by Catherine Snow.

Hargrove: This subdirectory contains a set of interviews in CHAT format between a speech therapist and 6 language-impaired children in the age range of 3 to 6. The files were contributed by Patricia Hargrove of Mankato State University.

Holland: This subdirectory contains a set of interviews with 40 recovering stroke patients who are suffering aphasic symptoms. The data were contributed by Audrey Holland of the University of Pittsburgh and are in CHAT format.

Hooshyar: This subdirectory contains CHAT files collected by Nahid Hooshyar of the Southwest Family Institute from 30 Down Syndrome children between the ages of 4 and 8.

Japanese: This subdirectory contains adult normal Japanese speech error data transcribed in CHAT by Yasushi Terao of Tsukuba University.

Rondal: This subdirectory contains data collected from 21 Down syndrome children in Minnesota by Jean Rondal of the University of Liège. The data have not yet been reformatted into CHAT.

Second Language Acquisition Data

ESF: This subdirectory contains data from the large project on second language learning by immigrant workers directed by Wolfgang Klein at the Max-Planck Institut in Nijmegen. The data are not yet in CHAT format.

Guthrie: This subdirectory contains data in CHAT collected by Larry Guthrie of the Far West Laboratory from three first grade classrooms of immigrant children in San Francisco.

Snow: This subdirectory contains picture descriptions and word definitions in both English and Spanish from 190 Puerto-Rican children in second through sixth grade bilingual classrooms transcribed in minCHAT format. The picture descriptions are coded for explicitness and narrativity. Similar data from an additional 18 fifth graders who are not in bilingual programs, and from 14 third graders who are monolingual Spanish speakers are also included. These data have been contributed by Catherine Snow.

Further information on these corpora can be found in MacWhinney (in press) and in on-line documentation files available with most of the data sets. Researchers can request copies of segments

of the database on either MS-DOS or Macintosh floppies. Copies are sent out free of charge from the Center at Carnegie Mellon and users are asked to return the floppies after copying the data to their hard disk. Copies of data can also be secured from Helmut Feldweg at the Max-Planck Institut für Psycholinguistik in Nijmegen. If members wish to have a complete copy of the database, they need to request data on magnetic tape or in forms compatible with certain specific mass storage devices available for the IBM/XT/AT or the Macintosh.

CHAT

The second major tool in the CHILDES workbench is the CHAT system for transcription and coding. The most conceptually difficult task involved in developing the CHILDES workbench was the creation of the CHAT system. Several years of work with a variety of earlier coding schemes and a great deal of input from our colleagues has led to the formation of the system we call CHAT (Codes for Human Analysis of Transcripts). As discussed in detail in MacWhinney (in press), no coding or transcription system can ever fully satisfy all the needs of all researchers. Nor can any transcription system ever hope to fully capture the richness of interactional behavior. Despite these limitations, the availability of a lingua franca for transcription can facilitate data exchange, data analysis, and the growth of scientific precision.

The CHAT system is designed to function on at least two levels. The simplest form of CHAT is called minCHAT. Use of minCHAT requires a minimum of coding decisions. This type of transcription looks very much like the intuitive types of transcription generally in use in child language and discourse analysis. A fragment of a file in minCHAT looks like this:

```
@Begin
@Participants: ROS Ross Child BRI Brian Father
*ROS: why isn't Mommy coming?
%com: Mother usually picks Ross up around 4 PM.
*BRI: don't worry.
*BRI: she'll be here soon.
*ROS: good.
@END
```

There are several points to note about this fragment. First, all of the characters in this fragments are ASCII characters. The @Begin and @End lines are used to guarantee that the file was not destroyed or shortened during copying between systems. Each line begins with a three-letter speaker code, a colon, and then a tab. Each line has only one utterance. However, if the utterance is longer than one line, it may continue onto the next line. A new utterance must be given a new speaker code. Commentary lines and other coding lines are indicated by the % symbol.

Beyond the level of minCHAT, there are a variety of advanced options that allow the user to attain increasing levels of precision in transcription and coding. Some of the major specifications available in the full CHAT system are:

1. File headers. CHAT specifies a set of 24 standard file headers such as "Age of Child," "Birth of Child," "Participants," "Location," and "Date" that document a variety of facts about the participants and the recording.
2. Word forms. CHAT specifies particular ways of transcribing learner forms, unidentifiable material, and incomplete words. It also provides conventions for standardizing spellings of shortenings, assimilations, interactional markers, colloquial forms, baby talk, and certain dialectal variants.
3. Morphemes. CHAT provides a system for morphemicization of complex words. Without such morphemicization, mean length of utterance is computer based on words, as defined orthographically.
4. Tone Units: CHAT provides a system for marking tone units, pauses, and contours.
5. Terminators: CHAT provides a set of symbols for marking utterance terminations and conversational linkings.
6. Scoping: CHAT uses a scoping convention to indicate stretches of overlaps, metalinguistic reference, retracings, and other complex patterns.

7. Dependent Tiers: CHAT provides definitions for 14 coding tiers. Coding for three dependent tiers have been worked out in detail.
   a. Phonological Coding: CHAT provides a single-character phonemic transcription system for English and several other languages called UNIBET. It also provides an ASCII translation for the extended IPA symbol set called PHONASCII. These systems were devised by George Allen of Purdue University.
   b. Error Coding: CHAT provides a full system for coding speech errors.
   c. Morphemic Coding: CHAT provides a system for morphemic and syntactic coding or interlinear glossing.

The full CHAT system is discussed in MacWhinney (in press).

CLAN

The third major tool in the CHILDES workbench is the CLAN package of analysis programs. The CLAN (Child Language Analysis) programs were written in the C programming language by Leonid Spektor at Carnegie Mellon University. They can be compiled to run under MS-DOS, UNIX, VMS, XENIX, or Macintosh operating systems. The Center at Carnegie Mellon provides members with executable versions of CLAN on floppies and with a manual for the programs. Most users install the programs on a hard disk along with CHAT files either from their own research projects or from the CHILDES database.

In MS-DOS and UNIX, CLAN commands are issued as single line commands to the operating system. For example the command

freq -f *.cha

runs the FREQ program on all the files in a given directory with the ".cha" extension. The "-f" switch indicates that the output of each analysis should be written to a file on the disk. Unless specifically given a file extension name, the FREQ program will figure out names for the new files.

Each of the CLAN programs is started up and run separately. The search programs contains options that allow one to focus the analysis on a particular speaker or a particular coding tier. Most of the programs also allow the user to limit the analysis to a particular numerical range of utterances, such as the first 100 utterances. The most useful CLAN programs are:

Check: This program performs a thorough check for adherence to the syntactic specifications of CHAT. However, the user can short-circuit full error checking in a variety of ways.

ChString: This program replaces specific strings in files with other strings. Although such changes can also be done in most text editors, ChString can effect a whole series of changes on a whole collection of files with a single command. The strings to be changed can be specified in a file that is created by the user.

Combo: This program conducts Boolean searches using a variety of logical operators and wild card symbols. For example, using Combo, one can search for all utterances with a wh-word followed somewhere else in the text by a present tense auxiliary. The user can specify the extent of material to be included in the window around the matching search string.

Freq: This program computes a variety of frequency analyses for the words in a file or corpus. The analyses can be for all the words in a corpus or for only those words matching certain search strings. Search strings can be specified with wildcards in a variety of ways. The shape of words can be varied by changing the nature of the punctuation set. Freq is particulary useful useful in providing data summaries for codes added to a transcript, when options for including coding tiers and excluding text tiers are used. A wide variety of statistics can be obtained with this program as with several of the other search programs.

Gem: This program is designed to allow the user to place important passages into a file for later analysis. Using a text editor, the user marks the passages to be stored. Gem then

uses these marks to determine what should be excised and placed in the "gems" file.

KWAL: This program performs an analysis that is somewhat similar to the key-word-and-line analysis found in concordance packages. However, it is not designed to produce a printed concordance, but rather a record that can be used by a researcher who is interested in testing hypotheses against examples. The program can be used with a file of search strings of words of a certain type, such as all the personal pronouns in the language.

MLT: This program allows the user to define words and turns in a variety of ways to provide means and standard deviations for the mean length of turn.

MLU: This program allows the user to define words and morphemes in a variety of ways to obtain different values for the mean length of utterance. The user can also simply use the standard definition of MLU as a default.

Retrace: This program is useful for tracking the extent to which one speaker repeats, corrects, or expands upon the speech of the previous speaker. The program was written by Jeffrey Sokolov of Carnegie Mellon University.

WdLen: This program tabulates word and utterance lengths and prints a histogram of those lengths. Data can also be output for statistical analysis.

In addition to these general-purpose programs there are also a variety of programs for special needs. Special-purpose programs include:

BackW: This program matches tier line codes with the corresponding main line text.

BibFind: This program finds selected entries in an CHILDES/Bib file. See the article by Higginson in this issue.

CapWd: This program prints all capitalized words. Useful for working with proper nouns.

ClanMan: This program types out documentation on a given CLAN program.

Dist: This program lists average distances between words or codes. This program is particular useful for conducting analyses of chains of anaphoric reference or tense marking chains.

Flo: This program adds a "flow" line to a transcript to represent the conversation without any coding or special symbols as a simple string of words.

KeyMap: This program creates an immediate contingency table for a given key search string.

RevConc: This program creates a reverse concordance. Revconc must be run twice together with one run of the Uniq program.

SaltIn: This program takes file in SALT format and converts it to CHAT format.

Uniq: This program displays unique lines. Is most useful when used with RevConc or Wheel.

Wheel: This program "rolls" through text finding word clusters. If the size of the wheel is set to "3," the program will find all clusters of three words within a given utterance. In its current shape this program can do a simple distributional or cooccurrence analysis.

The CapWd, Freq, MaxWd, Wdlen, and Wheel programs use some of the programming concepts found in programs of the same name developed in the HUM package written by Bill Tuttle for producing concordances. The full CLAN system is discussed in MacWhinney (in press).

The Organization of the System

Administratively, the System has three components: the Advisory Board, the centers, and the members.

The Advisory Board

The first meeting of the Advisory Board for the System was held in Concord, Massachusetts between March 15 and March 18. The meeting was organized by Catherine Snow. The board members present were Elizabeth Bates, Ursula Bellugi, Lois

Bloom, Melissa Bowerman, Robin Chapman, Eve Clark, Jane Edwards, Susan Ervin-Tripp, Paul Fletcher, Willem Levelt, Brian MacWhinney, Jon Miller, Ann Peters, Dan Slobin, and Catherine Snow. At this meeting, the Board sketched out the organization of the system, the shape of the database, and the types of programs to be used. No specific decisions were reached regarding a standard transcription system, although a variety of possibilities were explored. It was agreed that, if funding were available, meetings of the Advisory Board should be held every other year. Unfortunately, because of difficulties in securing funding for such meetings, it was only possible to convene subsets of the Board in 1985 and 1987. However, a full meeting is scheduled for the Fall of 1989 with users of the system who are willing to contribute their time and effort to its development and improvement. In addition to the input provided by the Advisory Board, we solicit suggestions from all researchers regarding modifications to CHAT or CLAN and possible additions to the database.

The Centers

Currently, complete copies of CHAT, CLAN, and the database are located at Carnegie Mellon University in Pittsburgh, Harvard University in Boston, Aarhus University in Denmark, and the Max-Planck Institut für Psycholinguistik in Nijmegen, The Netherlands. The most up-to-date versions of CHAT, CLAN, and the database are those at Carnegie Mellon. The other centers receive updates about twice a year. Further centers can be established wherever there are sufficient computational resources to store and transfer the database. The basic functions and shape of the system are duplicated at each center. The centers keep in correspondence by computer mail, phone, and regular mail for updating of files and task sharing.

The Members

Membership in the System is open. However, members must agree to abide by the rules of the System, not to distribute

copies of programs or files without permission, to abide by the wishes of data contributors, and to properly acknowledge the contributors and the system. Any article that uses the data from a particular corpus must cite a reference from the contributor of that corpus. The exact reference is given in a file called 00readme.doc which is distributed along with each data set. Members are urged to support the progress of the System through contributions of data, programming expertise, or professional advice. It is important for all researchers to understand that further development of the CHILDES tools depends entirely on funding support from government agencies and private foundations. Currently, support for the system comes from the National Institute of Child Health and Human Development at the National Institutes of Health. The best way to argue for such support is to show that the CHILDES tools are being used productively. This means that we need to get feedback from users about articles that have been published using CHILDES data or projects which are underway using the CHAT and CLAN tools.

We attempt to keep researchers informed about the development of the system in a variety of ways. From 1984 to 1987, we mailed out a newsletter that reported on a variety of issues in transcript analysis. Beginning in 1988, we decided that it would be better to use our resources to send out frequent updates of the manuals for CHAT and CLAN. We have also established an electronic mailing list which can be used to discuss issues relating to CHILDES work or other topics in child language development. In 1988, we ran three small workshops at Carnegie Mellon designed to familiarize researchers with the use of CHAT and CLAN. In June 1989, we ran a somewhat larger workshop at Harvard University. Similar workshops are planned for 1990 for Boston and for the International Child Language Association Meetings in Hungary. The 1990 Harvard Workshop will be particularly designed to promote use of CHILDES by researchers working in language disorders. There have also been CHILDES workshops in The Netherlands, Italy, and Denmark. We have also delivered brief presentations of key aspects of the system at child language conferences in Stanford, Austin,

and Boston. We have also placed announcements of the system into seven journals.

The CHILDES system is not for everybody. There are many important parts of child language research that remain outside the scope of the CHILDES system. Comprehension data and data from standardized tests are ignored in our current focus on production data. Moreover, some researchers are asking questions that cannot be addressed with anything but entirely new data. For such researchers, only the CHAT and CLAN tools may be interesting. There are still other researchers for whom none of the CHILDES tools are appropriate. There is, of course, no reason that the CHILDES tools should prove to be equally useful to all researchers. However, the increasing use of CHILDES tools in published research over the last two years indicates the extent to which these tools have begun to play an important role in our field.

Future Directions

Although we have completed a great deal of work in the past six years, there is still an enormous amount to be done. Our plans for the future focus on these goals:

1. We hope to complete the reformatting and checking of the current CHILDES database by 1991. Beginning in 1990, we hope that all new data that are to be added to the database will already by in the correct CHAT format and will have correctly run through the CHECK program.
2. Over the next few years, we expect the database to grow beyond the current focus on first language acquisition by normal children. In the future, the database will include large components of second language acquisition data, adult interactional data, and a variety of data on language disorders. Eventually, we may wish to distinguish between the CHILDES system and a larger Language Data Exchange System (LANDES).
3. During 1990, we will publish the CHAT and CLAN manuals in book format. This volume will also include a description

of the database. The publication will be done in a format that will allow us to issue new editions every one or two years, much as is done for statistical packages such as SPSS or BMDP. Work is currenlty underway to develop options that will have the CLAN programs ouput files in a format useful for analysis by statistical packages such as SPSS, SAS, or SYSTAT. Currently, the CLAN programs MLU, MLT, and FREQ generate a small data ouput file for each transcript file anlayzed. The new options will allow the data from each of the separate analyses to be listed together with a subject and/or session indentification number (read off a header line) in a fixed format data matrix which can be used directly for statistical analyses.

4. During the next few years, we will focus increased attention on the development of a parser-tagger for the semi-automatic analysis of morphological and syntactic structure. A simple version of this system already exists, but much more work will be needed before a full version is ready.
5. We hope to develop a workbench for phonological analysis, probably using the Macintosh computer. Parts of this tool such as digitization, signal analysis, and IPA fonts are already available as off-the-shelf products. We hope to put these tools together in a form that will allow researchers and their students to produce reliable phonological transcriptions which can be analyzed automatically.
6. With the basic tools of CHAT and CLAN, we are working on new ways of assessing language development. Together, these new measures and analyses will provide surer foundations for language assessment.

We encourage other researchers to join us in these goals, to make full use of the current CHILDES tools, and to propose to us new directions and possible improvements to the system. Please address correspondence on CHILDES to Brian MacWhinney, Department of Psychology, Carnegie Mellon University, Pittsburgh PA 15212 USA or send electronic mail to brian@andrew.bitnet or brian@andrew.cmu.edu.

Note

1. Support for the CHILDES system and the preparation of this report was provided by NIH grants HD 23388 and HD 23998.

References

Carterette, E., & Jones, M. H. (1974) *Informal speech.* Berkeley: University of California Press.

Haggerty, L. C. G. (1930) What a two-and-one-half-year-old child said in one day. *Journal of Genetic Psychology*, 38, 75-100.

Isaacs, S. (1930). *Intellectual development in young children.* London: Routledge, Kegan, Paul.

Isaacs, S. (1933). *Social development in young children.* New York: Harcourt, Brace, & Co.

MacWhinney, B. (in press). *Computational tools for language analysis: the CHILDES system.* Hillsdale, N.J.: Lawrence Erlbaum.

MacWhinney, B., & Snow, C. (1985). The Child Language Data Exchange System. *Journal of Child Language*, 12, 271-296.

Weir, R. (1970). *Language in the crib.* The Hague: Mouton.

A New Historical Corpus

Louis T. Milić
Cleveland State University

1. Introduction

The Century of Prose Corpus is a half-million word data-base incorporating prose compositions by 120 authors published between 1680 and 1780. These dates were chosen so as not to exclude authors who belong to the eighteenth century, but who wrote some or all of their works in the earlier period. The purpose of the Corpus is to provide a norm or resource for studies in language or style, for scholars interested in the eighteenth century, in linguistics, stylistics and related fields. The Corpus (hereafter COPC) is in two Parts: A and B.

2. Definitions

To avoid confusion between the two Parts of the COPC, special terms have been chosen to refer to its elements. In Part A, an *author* is one of the 20 major figures explained in 3 below; a *selection* is a group of sentences amounting to 5000 words drawn from one or more of his works; an *author-sample* is a gathering of three *selections* by one *author* totalling 15,000 words. In Part B, a *writing* is a group of sentences amounting to 2000 words drawn from one work of one *writer*; a *genre* is a kind of composition defined by its content or its form or purpose, of which there are ten (see 4 below); a *genre-sample* is a collection of ten *writings* in a single *genre*, consisting of 20,000 words; a *decade* is a unit of ten years within the century (1680-1780) and a *decade-sample* consists of ten *writings* in the different *genres* published during the same *decade* and

comprising 20,000 words. An identifying code is the first element of the *header-block*, which is 27 characters long in Part A and 21 in Part B and followed by a space and the sentence it identifies.

3. Part A

The first Part of the COPC consists of works from 20 major prose authors of the period, those whose styles are often the subject of study, specifically Addison, Berkeley, Boswell, Burke, Chesterfield, Defoe, Dryden, Fielding, Gibbon, Goldsmith, Hume, Johnson, Locke, Richardson, Adam Smith, Smollett, Steele, Sterne, Swift, Walpole. From each author, three selections of 5000 words have been drawn, covering as much of a range of genres or periods of each writer's production as is possible within those limits. Each selection may consist of more than one work: for example, each of Addison's selections consists of three or four separate items. Two of Gibbon's selections, on the other hand, are chapters from different volumes of his *History*. Each of Burke's three selections comes from a different work, with a gap of 18 years between the first and second and 13 years between the second and third. The 20 author-samples total some 300,000 words of expository prose. Although novelists (Fielding, Smollett, Sterne) are included, the selections chosen from their works are not fictional; that is, they are either non-fiction works or expository rather than narrative parts of fictional works (e.g., the critical introductory chapters of Fielding's *Joseph Andrews*). The selections can be studied individually or gathered together in 15,000-word author-samples. When a selection consists of several works, each can be separated from the others, if that is required, by means of the header block.

Selections are seldom complete works or units from complete works, except in the case of periodical essays or chapters from a novel or treatise. When chapters are selected, they may be entire or include part of another following chapter, as the main consideration is the length not the unity of the selection. Generally, selections are chosen by random methods and are

continuous: that is, selections are never made up of random pages.

Each sentence in the COPC is treated separately: that is, each is preceded by an identifying sequence of letters and numbers containing the following pieces of information:

1. An identifying code (e.g., FLD1:JANDR, which signifies: Fielding first selection, *Joseph Andrews*).
2. The publication date of the work.
3. The sequence number of the sentence in the selection.
4. The number of words in the sentence.
5. The paragraph number beginning with the first paragraph of the selection.

Thus the first sentence of the second selection from Gibbon looks like this:

> GIB2:DEC22(1781)001/032-P01 While the Romans languished under the ignominious tyranny of eunuchs and bishops, the praises of Julian were repeated with transport in every part of the empire, except in the palace of Constantius.

The identifying code for the author with the selection number is in this case followed by the code for the *History of the Decline and Fall of the Roman Empire* (DEC) and the chapter number (22). Each sentence begins on a new line; each header block begins at the left margin.

4. Part B

Unlike Part A, which is comprised of the works of the most noted authors, Part B contains one hundred representative writers. These are the writers who constitute the background prose of any period: the journalists, scholars, men of letters, popular fiction writers and others who produce the typical works, by contrast with the Part A authors, whose prose is noteworthy and recommended for study. They are arranged in a ten-by-ten matrix containing one hundred cells. The horizontal rows represent the ten decades of the period; the vertical columns, the

ten genres or subjects from which the writings were taken. These are:

A. Biography

B. Periodicals

D. Education

E. Essays

F. Fiction

G. History

H. Letters/Memoirs

K. Polemics

N. Science

Q. Travel

The contents of each cell were chosen by reference to the *New Cambridge Bibliography of English Literature* (Volume II) and the *Annals of English Literature* (Second Edition), which is conveniently laid out year by year, though not according to category. These two sources and a good deal of trial and error, contingent on availability and the willingness of librarians provided the basis for the list. The header block for a Part B writing consists of the same elements as in Part A, except that the identifying code is different. Thus the fifth sentence of White Kennett's *Complete History of England* (1706) is represented as follows:

> 3G00(1706)0005/026-P0 He did, with a good presence of mind, and competent knowledge of the laws, and a readiness of speech, make a long, but a frivolous defense.

The first digit of the identifying code refers to the third decade of the century (1700-1709). The second digit identifies the writing as a member of the History genre (G). And the "00" means that its accession number is the last, or one-hundredth of Part B. All else is the same as in Part A, except that the paragraph number indicates only that the sentence did not begin

a new paragraph.[1] Each of the 100 writings in this Part contains 2000 words, so the entire Part totals 200,000.

5. Encoding Conventions

For a variety of reasons, the original spelling of the texts has not been reproduced exactly. Spelling and capitalization have been regularized to the American standard, as represented by the Merriam-Webster dictionaries. As a result, the variety of orthography characteristic of eighteenth-century print shops has been reduced and word-indexes and their frequencies are much more reliable. In particular, hyphenation which even today is more a matter of taste than of rule, is less variable than it is in the original texts. Moreover, in the course of the century in question, spelling, capitalization, and hyphenation became more stable, and Gibbon's spelling is nearly modern whereas Dryden's and Temple's is clearly that of an earlier era. Reducing them to a single standard undoubtedly loses some precision for certain kinds of inquiry, but gains a great deal in homogeneity and makes lexical and syntactic analysis much easier.

6. History

As far as I know no "period corpus," or compilation reflecting the usage of a non-modern time, existed before I devised the "Augustan Prose Sample."[2] This latter was created for the purpose of serving as a norm for a study of Addison and Steele's periodical writings and covered (inadequately) the period 1675-1725 with some 52 selections amounting to about 80,000 words.[3] It was the inadequacy of this data-base which led me to construct the present one.

7. Procedure

Although one might desire to construct a data-base in which each year of a certain period is represented by a work written or published during that year, that is not always possible. Some years are empty of the works one is looking for and others

are rich with them. Some compromise may be necessary: one may have to poach on the edge of neighboring periods. More important, there is no necessary relationship between the date of composition and the date of publication. Swift's *History of the Four Last Years of the Queen*, his account of the Tory ministry in which he played a part, was published in 1758, thirteen years after his death, although it was written before 1714. Within limits, the best plan is to accept the publication date as the one that matters, since the language and usage are more likely to reflect the practice of the publication date than of the date of composition, which can seldom be established with any precision. There are several exceptions, however, primarily in the case of manuscript materials published long after the work was written.

There is also the matter of the integrity of the texts. For the authors in Part A, there are scholarly editions for nearly every work. But for the less well-known – more obscure – writers in Part B, it is often necessary to seek out the originals. First editions of such works are not readily lent by the libraries that own them via Inter-Library Loan services. As a result, one must make do with microfilm, microfiche, micro-opaque renditions or request xerographic copies of selected pages, as an alternative to visiting the sites where originals are to be found. The former is inconvenient for one set of reasons, the latter for another. Apart from the expense and delay inherent in visiting libraries in other locations, it is unlikely that good copies can be made of books in poor condition. The alternative is to copy them out in longhand (2000 words at a stretch) or to keyboard them on a laptop computer – probably the best plan since it does not involve transcription more than once, but it does not allow verification or proofreading – or to make an oral recording to be transcribed later, which is not foolproof either. These considerations may help to explain why it has taken nearly ten years to reach this not quite final point on the road to completion. To be specific, the 100 writings in Part B were acquired in the following ways: 14 were keyboarded direct from first editions, 25 from facsimile reproductions, 18 from microfilms, and 8 from xerographic copies – all 65 of

which represent access to original editions. The remaining 35 included 11 from later editions, 13 from edited versions, and 11 from reprints (the least reliable). The libraries in Cleveland (Cleveland State University, Cleveland Public Library and Case Western Reserve University) contributed 32 titles; others in the State of Ohio (Akron, Kent, Ohio State, Ohio University, Youngstown, Oberlin, Bowling Green, Central State, Toledo, Cincinnati, Hiram) a total of 22; the remaining 46 came from Berkeley, Columbia, Indiana, Cambridge, Oxford, Yale, Vermont and a number of others all over the United States; the remaining nine are of unrecorded provenance.

I leave out of account the decisions that have to be made while keyboarding the text: such matters as capitalization – is it necessary? ; how to treat names – with underlines binding its several elements so that Constantius_II appears as a unit and not as two in the word-index; how to manage titles, quotations – leave them out if they are more than one line long ; foreign words – enclose in angle brackets – dialogue – omit – and other extraneous (non-text) material. Because there is no agreed-upon standard, there is in this matter no right way to do things: one must choose a method and then stay with it. Consistency is the main virtue. Nothing is worse than contemplating going back over thousands of words of text because of some unforeseen contingency.

As for data entry, which I have here referred to as keyboarding though there are other methods, my experience may be helpful. Apart from keyboarding, data entry may be achieved through electronic scanning, using the Kurzweil reader or any of its descendants. Scanning is much faster than keyboarding but entails the risk that the scanner will decode a smudge as a letter, and it requires the presence of the person responsible for the text to proofread and correct after the scanning process ends.

Keyboarding can be done oneself or by a surrogate. Each method has disadvantages, but on the whole the compiler is best off doing the work himself. For one thing, he will have to proofread and verify it anyway, so he might as well strive for accuracy during the initial stage. He knows the text, the

conventions, the usages better than any surrogate, and he is responsible for decisions about details of data entry. More important, there is a considerable learning benefit to be derived from doing the keyboarding, allowing the texture of the prose to become thoroughly familiar.

Nine-tenths of the Part B texts (which were produced first) were first entered and deposited on tape in the University's Computer Center via the CMS link to our IBM mainframes (370/158, then 4341, then 3081). From there they were eventually downloaded using PROCOMM to my AT&T 3B1, which functions under the UNIX operating system. The texts were manipulated by means of programs written in SPITBOL 68K. Part B is now complete. Some of the Part A texts were entered on an IBM-PC clone using KEDIT software but mostly on the 3B1 with the UNIX vi text editor. When it is complete (expected date: 31 October 1990) the Corpus will fill twelve 360K diskettes to capacity, but it will also be available on magnetic tape for use on mainframes.

8. Function

Today, when we are overwhelmed with information – subject to "information overload" – largely as a result of technological advances, one ought to ask anyone who produces a large new body of text: what use is it? In the case of the CENTURY OF PROSE CORPUS, the question is not very difficult to answer, but the answer applies to more than just this specimen of a corpus. Essentially a corpus type of database is effective for two kinds of inquiry: 1) to provide examples of certain kinds of structures or contexts characteristic of a certain time/place; 2) to make available comparative data which can define the relation between some variable and the norm. The Corpus is also an anthology of texts representing a vast range of subject matter, a census of viewpoints, attitudes, key words and names, a kind of encyclopaedia without definitions. It is a book exhibiting the spirit of its time and place.

9. Analysis

In speaking of the COPC I refer of course to the machine-readable version available on hard-drive, tape, diskette or other medium. But there is another format in which its characteristics can be displayed. In this the results of an analysis of the text are presented in statistical and tabular form. Apart from the usual rank-order and alphabetical frequency tables of the words in the text, average word-length by decade, by genre and overall will be given, along with sentence-length, type-token and *hapax legomena* ratios for the same measures. Similarly, these data for each writer in Part A and for each selection of each writer can also be provided. Though these are the most common parameters, they are not exhaustive by any means. The only limitation is the size of the volume in which the results are presented. Of course, there is no way to predict the kind of need for data that the user of the COPC may have. After the most obvious facts have been made available, the user must engage the Corpus itself by the ordinary use of special programs or such packages as OCP or Word Cruncher.

10. Conclusion

If the use of this Corpus justifies the amount of effort that has gone into its making, it seems reasonable to hope that other period corpora for British, American and other literatures will be made available. At the moment, most existing corpora are modern and most text data-bases have a specific purpose or were designed for a specific research project.[4] The present work is a general-purpose resource whose usefulness can only be determined by the use that scholars and students make of it.[5]

Notes

1 A complete list of the writers and writings can be found in the Appendix.

2 The current listing of the Oxford Text Archive mentions a body of (English) Civil War polemics, the complete corpus of Old English, Michigan Early English materials, a Collection of 18C. verse (both of restricted availability), but no systematic grouping of British prose of any period. But the Helsinki Corpus will apparently soon offer a body of text earlier than the COPC.

3 Briefly described in *ICAME News*, No.4 (Sept. 1980), 11-12.

4 See the listing of available resources by the Oxford Computing Laboratory: The Oxford Text Archive.

5 Copies may be requested from the compiler at cost. Send for a request form by mail at Department of English, Cleveland State University, Cleveland, Ohio 44118, U.S.A. or by e-mail at R0097@CSUOHIO via Bitnet.

Appendix

List of writers and writings in Part B

Note: Writers' names when known are given even for writings published anonymously; titles are often abbreviated; date is date of original publication; if manuscript source, estimated date of composition is given; identifying code is shown following date.

AIKIN, John, Essay on the Application of Natural History to Poetry (1777-0E63)

ANONYMOUS1, Adventures of Lindamira (1702-3F36)

ANONYMOUS2, An Enquiry whether a General Practice of Virtue Tends to Wealth or Poverty (1725-5K77)

ATTERBURY, Francis, English Advice to the Freeholders of England (1714-4K91)

BARON, S., A Description of the Kingdom of Tonqueen (1700-3Q23)

BATTIE, William, Treatise on Madness (1758-8N18)

BEHN, Aphra (?), The Ten Pleasures of Marriage (1682-1F44)

BELL, John, A Journey from St.Petersburg to Pekin (1719-4Q10)

BLACKMORE, Richard, Essays upon Several Subjects (1716-4E99)

BOWER, Archibald, History of the Popes (1749-7G64)

BOYER, Abel, Memoirs of the Life and Negotiations of Sir William Temple (1714-4A59)

BRADLEY, James, A Letter concerning an apparent motion observed in some of the fixed stars (1748-7N02)

BRADY, Robert, Complete History of England (1685-1G57)

CARTER, Elizabeth, A Series of Letters between Mrs. Elizabeth Carter and Miss Catherine Talbot (1750-8H88)

CHAMBERS, Ephraim, Cyclopaedia (1728-5N03)

CHANDLER, Robert, Ionian Antiquities (1769-9Q96)

CHELSUM, James, Remarks on the Two Last Chapters of Mr. Gibbon's History (1'.78-0K89)

CHETWOOD, William, The Adventures of Captain Richard Falconer (1720-5F38)

CIBBER, Colley, Another Occasional Letter to Mr. Pope (1744-7K01)

CLARKE, John, Essay upon Study (1731-5D19)

CLARKE, Samuel, A Letter to Mr. Dodwell (1706-3K97)

CLOGIE, Alexander, Speculum Episcoporum (c.1679-1A28)

COLLIER, Arthur, Clavis Universalis (1713-4G79)

COSTEKER, John, The Constant Lovers (1731-6F37)

COVENTRY, Francis, Pompey the Little (1752-8F49)

Daily Advertiser (1741-7B62)

DALRYMPLE, John, Memoirs of Great Britain and Ireland (1771-0G33)

ECHARD, Laurence, The Roman History (1695-2G31)

EGMONT, Earl of, Diary (1739-6H85)

ELLIS, Henry, Original Letters Illustrative of English History (1688-1H22)

EVELYN, John, Memoirs for my Grand-son (1704-3H06)

FIELDING, Sarah, The Governess (1749-7F14)

The Free-Thinker (1718-4B92)

GALLY, Henry, A Critical Essay on Characteristic-Writing (1725-5E70)

GARRICK, David, Letters of David Garrick (1771-0H21)

The Gentleman's Journal (1691-2B60)

GILDON, Charles, Miscellaneous Letters and Essays (1694-2E80)

GRANGER, J., Biographical History of England (1769-9A47)

Gray's Inn Journal (1753-8B87)

GREGORY, John, A Father's Legacy to His Daughters (1774-0D12)

GREW, Nehemiah, The Anatomy of Plants (1682-1N11)

HALES, Stephen, Philosophical Experiments (1734-6N07)

HALIFAX, George Savile, Marquis of, Letter to a Dissenter (1687-1K81)

HAMILTON, Alexander, A New Account of the East Indies (1727-5Q24)

HARRIS, James, Hermes (1751-8G27)

HARTLEY, David, Observations on Man (1749-7D46)

HEARNE, Mary, The Lover's Week (1718-4F35)

HOLMES, John, The Art of Rhetoric Made Easy (1739-6D25)

JENNER, Charles, The Placid Man (1770-0F43)

JOHNSON, Richard, Grammatical Commentaries (1706-3D05)

JOHNSTONE, Charles, Chrysal (1761-9F48)

JONES, William, Essay on the First Principles of Natural Philosophy (1762-9N55)

KENNETT, White, Complete History of England (1706-3G00)

KER, John, Memoirs (1726-5H39)

KING, William, A Journey to London (1705-3E65)

LANGHORNE, John, Letters that Passed Between Theodosius and Constantia (1764-9E94)

LAW, Edmund, Life of John Locke (1777-0A72)

LEDIARD, Thomas, Life of John Duke of Marlborough (1736-6A76)

LEWIS, John, Life and Sufferings of John Wicliffe (1720-5A56)

LOCKHART, George, Memoirs concerning the Affairs of Scotland (1714-4H54)

London Gazette (1681-1B61)

London Journal (1723-5B86)

London Magazine (1736-6B67)

LUDLOW, Edmund, Memoirs (1698-2H30)

LUXBOROUGH, Lady, Letters to William Shenstone (1748-7H84)

MASSEY, W., Origin and Progress of Letters (1763-9D15)

MEAD, Richard, A Mechanical Account of Poisons (1702-3N40)

MELMOTH, William, Letters of Sir Thomas Fitzosborne (1742-7E08)

MOLESWORTH, Robert, An Account of Denmark (1694-2Q42)

Monthly Miscellany (1707-3B68)

NEAL, Daniel, History of the Puritans (1732-6G50)

NEWTON, John, The English Academy (1677-1D75)

NORRIS, John, Cursory Reflections on an Essay Concerning Human Understanding (1690-2K95)

NORTH, Roger, Life of Francis North (1742-7A52)

OLDMIXON, John, Critical History of England (1724-5G53)

OSBORN, Sarah Byng, Political and Social Letters (1766-9H93)

PASLEY, Thomas, Private Sea Journals (1778-0Q13)

PERRY, Charles, View of the Levant (1743-7Q58)

PRIESTLEY, Joseph, Essay on the First Principles of Government (1768-9G32)

RAY, John, Philosophical Letters (1718-4N04)

RICHARDSON, J., Notes and Remarks on Milton's Paradise Lost (1734-6E41)

ROWE, Nicholas, Some Account of the Life of Mr. William Shakespear (1709-3A69)

ST.JOHN, Henry, Letters on the Study and Use of History (1752-8E71)

SHERIDAN, Thomas, British Education (1756-8D20)

STEVENS, John, Journal (1689-1Q17)

STRYPE, John, Memorials of Cranmer (1694-2A26)

STUKELEY, William, Memoirs of Sir Isaac Newton's Life (1752-8A29)

TEMPLE, William, Upon the Gardens of Epicurus (1685-1E66)

TILLOTSON, John, Of the Education of Children (1694-2D51)

TROTTER, Catherine, Olinda's Adventures (1693-2F34)

TYSON, Edward, Orang-Outang (1699-2N09)

TYTLER, William, Enquiry into the Evidence Against Mary Queen of Scots (1760-9K98)

Universal Museum (1762-9B73)

WALPOLE, Robert, Speech on a Motion to Repeal the Septennial Bill (1734-6K82)

WARBURTON, William, Letter to the Editor of the Letters on the Spirit of Patriotism (1749-8K78)

WARD, John, Young Mathematician's Guide (1719-4D83)

WATSON, Richard, Essay on the Subjects of Chemistry (1771-0N74)

Westminster Magazine (1776-0B90)

WOOD, Robert, The Ruins of Balbec (1757-8Q16)

WRIGHT, Edward, Observations Made in Travelling Through France and Italy (1730-6Q45)

ICAME 10TH

The 10th International Conference on English Language Research on Computerized Corpora, Bergen, 1-4 June, 1989

The conference was held in Bergen, to mark the tenth anniversary of the first ICAME conference. It was a pleasure for the organising committee (Jostein Hauge, Knut Hofland, Stig Johansson, and Anna-Brita Stenström) to welcome some sixty participants, including a few who were there in 1979: Jan Aarts, W. Nelson Francis, Henry Kučera, Geoffrey Leech, Willem Meijs, and Jan Svartvik. Michael Halliday, who was invited as guest speaker, gave a talk on "Using the results of frequency analysis in a probabilistic grammar of English". More than thirty papers were read on a variety of aspects of corpus work; see the list of papers below, followed by abstracts for most of them. Selected papers will appear in a volume edited by Stig Johansson and Anna-Brita Stenström (Berlin: Mouton de Gruyter).

The packed programme included a number of demonstrations: Micro-OCP, the Oxford Shakespeare, and the OED on CD-ROM (by Sarah Tulloch, Oxford University Press), Correct Grammar (by Henry Kučera, Brown University); the tagged LOB Corpus indexed by WordCruncher and other demonstrations (Knut Hofland, Norwegian Computing Centre for the Humanities); etc. The highlight of the social programme was a boat trip on the fjord with a visit to Lysøen, the small island where the Norwegian composer Ole Bull had his summer residence. There was a

concert with Norwegian folk music, at the end of which Professor Bertil Sundby – distinguished anglicist and gifted musician – was presented with a festschrift on the occasion of his 70th birthday.

In the course of the ten years since the first ICAME conference the circle of corpus workers has grown, and so has the level of sophistication of corpus work. It was encouraging to note the enthusiasm of those involved in the International Corpus of English project (see the contribution by Sidney Greenbaum elsewhere in this journal). We hope to see continued progress at next year's conference in Berlin (June 1990).

A more detailed report from the conference (by Clive Souter, Leeds) is given in the newsletter *Computer Corpora des Englischen in Forschung, Lehre und Anwendungen* (1989), edited by Gerhard Leitner, Freie Universität Berlin.

List of papers

Jan Aarts (Nijmegen): The corpus grammarian's dilemma

Karin Aijmer (Lund): Work in progress within the project Conversational Phrases in English – see abstract

Bengt Altenberg (Lund): Collocations in the London-Lund Corpus – see abstract

Nancy Belmore (Montreal): Tagging BROWN with the LOB tagging suite – see abstract

Magnar Brekke (Bergen): Automatic parsing meets the pragmatic wall

W. Nick Campbell (Winchester): Timing and speech

Mats Eeg-Olofsson (Lund): Collocations in the London-Lund Corpus (computational aspects)

Dorrit Faber (Copenhagen) and Karen M. Lauridsen (Aarhus): The compilation of an English-French-Danish corpus in contract law – see abstract

Margery Fee (Kingston, Canada): The use of the machine-readable versions of the OED in the study of Canadian English – see abstract

Pieter de Haan (Nijmegen): Postmodifying clauses in the English noun phrase – see abstract

Michael Halliday (Killara, Australia): Using the results of frequency analysis in a probabilistic model of grammar – see abstract

Ossi Ihalainen (Helsinki): The grammatical subject in educated and dialectal English – see abstract

Sylvia Janssen (Amsterdam): Enriching syntactically analysed corpora with semantic data – see abstract

Geoffrey Kaye (Winchester): A corpus builder and browser for tagged and bilingual texts – see abstract

Françoise J. Keulen (Nijmegen): Developments and future extensions of the CELEX English database

John Kirk and George Munroe (Belfast): Dialectometry: From corpus lists to analysed maps – see abstract

Gerry Knowles (Lancaster): The presentation of spoken corpora: Prosodic labelling – see abstract

Geoffrey Leech (Lancaster): Running a grammar factory: The compilation of treebanks – see abstract

Gerhard Leitner (Berlin): Report on research with the Kolhapur Corpus of Indian English – see abstract

Magnus Ljung (Stockholm): Evaluating TEFL vocabulary by means of computer – see abstract

Christian Mair (Innsbruck): Quantitative or qualitative corpus analysis? Infinitival complement clauses in the SEU corpus – see abstract

Willem Meijs (Amsterdam): Extracting semantic information from large corpora – see abstract

Dieter Mindt (Berlin): Computational procedures for the comparison of texts

Charles Meyer (Boston): Apposition in a corpus of British and American English – see abstract

Jacques Noël (Liège): Corpora and dictionaries in multiline records – see abstract

Lise Opdahl (Bergen): *-ly* as adverbial suffix: Corpus and elicited material compared – see abstract

Pam Peters (Sydney, Australia): Australian and Canadian English: Some unsettled points of orthography – see abstract

Antoinette Renouf (Birmingham): Progress report on corpus linguistics at Birmingham – see abstract

Matti Rissanen and Merja Kytö (Helsinki): Testing the diachronic part of the Helsinki Corpus

Josef Schmied (Bayreuth): Relative *whose* in the Kolhapur Corpus of Indian English – see abstract

Clive Souter and Tim O'Donoghue (Leeds): The COMMUNAL RAP: A realistic approach to probabilistic parsing – see abstract

Piek Vossen (Amsterdam): Polysemy and vagueness of meaning descriptions as found in LDOCE – see abstract

Anne Wichmann (Lancaster): The prosodic structuring of texts in the Lancaster/IBM Spoken English Corpus – see abstract

Abstracts

Work in progress within the project Conversational Phrases in English

Karin Aijmer
Lund University

Studies of discourse elements have mainly dealt with single words such as *well* and *now* and with a few fixed two-word units (eg *you know*, *sort of*, *I mean*). Pragmatic function is not restricted to these elements, however; even larger stretches of language can have pragmatic functions. Words, phrases, and larger patterns with a discourse function I have called conversational phrases. Conversational phrases can have a social function (eg phrases used for thanking and apologising) or a textual function.

It is not clear, however, how one recognizes a conversational phrase or how the phrases should be analysed. Furthermore no accepted terminology exists for these signals although phrases with a textual function are sometimes referred to as 'gambits' (Keller 1981). I shall use the more neutral term 'structural signal' (cf Vincent 1983). Structural signals are mainly forward-looking and guide the hearer in the process of interpreting and structuring the message.

The aim of my study is to describe the structural signals in the London-Lund Corpus of Spoken English in such a way that

the description can be useful for language teaching and for the compilation of a dictionary of conversational phrases in English.

One cannot define structural signals only from the point of view of what they are doing in the discourse. Structuring signals are elements of variable length and position in the utterance. They may occur with a characteristic prosody and represent certain meanings and grammatical patterns. The phrases below introduce an utterance which the speaker regards as offensive or unpleasant to the hearer. Grammatically and semantically they represent a heterogeneous group however.

let's face it
the last thing I wanted to say was this
to tell you the fact
to be honest
honestly
quite frankly
to tell you frankly
to put it mildly
to tell you the (honest) truth
to tell you the fact
to be fair
to be frank
I don't want to be personal but

Although they can be analysed syntactically the structural signals above should not be defined as ordinary finite or non-finite clauses, adverbs, prepositional phrases etc but as fixed lexical units with special discourse functions.

The notion of fixedness (cohesion, unity) is however problematic. One of the reasons that fixedness is so difficult to deal with is that there is a continuum between completely fixed phrases and 'novel' phrases. Furthermore the criteria for fixedness have to do with syntactic structure, meaning, prosody and function. The criteria themselves may be fuzzy and more or less strong. A grammatical criterion strongly indicating that the phrase is fixed is for instance syntactic irregularity. In *frankly speaking* fixedness is achieved by the irregular word order (cf the regular 'he spoke frankly'). The unity of the phrase is

indicated more weakly if the semantic content is placed in an invariable, recurrent syntactic frame. *Let's face it* consists of a partly fixed syntactic form (*let's*) with a slot for variable lexical content.

The factors which are of interest to analyse the structural signals can be coded and investigated by means of a database program. The following questions will be asked:

1) How long is the phrase? (Cf *actually – what I was going to say was this*)
2) What position in the utterance (initial, medial, final) does it have?
3) Does it have referential meaning? (Cf *well – let me say this*)
4) Does it have metalinguistic content?
5) What discourse function does it have?
6) Is it a separate tone unit?
7) What intonation contour does it have?
8) Is it syntactically detachable?
9) Is it syntactically irregular?
10) Is the semantic content put into a syntactic frame?
11) Does it recur in the same social situation?
12) Can the phrase be extended? (Cf *to tell you the truth – to tell you the honest truth*).

References

Keller, E. 1981. Gambits: Conversational strategy signals. In Coulmas, F. (ed.) *Conversational routine. Explorations in standardized communication situations and prepatterned speech.*

Vincent, D. 1983. Les ponctuants de la langue. Thèse. Université de Montréal. Département d'Anthropologie. Faculté des Arts et des Sciences.

Collocations in the London-Lund Corpus

Bengt Altenberg
Lund University

The project 'Phraseology in Spoken English' at Lund University, supported by the Bank of Sweden Tercentenary Foundation, as for its aim to describe the types and functions of recurrent word combinations in the London-Lund Corpus of Spoken English. The first phase of the project has yielded a working material of just over 200,000 recurrent examples representing 68,000 different combination types. The material is still preliminary in some respects: it is not grammatically tagged and homograph-separated, and it contains a number of phraseologically uninteresting examples (eg fragments like *in the, and a*). Refining the material in these respects will be the main concern of the second phase of the project.

The high frequency of recurrent word combinations in the corpus underlines the 'repetitive' nature of speech: roughly 70% of the running words in the corpus form part of recurrent word combinations of some kind. The combinations vary greatly in length and frequency. Generally speaking, their length is inversely related to their frequency: the great majority are fairly short (2-3 words), while longer ones (5 words or more) are comparatively rare (3% of the material). The most frequent combinations are also chiefly composed of closed-class words, while those containing open-class words are less common.

To test the material and evaluate its usefulness for collocational studies, an analysis was made of combinations containing 'maximizing' intensifiers, ie degree adverbs like *absolutely, completely, entirely, quite, perfectly*. Six features were examined:

(a) the relative frequency of the maximizers (of seventeen examined maximizers, twelve were represented in the material, their frequency ranging from 230 instances of *quite* to two instances of *dead* and *utterly*);

(b) the range and frequency of word combinations containing a maximizer (135 different types of combinations were found, representing 622 examples);

(c) the type of items intensified by maximizers (ranging from 57% adjectives down to 1% determiners);

(d) the range of items collocating with each maximizer (ranging from 45 items with *quite* to one with *dead* (*against*) and one with *utterly* (*powerless*));

(e) collocational restrictions exhibited by different maximizers (eg *completely different* but not *completely difficult*);

(f) collocational preferences revealed by certain intensified items (eg *different* occurring 12 times with *quite*, 6 with *totally*, 5 with *completely* and 2 with *entirely*).

Despite the limited material, the study shows that a corpus of recurrent word combinations can provide interesting information about the use of intensifiers in speech. It can serve to enrich existing descriptions in grammars and dictionaries, provide a basis for comparisons with other varieties, and highlight areas where supplementary corpus or elicitation studies are needed.

Tagging BROWN with the LOB tagging suite

Nancy Belmore
Concordia University

My purpose in using the corpora is somewhat different from that of most other users. The aim is to use them as a research tool in arriving at a word classification system which has a sufficient number of informationally-relevant classes and subclasses to be adequate for natural language understanding systems. The number of such classes is much larger than would normally be considered adequate for grammatical description.

There are numerous ways in which tagged corpora can be used to achieve this aim. In an initial pilot study at the University of Amsterdam, I used the QUERY program to determine the usefulness of pattern extraction from tagged corpora in order to get precise information on the exact circumstances in which problems in defining particular word classes arise and what would be required to resolve them. In another, at the University of Lancaster, I used the original LOB tagging suite, CLAWS1, to tag a subset of sentences exemplifying some of the major problems the QUERY output had revealed. CLAWS1 was used to tag the LOB Corpus available from ICAME.

The two studies demonstrated the value of locating authentic examples of problematic tagging decisions through pattern extraction from tagged corpora. They also showed the value of finding out how different systems tag the same set of problematic examples, or even examples which do not appear to present a problem. In the latter case unanticipated differences in tagging decisions may reveal unexpected insufficiencies in word class definitions.

I therefore decided to undertake tagging the entire BROWN Corpus with CLAWS1. This will make it possible to compare

objectively the differences between the BROWN Corpus tagged by CLAWS1 and the BROWN Corpus as originally tagged at Brown University. My aim is to achieve this, insofar as possible, from my own desktop, using a Macintosh II.

I had previously downloaded, edited and re-formatted a sample file from the tagged BROWN Corpus, chosen because it was the smallest file in the corpus and yet large enough (about 2500 words) to provide adequate test data. I later downloaded the untagged counterpart of this file from the typographically enhanced version of the corpus available from ICAME called Bergen I.

The major task was to edit Bergen I so that it would conform to the expectations of CLAWS1. This requires knowledge of the relations between the Bergen I coding conventions, the conventions for coding the original BROWN Corpus, and the conventions used in preparing the LOB Corpus for input to CLAWS1. It also requires knowledge of the extensive manual editing of the LOB Corpus which occurred after running the PREEDIT program, the first program in CLAWS1. Several different types of editing was necessary:

1. Restoration of information no longer explicitly indicated. For example, in Bergen I a number of compound symbols in the original corpus have been replaced by a single character, sometimes suppressing information, like the difference between a begin-quote and an end-quote, which CLAWS1 expects. In such cases, I had to restore the original BROWN compound symbols.

2. Insertion of meta-symbols which CLAWS1 expects, e.g. a sentence initial marker and a paragraph marker.

3. Prevention of clashes.
   For example, in Bergen a "~" indicates an acronym while to CLAWS it indicates an included sentence. More subtle was the discovery that hyphens in Bergen I must be recorded as dashes when they represent a mark of punctuation, but not when they represent a preposition as in the phrase 'from 1-100'.

4. Reformatting.
   For example, CLAWS1, expects headings and paragraphs to begin on a separate line, but in Bergen I they do not necessarily do so.
5. Performing automatically, *before* running the PREEDIT program, as many as possible of the manual editing tasks which originally followed the PREEDIT program.

Doing the editing required a a word processor with sophisticated search and replace capabilities. I used MindWrite from Delta-Point.

The adequacy of the editing was checked by running PREEDIT with successive versions of the edited sample as input. Most of the required editing was completed before the first test. A few more steps were completed before each of two additional tests.

While I edited the test data on the Macintosh II, a colleague at our Computer Centre, Anne G. Barkman, worked on making the necessary revisions in CLAWS1 so that it would run on a VAX (the original programs were run on an ICL). Only minor changes to PREEDIT were required. The succeeding programs also required changes, most of which occurred because of differences in the EBCDIC and ASCII sorting sequences.

The next-to-last program in CLAWS1 assigns one or more tags, along with the probability that it is the correct tag, to each word. When run over the edited test data, this program reported that of the 2514 input words, 1709 had been unambigously tagged by earlier programs in the tagging suite while 805 were still ambiguous. It resolved the ambiguity of a further 529, assigning more than one tag, together with a probability, to only 276 words.

Before CLAWS1 can be run over the *entire* one-million word BROWN Corpus, there is, of course, further work. Samples from other genres in the untagged corpus need to be edited with MindWrite until there is reasonable certainty that all necessary editing has been identified. After that, a program can be written to carry out the editing with almost no manual intervention.

The most important task remaining is to set up a database in which to record the tagging decisions for each word in the corpus. I have described elsewhere the advantages of the very powerful Fourth Dimension (4D) database management and programming language from ACIUS. A 4D database will facilitate answering two key questions: When is the tag which CLAWS1 designates as the most likely tag different from the original BROWN tag? When it is different, is the original BROWN tag an alternative CLAWS1 tag?

Stig Johansson noted in *The Tagged LOB Corpus Users' Manual* (1986: 26): 'A particular problem has been that we have chosen to draw a borderline and assign a single tag for each occurrence of a word, though we know that gradience and fuzzy borderlines are characteristic of language.'

He also observed (1986: 26): 'While an attempt has been made to find a classification which is linguistically justifiable, this has not always been possible. For one thing, this would have meant tackling grammatical problems which are still awaiting a solution.'

My hope is that by making it possible to compare objectively the results of two different tagging procedures applied to the same corpus, we will have yet another tool to help us state more precisely where the fuzzy borderlines are and to make further progress in solving the many grammatical problems which are still awaiting a solution.

The compilation of an English-French-Danish corpus in contract law

Dorrit Faber
The Copenhagen Business School

Karen M. Lauridsen
The Aarhus School of Business

In 1987 the Danish Research Council for the Humanities decided to promote research activities in LSP and LSP communication. One of the results of this initiative is the creation of three machine-readable corpora in Danish, English and French which are to serve as the empirical basis for a number of language specific as well as contrastive analyses of LSP texts.

The three corpora – each of 1 m running words – were compiled by a group of people working at the Copenhagen and Aarhus Business Schools from 1987 to 1989. The LSP textual universe chosen was that of contract law, with all texts dating from the 10-year period 1978–1987. The bibliographies developed were structured according to a rough text typology covering six text types which were considered relevant to the subject area. The text selection was based on

i) equal representation of each text type

ii) a thematic classification of the subject area with seven main themes

The combination of these two sets of criteria may be claimed to result in bodies of text that are reasonably representative of the textual universe chosen and of legal language usage in

contract law texts. Generally speaking, the length of the texts does not exceed 5,000 words, but since the corpora were only to include complete texts, this limit has not been observed at all times.

The corpora are available as ascii-files to linguists using them for non-commercial linguistic research. The files will be supplied with an introduction to the corpora and a list of the texts included. At the moment the three bibliographies are only on-line at the Copenhagen Business School. The corpora may be obtained through Karen M. Lauridsen, The Aarhus School of Business, Fuglesangs allé 4, DK-8210 Aarhus V.

The use of the machine-readable versions of the OED in the study of Canadian English

Margery Fee
Strathy Language Unit
Queen's University

In producing machine-readable versions of the OED, Oxford University Press has vastly increased the information researchers can find efficiently. The first 12-volume edition (1928) is now available on CD-ROM and the database of the second (1989) edition is available to researchers at the University of Waterloo's Centre for the New Oxford English Dictionary. (These versions were discussed by Stig Johansson in 'The New Oxford English Dictionary Project,' *ICAME Journal* 12 (April 1988)). This preliminary survey was made to discover how much Canadian English can be found in the two versions and how easy it is

to find it. The study focused on 'Canadianisms' in the expectation that problems arising in looking for them would be helpful in discovering how best to extract more general samples of Canadian English, such as the quotations used to support ordinary word senses. The first edition contains relatively few entries labelled 'Canada,' 'Canadian' or 'Canadianism' and the bibliography gives little envidence that many Canadian sources were consulted. However, no OED bibliography is a complete record of works cited in the body of the dictionary, nor does any bibliography give the place of publication in the bibliography entry. This makes it difficult to access the samples of Canadian English in either version in a systematic way. The speed and efficiency of PAT, search software designed at Waterloo, allows one to compensate when searching the OED, however, and to discover that the OED does contain a wide range of Canadian English. Many of the examples examined raise interesting questions for further lexicographical research. The findings of this survey show, not surprisingly, that once the OED is more widely available in machine-readable form in a year or two, it will be a efficient research tool for the study of many varieties of English.

Postmodifying clauses in the English noun phrase - Final report

Pieter de Haan
University of Nijmegen

This paper reports on a project that has recently been completed. The aim of the project was to give a detailed description of

a number of syntactic properties of *postmodifying clauses* (PMCs) in the English noun phrase and to look at the way in which some of these properties are related to each other. The study is based on an examination of corpus data, which implies that only surface structures have been considered.

Following Quirk et al. (1985), I have analysed noun phrases basically in terms of four constituents:

determiner premodifier head postmodifier

The object of study were those noun phrases in which the postmodifier consists of at least one clause.

A comprehensive numerical coding system was designed (see De Haan 1984), accounting for 26 different variables. The numerical codes were stored as a set of computer data and subsequently processed by means of SPSS.

A full account of the project, and a detailed analysis of the data can be found in De Haan (1989). This paper reports on a number of more general results. It is shown that subject noun phrases disfavour PMCs. As is shown in De Haan (1987a), this is due to the fact that more complex structures are avoided in non-final positions in the sentence. The same goes for physically longer structures (see De Haan & Van Hout 1986). As subject noun phrases are usually found in non-final positions in the sentence, it stands to reason that they tend to be as simple as possible.

This conclusion is confirmed by the observation that PMCs usually have relatively simple clause patterns, but that this tendency is stronger in non-final PMCs than in in final PMCs. These tendencies, however, are not found in non-restrictive relative clauses, which are found in considerable numbers in subject noun phrases, and which are significantly longer than the other types of PMCs.

Surprisingly, I found a large number of indefinite noun phrases with PMCs, both restrictive and non-restrictive. Especially for restrictive relative clauses this means that the way these clauses are often said to function (in both pedagogical and descriptive grammars), viz. as defining or identifying the antecedent, needs to be reconsidered. For this cannot cannot be their function in

indefinite noun phrases, as these noun phrases remain unidentified. If they were identified, they would no longer be indefinite. In indefinite noun phrases the function of restrictive relative clauses (or restrictive modifiers in general) is often what I prefer to call classifying, viz. restricting the reference of the head noun to a (member of a) subclass.

These are but a few of the results of the project, which could be characterised partly as a 'fact-finding mission', and partly as an attempt to demonstrate the usefulness of the descriptive model that was used. It has also provided some insight into language performance, and has shown relationships that have hitherto been unnoticed.

References

Aarts, J. & W. Meijs (eds.) 1984. *Corpus linguistics*. Amsterdam: Rodopi.

Aarts, J. & W. Meijs (eds.) 1986. *Corpus linguistics II*. Amsterdam: Rodopi.

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

Haan, P. de 1987a. Exploring the Linguistic Database: Noun phrase complexity and language variation. In Meijs, W. (ed.) (1987). 151-165.

Haan, P. de 1987b. Relative clauses in indefinite noun phrases. *English Studies*, 68: 171-190.

Haan, P. de 1989. *Postmodifying clauses in the English noun phrase: A corpus-based study*. Amsterdam: Rodopi.

Haan, P. de & R. van Hout. 1986 Statistics and corpus analysis. In Aarts, J. & W. Meijs (eds.) (1986). 79-97.

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

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

Using the results of frequency analysis in a probabilistic model of grammar

Michael Halliday
Killara, Australia

Systemic-functional grammar is a paradigmatic grammar which interprets a language as a set of interrelated options (formally a 'system network'). The options that make up any one system are (in my view) inherently probabilistic: for example, the system of 'polarity' consists not simply of the options 'positive/negative' but of these options together with the probability of selecting one or the other. Probability in the system is realized as frequency in the text. Only recently has it been possible to establish grammatical frequencies, through the study of large corpora. Such studies are essential for the further development of grammatical theory; the paper discusses how this information is to be used.

The grammatical subject in educated and dialectal English: Comparing the LLC and the Helsinki Corpus of Modern English Dialects

Ossi Ihalainen
University of Helsinki

This paper discusses the structure of the grammatical subject in two varieties of spoken English: Standard English and regional English. The source of the first type is the London-Lund Corpus and the source of the regional variety is the dialectal part of the Helsinki Corpus. The Helsinki texts are transcriptions of tape recordings of working-class English from the 1970's. The type of dialectal English studied here might be called conservative rural vernacular.

A sample of 6400 words was taken from each corpus. The samples were chosen on the basis of their subject matter to make them 'topically' and stylistically comparable. For instance, both samples have a story about a fire in the kitchen as told informally to a friend. The dialect sample is Somerset English. The LLC sample comes from text category S.4.

The samples yielded more than a thousand grammatical subjects for each variety. The data were analysed into NP subjects, personal pronoun subjects, non-personal pronoun subjects, existential subjects, and relative pronoun subjects. The frequencies are listed below:

Type of subject	Somerset	London-Lund
personal pronoun	687	805
non-personal pronoun	39	60
existential *there*	27	22
noun phrase	107	72
relative pronoun	6	25
subject ellipsis	179	29
Total	1045	1013

ChiSq = 6.58 + 6.79 +
2.53 + 2.61 +
0.18 + 0.19 +
2.85 + 2.95 +
6.03 + 6.22 +
50.99 + 52.60 = 140.50

df = 5

As the figures in the table above show, the most striking difference between the two varieties is ellipsis. Although it turned out that Somerset English allows ellipsis in contexts where standard English does not, there were not strikingly many such contexts in this particular sample, so that they alone cannot account for the great difference between educated English and Somerset English. Rather, the discrepancy is accounted for by the greater frequency of subject ellipsis in Somerset English in general (in particular, in contexts where educated English has prominal subjects, as suggested by the above figures).

NP subjects turned out to be structurally similar in both samples. Also, they were structurally rather simple. Whatever complexity there is in spoken English, it is clearly not located in the subject. This is clearly seen from the fact that although there were 26 subject relative prounouns in the LLC sample, the NP subject was modified only twice by a relative clause.

The greater number of NP subjects in the dialect sample is probably accounted for by the frequent occurrence of family words like *the missus*, or *the wife*, *dad*, etc., or words like *the*

boss, the governor or *the farmer* (usually without an article, as in *Farmer said to me one day, he said...*).

Although there were no striking structural differences between the two samples as far as the internal structure of the NP subjects was concerned, there was a striking syntactic difference. In the dialect sample, there were 12 instances of dislocation of the type *My mother, she was a good mother* and one instances of right dislocation *They go mad over it, bullocks would.* This contrasts sharply with the LLC sample, which showed only one instance of dislocation and even this involved an infinitive, not a noun (*This might be difficult – to climb up there*). Karin Aijmer's figures, based on a much larger sample, show quite clearly that dislocation is rare in the LLC in general, so that from the viewpoint of dislocation the 6400-word sample seems to be quite representative ('Themes and tails: The discourse functions of dislocated elements', in *Nordic Journal of Linguistics*, 12:2, 1989).

Although personal pronouns are used significantly more often as subjects in the LLC sample than they are in the Somerset sample (the figures are 805 and 687 respectively), in the case of the pronoun *they* the figures are reversed: there are 83 instances of *they* in the Somerset sample but only 52 in the LLC sample. This reversal is mainly accounted for by the fact that indefinite *they* is frequently used by dialect speakers to replace a passive construction, as in *They got to skin 'em* instead of *They had to be skinned.* An active construction is particularly common if the verb phrase is complex; that is, although forms like *They were skinned* are quite common, forms like *They had/got to be skinned* are less so, and tend to be replaced by forms like *They had/got to skin 'em.* To show this, all the sentences with *they* as subject that could have been replaced by a passive sentence were counted (i.e. sentences like *Spars they call them*, which could have been replaced by *Spars they're called*; both variants actually occur with the verb *call*). It turned out that none of the 52 sentences with *they* as subject in the LLC text was of this type, whereas 17 sentences in the dialect sample were.

Enriching syntactically analysed corpora with semantic data

Sylvia Janssen
University of Amsterdam

Automatic analysis of computerized corpora has mainly been carried out to reveal the syntactic structure of the input text. Apart from structural data, corpora also contain a wealth of semantic data. Collocations and other contextual data are now retrieved from corpora to support dictionary entries as in the *Collins COBUILD English language dictionary* and to provide statistical data about word cooccurrences. All retrieved data have to be individually inspected to decide upon their precise semantic interpretation, however, since the majority of of words are homonymous and/or polysemous. The aim of the current project is to develop a system which can automatically discriminate senses of words occurring in syntactically analysed corpora. A second goal is the subsequent enhancement of the lexically disambiguated words with refined characterization of verb and argument types. The system makes full use of the semantic data accompanying the entries of a machine-readable dictionary, namely the *Longman dictionary of contemporary English* (LDOCE), a lexicon which has been extensivly studied and used in our department (the ASCOT and LINKS projects).

We shall report on the first phase of the project in which a module is being developed which provides the semantic interpreter with some 'expectations' about likely senses by means of a semantic priming mechanism. The idea behind this approach is that – as in human sentence processing – sense discrimination is guided by both intra- and extrasentential clues within the text. These local and global effects on sense selection are referred to as to as instances of *associative* and *contextual* priming in psycholinguistic literature. A system carrying out

automatic semantic analysis should preferably do so in an 'intelligent' manner, i.e. similar to the way humans process text. To simulate the effect of contextual priming in our system, we decided to carry out a so-called subject field code frequency count to discover the *topic* of a given text – in this example a sample from the tagged LOB Corpus. Although we intend to use a fully analyzed corpus as input to our semantic interpreter (probably the Nijmegen Corpus), the priming module was initially tested out on a tagged corpus sample. Once the topic has been established, the system would then be directed to activate senses of lexically ambiguous words that relate to the established semantic domain and and deactivate senses that could not be related. In this way, an 'expectation-based' environment is created in which likely senses are tried out first by the semantic checker.

The sample we were going to use had to meet two requirements:

1. it should be a domain-restricted test
2. the semantic domain should be well-represented by the lexicon

A text covering a general topic would not contain a substantial amount of jargon words and expressions, resulting in a high type frequency and a low token frequency. In domain-restricted text samples, some subject field code types are expected to occur very frequently, representing the topic domain. The second requirement is inherent to the use of a (learner's) machine-readable dictionary such as LDOCE. The codes with a large number of assignments represent 'popular' topics such as 'FO', (FOod), 'SP' (SPort) and 'LW' (LaW). We therefore selected a sample from the LOB Corpus covering the law domain, viz. a government document on borstal training.

The first frequency count we carried out clearly revealed 'LW' as the main topic scoring a token frequency of 94. Priming 'LW' resulted in a high score for 'PL' (PoLitics), and after priming this code a 'subtopic' was revealed. viz. 'ED'. It was also decided to prime code 'SOZC' (SOciology, subfield criminology) despite its low frequency score because this score

could be directly related to the main code 'LW'. After four subsequent priming sessions, the type frequency was halved from 73 to 37. Of the remaining codes, 27 types scored a frequency of 5 or less and qualified to be deactivated. Although these results were promising, a number of observations were made which will considerably affect a new implementation of the priming module. The planned enhancements involve a statistical analysis of the frequency count results (the results have so far been interpreted manually), the creation of a semantic network of subject field codes so that codes can be related to one another, and the measuring of the relative distances between the token occurrences to detect a possible local 'change of topic'.

References

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

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

A corpus builder and browser for tagged and bilingual texts

Geoffrey Kaye
IBM UK Scientific Centre, Winchester

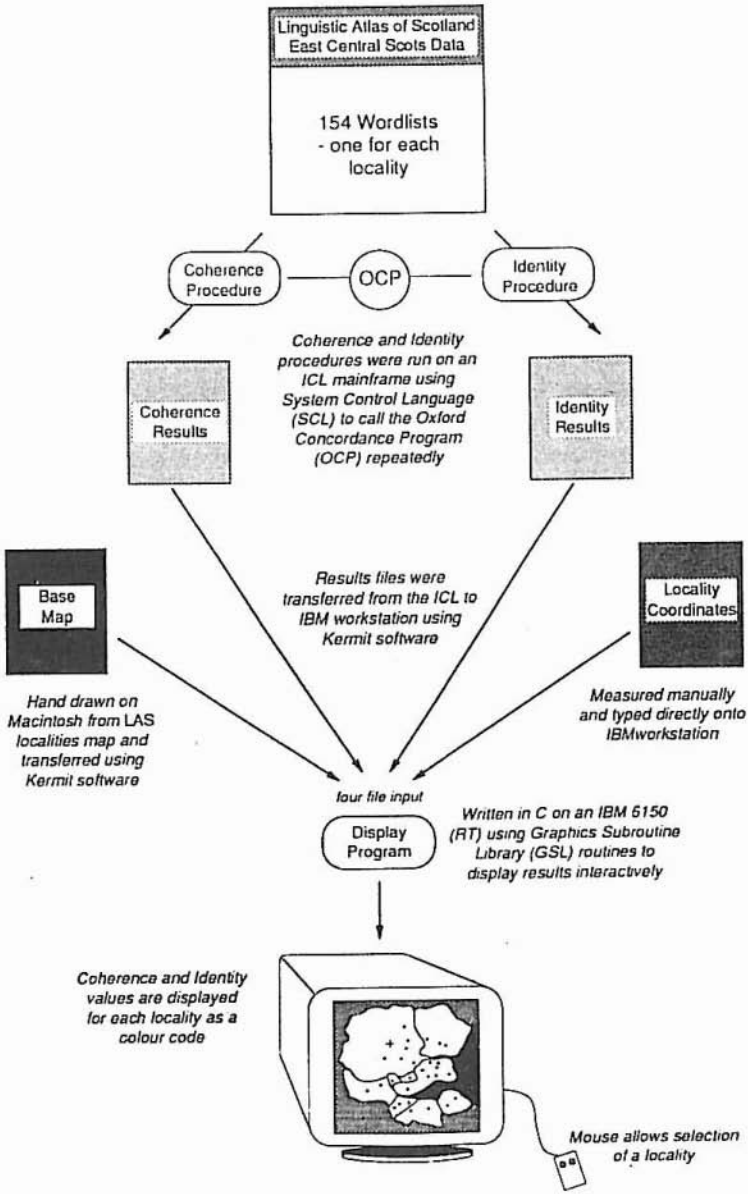
A corpus building system and KWIC (Key Word In Context) concordance browser has been designed for use on an IBM Personal Computer. It provides the facilities for storing the texts of a corpus and building a concordance index to them to give easy and fast access to lexical items in their context. The lexical items are retrieved in real-time and are presented to the user as conventional concordance lines and also as a complete sentence with the lexical search items highlighted. The sentence can be viewed in isolation or in the full context of the text containing it. The system can handle plain texts, transcribed speech with prosodic markings and other features of speech, grammatically tagged texts such as the Brown and LOB Corpora, and bilingual texts such as the glossed Old English texts of the Toronto Corpus. Word frequency indexes are built for all lexical items. Optionally, each language of a bilingual text can be separately concordanced and browsed. When the languages are separately indexed, a word frequency table is produced for each language. Irrespective of how a bilingual text is indexed, when it is browsed the text is displayed showing the primary text and its gloss. The overall objective has been to produce a system in which, after the text has been prepared following the rules for the system, it can be indexed (concordanced) and word frequency counts produced without the need for any manual adjustment of the results. No manual post-editing is needed or possible. The design of the whole system is such as to make it easy and obvious to use by scholars of English who may have little previous knowledge of using a corpus and concordance, and should need only a

basic understanding of the computer they are using. Much to the distress of some users it cannot deal with total ignorance.

Dialectometry: From corpus lists to analysed maps

John M. Kirk and George Munroe
The Queen's University of Belfast

Our aim was to devise a method for doing dialectometry. Information contained within the *Linguistic Atlas of Scotland* was transposed to a computer-based interactive mapping system. A display program enabled an investigation of the area distribution of *coherence* and *identity* values. Our approach further enabled a fresh assessment of the dialect of East Central Scots, and of the status and value of the lexical data gathered by the Linguistic Survey of Scotland. See the figure on p. 67.


Linguistic Atlas of Scotland
East Central Scots Data
154 Wordlists
- one for each
locality
Coherence
Procedure
OCP
Identity
Procedure
Coherence and Identity procedures were run on an ICL mainframe using System Control Language (SCL) to call the Oxford Concordance Program (OCP) repeatedly
Coherence
Results
Identity
Results
Base
Map
Results files were transferred from the ICL to IBM workstation using Kermit software
Locality
Coordinates
Hand drawn on Macintosh from LAS localities map and transferred using Kermit software
Measured manually and typed directly onto IBMworkstation
four file input
Display
Program
Written in C on an IBM 6150 (RT) using Graphics Subroutine Library (GSL) routines to display results interactively
Coherence and Identity values are displayed for each locality as a colour code
Mouse allows selection of a locality

The presentation of spoken corpora: Prosodic labelling

Gerry Knowles
University of Lancaster

The work reported in this paper sought to go behind the tone group boundaries marked in the prosodic transcription of the Lancaster/IBM Spoken English Corpus, and identify the discontinuities involved at the phonetic level. In the overlap passages of texts E01 and G01 – that is, sections transcribed independently by two phoneticians, BW and GK – several different kinds of discontinuity were found, in three main categories:

1. temporal, including pause and pre-final lengthening;
2. pitch, usually a jump up or down after the boundary;
3. segmental e.g. absence of assimilation, elision, and linking-r, or the use of a glottal stop before a vowel.

The order above ranks them according to salience; they are also related so that a temporal discontinuity is usually accompanied by a pitch discontinuity, and a pitch discontinuity by a segmental discontinuity. It was found that where the transcribers disagreed, GK was generally prepared to mark a boundary in response to a less salient discontinuity than BW. G01 was spoken with a faster speech rate than E01, and both transcribers identified less salient discontinuities as tone group boundaries in G01 than in E01.

The problem for corpus linguistics is that there is not one phonetic event that can be identified as a tone group boundary: it is a matter of judgement, in the light of a particular theory of intonation. It would therefore be preferable, when presenting a spoken corpus, to label the discontinuities themselves, and leave the interpretation in terms of tone groups or other prosodic structures to the user of the corpus.

Running a grammar factory: The compilation of treebanks

Geoffrey Leech
University of Lancaster

At Lancaster we have been engaged over the past five years in producing annotated text in the form of *treebanks*, i.e. computer corpora in which each sentence is stored with its (hopefully uncontroversial) phrase structure analysis in the form of labeled bracketing. The purpose of such treebanks is to provide a training corpus for probabilistic parsing, or to provide a testbed for evaluating the coverage and accuracy of grammars. This work has now reached the stage where large quantities of text are being processed fairly rapidly by a team of 'grammarians' – hence the 'grammar factory' of the title.

The Lancaster-Leeds Treebank of c. 45,000 words of the LOB Corpus was the result of detailed manual analysis by a skilled linguist (Geoffrey Sampson).

The LOB Corpus Treebank (c. 140,000 words of the LOB Corpus) was developed by extensive post-editing of the error-prone output of automatic parsing. Its analysis is less detailed than that of the Lancaster-Leeds Treebank, but the corpus is more extensive, in order to provide more adequate statistics for a probabilistic parser. The LOB Corpus Treebank is now undergoing a final stage of checking.

At present effort is concentrated on *skeleton parsing*: an outline phrase structure analysis which can be undertaken by a team of 'grammarians' who need relatively little training on the specific form of the grammar. Altogether, c. 1,500,000 words of text have been annotated, including the Lancaster/IBM Spoken English Corpus. A fast input program, written by Roger Garside, enables analysis to proceed at less than one sentence per minute. The objective is for a very large treebank to be

built up relatively cheaply in terms of time and effort. The labelled bracketing provided is necessarily of a simplified, general purpose type, but the skeleton-parsed sentences can subsequently be annotated in more detail automatically, by a syntax enrichment program.

The three treebanks above represent different techniques aiming roughly at the same goal. The Lancaster-Leeds Treebank was small, but detailed and accurate. The skeleton parsing technique sacrifices detail, and some accuracy, to speed and quantity. Our present opinion is that the last method, that of skeleton parsing, is the most promising way of building up a corpus sufficiently large to provide statistics for probabilistic corpus-based parsing.

Report on research with the Kolhapur Corpus of Indian English

Gerhard Leitner
Freie Universität Berlin

Our research with this corpus has been guided by two questions: (1) in what ways do different varieties of (educated) English differ from each other and (2) how, and if, can the notion of 'new' Englishes be given quantified linguistic substance on the basis of empirical evidence? We have looked at quantitative aspects of the following areas, comparing them with data from the Brown and LOB corpora wherever possible:

1. lexicological aspects
2. complex prepositions
3. the subjunctive

4. *modal verbs*
5. the present perfect and the simple past tense

The table below gives figures for the fifteen most frequent words in the three corpora and shows marked differences in absolute frequencies but only minor ones in their rank order. What is noteworthy, but not illustrated in the table, is that nativized Indian English terms, such as *bandh* 'strike', occur very infrequently and cluster in selected text categories and text types.

Our work is still at an early stage, but it seems that the question of whether the 'new' Englishes are different not only in terms of a presumed norm or a consciousness thereof but in terms of linguistic structure will require a very complex answer. Nativization in core grammatical areas seems to progress very slowly, to say the least, and mainly seems to affect the concept of textual norms. We, therefore, feel we must complement our research later with text-linguistic analyses to see in what ways these aspects are affected and if they can account for the quantitative differences we have observed.

	RANK ORDER			ABSOL. FRQ		
	KOL	LOB	BRW	KOL	LOB	BRW
the	1	1	1	74.584	68.315	69.968
of	2	2	2	39.950	35.716	28.856
and	3	3	3	29.573	27.856	28.856
to	4	4	4	26.496	26.760	26.143
a	6	5	5	21.626	22.744	23.486
in	5	6	6	22.934	21.108	21.341
that	8	7	7	10.099	11.188	10.581
is	7	8	8	11.754	10.987	10.093
was	9	9	9	9.820	10.499	9.808
it	10	10	12	8.812	10.010	8.724
for	11	11	11	8.596	9.299	9.485
he	12	12	10	7.245	8.776	9.540
as	13	13	14	7.072	7.337	7.250
with	14	14	13	6.933	7.197	7.286
be	15	15	17	6.403	7.186	6.373

Evaluating TEFL vocabulary by means of a computer

Magnus Ljung
Stockholm University

Swedish university students of English often have considerable difficulty mastering the vocabulary used in non-fictional British and American texts, like e.g. the English used in non-technical reports or in publications like *The Observer*, *Time* and *Newsweek*.

Since this may be taken to reflect the vocabulary used in the TEFL texts used in the secondary schools, an evaluation was carried out of the vocabulary used in 56 widely used Swedish TEFL text-books.

The books were converted to machine-readable form with a scanner, the result being a corpus of 1.5 M words. This corpus was then compared with the vocabulary in the 18 M-word corpus compiled for the *Collins COBUILD English language dictionary.*

The results yielded by the comparison indicate that the TEFL texts contain a skewed vocabulary, with an exaggerated proportion of simple, concrete terms and too few complex abstract terms.

It was also found that the schoolbooks favoured certain constructions characteristic of spoken language, like contracted forms.

The paper concludes with a discussion of these results and it is argued that TEFL texts at this stage should reflect the composition of large representative corpora like the COBUILD corpus. It is also suggested that this calls for a change of strategy for the compilation of TEFL text-books.

Quantitative or qualitative corpus analysis? – Infinitival complement clauses in the SEU corpus

Christian Mair
University of Innsbruck

On a descriptive level, the paper is a contribution to the study of Modern English infinitival complement clauses, mainly based on the corpus of the Survey of English Usage at University College London. Methodologically, I argue that the quantitative-statistical machine-driven evaluation of text corpora should be complemented by a qualitative-textlinguistic approach, because only the close study of authentic utterances in their original disourse context enables the analyst adequately to describe the syntax-discourse interface with its manifold trade-offs between grammatical well-formedness at *langue*-level and the requirements of functional and communicative efficiency in *parole*.

Extracting semantic information from large corpora

Willem Meijs
University of Amsterdam

1. Background

As techniques and instruments become more refined, we can see a tendency to widen the scope of corpus linguistics from purely categorial and syntactic processing to include semantic and functionai information as well. In the institutional situation at Amsterdam University this is reflected in the projects that have been and are being undertaken. In the ASCOT project (finished in 1988) our sole concern was the extraction and 'streamlining' of categorial and subcategorial information from the *Longman dictionary of contemporary English* (LDOCE), in the LINKS project (to be concluded towards the end of 1989) we have systematized the semantic information from the same source, and in LEXALIZA (begun in 1988, finishing in 1992) we are trying to apply both kinds of information in a functional way to the semantic disambiguation and textual analysis of corpus-material.

In addition to these projects we are also involved in two others on a larger scale, ACQUILEX and DLT, in which semantic and functional aspects play an important part. ACQUILEX (short for 'Acquisition of Lexical Knowledge for Natural Language Processing Systems') is an ESPRIT project commencing August 1989 in which we collaborate with researchers in Cambridge, Dublin and Pisa on the feasibility of the (semi-)automatic extraction of a lexical semantic knowledge base from various monolingual and bilingual machine-readable dictionaries (see Boguraev *et al.* 1989). DLT ('Distributed Language Translation') is a (partly government-subsidized) long-term project undertaken

by the Duch software company BSO, aiming at the development of a largely automated multi-language translation system for use in office-environments, in which a great deal of effort is put into developing methods for determining the choice of semantically adequate translation-equivalents.

Since ASCOT, LINKS and LEXALIZA have been reported on elsewhere (see *inter alia* Akkerman *et al.* 1985, 1988, Vossen *et al.* 1989, and also the abstracts by Vossen and Janssen in this volume), the emphasis here will be on some ideas related to the other two projects. Notice that the account to be given basically represents my personal view of the theoretical issues involved and their possible practical applications, and is hence not to be taken as a description of the projects as such.

2. *Explicit vs implicit semantics*

For the sake of argument we may distinguish between explicit and implicit lexical semantics. An explicit semantic description is a systematic description of the meanings of the words in the language and their interrelations. In theoretical approach this usually takes the form of a description in which the words are reduced to, or decomposed into, a restricted set of abstract semantic elements (features), some of which may have the status of supposedly irreducible semantic primitives. A full-blown account will also include functional information relative to the interaction of the words concerned with syntax: predicate-argument patterns, (case-)roles etc.

An implicit semantics would be one based on the way lexical items combine with the other lexical items of the language in actual language-use, i.e. their collocational patternings in the language in a very wide sense (to be specified in more detail below). By their nature, corpora lend themselves to study of implicit semantics in this sense.

In practice, of course, one form of semantics cannot do without the other: an explicit theoretical semantics must have some basis in empirical observation of language data, and alternatively, since data do not explain themselves, we will need some theoretical preconceptions to know what to look for

if we want to discover the implicit semantics of corpus-material. In other words, the two kinds of semantics should complement each other.

Dictionaries stand halfway between the two kinds of semantics. In a monolingual dictionary the description is in terms of an (ultimately circular) explanation of the meanings of words by means of other words of the language, while example-sentences further illustrate their implicit semantics. In bilingual dictionaries the one-to-many relations which often hold in either direction may help to make some of the implicit semantics in monolingual dictionaries of the languages involved explicit.

LDOCE occupies a special position here for two reasons: the definitions are all expressed by means of a basic vocabulary of some 2200 items and their derivatives, which can thus (with some stretching of the imagination, admittedly) be regarded as a set of 'primitives', and secondly it includes (in the machine-readable version only) a great deal of information of the theoretical explicit kind – labels like Concrete, Human, Abstract etc., semantic-field indications, and information relevant for predicate-argument patterns, case-roles etc. In the ACQUILEX-project we will investigate, among other things, how much of this kind of information can be preserved (or transferred) across different languages.

3. Analysed corpora as a source of implicit semantics

The standard notion of collocations as sequentially adjacent combinations of two or more words will not do to get at the implicit semantics contained in corpus-data. What is needed are syntactically fully-analyzed corpora, in which it is possible to trace the connections between lexical items via significant hierarchically-structured relationships such as Subject-Main Verb, Main Verb-(Head of) Direct Object, Attributive Adjective-Headword Noun, Headword Noun-Postmodifying Prepositional Phrase etc. Given sufficiently large and varied analyzed corpora of this kind, in combination with a dictionary with a descriptive depth comparable to that of LDOCE, it will then be possible,

relative to new combinations encountered in a text, to link up the implicit semantics contained in the corpus-data with the explicit semantics contained in the dictionary.

In order to do this we need computationally simple but powerful (and fast!) software that can scan all of the available corpus material in order to establish which of the registered (structured) combinations are most like a newly-encountered semantically ambiguous combination in an input-text. Thus a sufficiently rich corpus data-base would show that the word *interest* goes with governing verbs like *show, arouse* and *express* in the 'attention' sense, sharing this kind of environment with object-nouns like *feeling, opinion, fear* etc., while in its 'money'-sense it goes with governing words like *pay, charge, levy* etc., in common with other object-nouns like *fee, tax, amount* etc. Similarly, such a data-base would show that *interest* goes with *public, special, intense* etc. as premodifying attributive adjective in the 'attention' sense (in common with *attention, feeling, opinion* etc.), and with *high, annual, accrued* etc. in the 'money' sense (competing with words like *amount, payment, budget* etc.). Given a fast matching procedure it should be possible to show that a newly-encountered combination like *exhibit an acute interest in (sth)* sides with the 'attention' rather than the 'money' patterns of *interest* and thus link it up with the appropriate explicit semantics from the dictionary.

Something of this kind is being done in the present (prototype) DLT system, in which ambiguous word-pairs are matched on-line against similarly structured word-pairs in the so-called 'knowledge-bank', to arrive at the most likely translation. In the next stage of the DLT project the same principle will be applied on a much larger scale on the basis of 3-million word (analysed) translation corpora (cf. Sadler 1989). A similar approach could be applied to semantic disambiguation in monolingual language-analysis. The emergence of parallel processing in recent years appears to provide an ideal context for the massive pattern-matching and statistical computation which such an approach would require.

References

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

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

Boguraev, B., E. Briscoe, N. Calzolari, A. Cater, W. Meijs, E. Picchi & A. Zampolli 1989. *Acquisition of lexical knowledge for natural language processing systems.* ESPRIT Project Description.

Boguraev, B. & E. Briscoe (eds) 1989. *Computational lexicography for natural language processing systems.* London: Longman.

Sadler, V. 1989. *The bilingual knowledge bank: A new conceptual basis for MT.* Utrecht: BSO/Research.

Vossen, P., W.J. Meijs & M. den Broeder 1989. Meaning and structure in dictionary definitions. In Boguraev & Briscoe (1989). 170-192.

Apposition in a corpus of British and American English

Charles Meyer
University of Massachusetts

Apposition is a grammatical category whose structure has been discussed in most scholarly grammars, from Jespersen's *Modern English grammar on historical principles* to Quirk *et al.*'s *Comprehensive grammar of the English language.* But despite

the fact that apposition has been widely discussed, it remains a category that is poorly understood. A casual glance at Jespersen, Quirk *et al.*, or any of the other sources that discuss apposition reveals numerous disagreements about how apposition should be defined and a wide variety of different constructions that are considered appositions.

To arrive at a better understanding, I analysed the appositions in sections of three computer corpora of English: the Brown Corpus, the London-Lund Corpus, and the corpus of the Survey of English Usage. This analysis yielded information such as the following about appositions in English. While appositions can have many syntactic forms, most consist of two units that are noun phrases. Because appositions are syntactically heavy constructions, they tend to have syntactic functions (such as direct object) associated with heavy noun phrases. And since the second unit of an apposition adds new information about the first unit, appositions occur most frequently in press reportage, a genre in which it is necessary for journalists to use appositions (e.g. *the president of the corporation, Susan Smith*) to convey important information to their readers.

References

Jespersen, O. 1909-49. *A Modern English grammar on historical principles.* Copenhagen: Munksgaard.

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

Corpora and dictionaries in multiline records

Jacques Noël
Universite de Liège

Like the CLAN program from Carnegie-Mellon University, MKS Toolkit – a UNIX toolbox designed for MS-DOS machines – offers the linguist-lexicographer the best of both worlds: the well-known advantages of a user-friendly MS-DOS environment, and text processing utilities making full use of the power of UNIX to process lines. In addition, the AWK programming language, which comes as part of the MKS package, permits the creation of multiline records separated by blank lines, which can be processed and queried as ordinary text lines by UNIX, and similar utilities. This paper focuses on the transformation of major corpora (Brown, LOB, etc.) and well-known mono- and bilingual dictionaries into databases consisting of such 'superlines' and on what can be done with these databases. In a concluding section I examine the respective merits of this UNIX approach and of a standard text retrieval package such as WordCruncher. I also examine the use of CLAN in conjunction with UNIX and AWK, and its application to the output of AWK programs.

-ly as adverbial suffix: Corpus and elicited material compared

Lise Opdahl
University of Bergen

This paper presents data from a project investigating factors underlying the use/non-use of the suffix *-ly* with certain verb-modifying adverbs in English. The items studied are 20 adverbial pairs of relatively low frequency, and the approach is mainly quantitative. The aim is to show what kind of tentative conclusions can be made on such a material on the basis of corpus data alone as well as on a juxtaposition of corpus and elicited material. Only a few aspects of the project are presented in this report on work in progress.

The corpus material is based on four different corpora: the LOB and Brown Corpora, the London-Lund Corpus, and a corpus of mainly post-war machine-readable British novels from the Oxford Text Archive.

The corpus study shows a wide span in the number of occurrences of the adverbial pairs investigated. For most adverbs LOB and Brown give very similar results, both as totals and as regards distribution of the forms with and without *-ly* (also called plus- and minus-forms, respectively). However, LOB contains more minus-forms than Brown, a result going counter to the claim by several scholars that the minus-form is more frequently used in American than in British English.

The elicited material was obtained through a questionnaire consisting of some 180 items in judgment test form, given to 100 native speakers of English, 50 British and 50 American, of both sexes and of varying ages and educational backgrounds.

The adverbial pairs *low - lowly* and *direct - directly* were selected for more detailed analysis, and a combination of corpus and elicited data revealed several features concerning the use

of the plus and minus-forms which it would have been difficult to reach on the basis of corpus study alone.[1]

Thus, although the use of a quantitative approach may fail to reveal a great many points, this method is presented as a useful tool for making diagnostic charts and as a sensible starting point for exploratory studies.

The scales of acceptability and frequency in use do not necessarily coincide, but it is reasonable to expect some relation between them, a relationship which should be explored. Thus a strengthening of the link between corpus and elicitation studies is advocated, e.g. through developing guidelines for the sampling of sentences for elicitation from corpora.

Note

1. For details on *direct – directly*, see L. Opdahl, 'Did they purchase it direct or directly?' On *direct* and *directly* as verb modifiers in British and American English. In L.E. Breivik, A. Hille, and S. Johansson (eds.) *Essays on English language in honour of Bertil Sundby*, Studia Anglistica Norvegica 4 (1989), 245-257. Oslo: Novus.

Australian and Canadian English: Some unsettled points of orthography

Pam Peters
Macquarie University

Can we demonstrate the relative influence of American English on Australian and Canadian English, as they currently appear

in print? How resilient in either of them are features of British English? The primary data for this study came from a computerised sample of two mass-circulating magazines – the *Bulletin* (Australia) and *Maclean's* (Canada), both of which resemble the US *Time* magazine. Frequencies from the *Bulletin* and *Maclean's* were compared on a number of items which have been shown to be indexical of the difference between British and American English. The magazine frequencies were also compared with those from the BROWN & LOB corpora, and from the corpus of Australian newspapers at Macquarie University.

On matters of morphology and spelling, both magazines accommodated themselves to standard US practices, and more consistently than other publications at large in Australia and North America. However in its hyphenation the *Bulletin* used consistent and conservatively British practices, whereas *Maclean's* was erratic in its adoption of American practices. Collectively this data showed increasing conformity to American norms where they are clearly standardised (as in morphology and spelling), but americanisation was less evident in hyphenation.

Progess report on corpus linguistics at Birmingham

Antoinette Renouf
University of Birmingham

A recent house move for Cobuild Ltd and the Development Unit has occasioned the transfer of corpus and other data to an AIX-based, IBM PC RT network, and the adaptation of existing software to the new system. Some text-handling en-

hancements have been incorporated at the same time. These include an on-line access facility to ten million words of corpus data, and off-peak access to twenty million words, in each case for the retrieval of concordances both to individual words and to word combinations across a specified span. Searches based on partial string matching have been introduced, and screen facilities have been improved.

The Development Unit has also been concerned with the transfer of data for the School of English to the new University mainframe, an IBM 3090.

At present, the Development Unit holds approximately 38 million words of corpus text on the two systems. It nevertheless continues to collect and process material, and is currently focussing on two areas: spoken language in transcription, and examination scripts for English language and language-based sciences. Work is also in progress on the development of software routines to process the long-planned 'monitor corpus', and simple searches for new word occurrences are already running on 'Times' newspaper data.

Other Development Unit research includes a colloborative project, with Dr Michael Hoey, to develop a system of automatic text abstracting. Manual models already exist for English and other European languages. Professor Sinclair has also begun to work on the parsing of lexical statement in dictionary entries, with particular reference to the mechanisms of rephrasing that occur.

Relative whose in the Kolhapur Corpus of Indian English

Josef Schmied
University of Bayreuth

1. Complex grammatical relations

Computer corpora can be used in three ways in grammatical analysis:

a) They can provide sample sentences for analytical categories, proving that certain constructions are actually used, however unlikely they may appear.

b) A representative corpus of authentic text material can serve as a basis for simple statistical (quantitative) analyses, providing absolute and relative frequencies of comparable forms and structures.

c) Applying a relatively stringent framework to categorize different structures, more complex statistical procedures can be used to show which elements co-occur with which other elements (Sankoff 1988).

Here I want to show how these three ways of analysis can be carried out using tools that are relatively simple and easily available. Relative constructions[1] have been chosen as an example because they frequently exhibit particularly complex relations between numerous variables. My material has been taken from the Kolhapur Corpus of Indian English (see Shastri 1988).

2. Methodology[2]

The well-known Statistical Package for the Social Sciences in the version for Personal Computers (SPSS/PC+) was employed with its coordinated programmes DATA ENTRY for the conve-

nient creation of an SPSS data file and GRAPH for the graphical presentation of the results (with MS-Chart).

In this set-up a data entry sheet is provided to code the values for the grammatical categories analysed, which also shows the variable names, the value entry fields and the value labels with their explanations. Data entry offers the possibility of including 'skip and fill' rules, which skip irrelevant variable fields and fill in unambiguously dependent fields, and cleaning specifications, which reject impossible or unlikely value entries.

3. *Unidimensional results*

Of the many factors that play a role in the complex system of relative constructions more than 30 could be distinguished, some of which have already been mentioned by Randolph Quirk (1957) in his pionering corpus-based article on relative clauses in spoken English. Only a few can be illustrated here, such as the (average) length of *whose* relative clauses. The result obtained is that the mean length is 10 words, which is surprisingly high. Quirk (1957: 108) points out that 'there are no zero clauses of more than nine words, whereas three percent of the *that* clauses and eleven percent of the *which* clauses are no longer than nine words.' In the case of *whose* in the Kolhapur Corpus 40 % are longer (the mode and median are 9 words, the longest relative clause has as many as 35 words), although it must be taken into consideration that Quirk did not count the relativizer.

The relative frequency of the syntactic function of the lexemes modified by the genitive relativizer can also be calculated; this shows clearly that the overwhelming majority (127, i.e. 88%) are the subject of the relative clause, only 12 *whose* being attached to an object and 6 to prepositional phrases. By contrast, an examination of the syntactic function of the relative head in the matrix clause shows, not surprisingly, that only 35 (i.e. 24%) of these are in subject position.

4. Multidimensional results

This example will show how the interrelationship between different variables can be tested statistically, by crosstabulation:

CROSSTABS TABLES = RCLENGTH BY HSYN/STATISTICS = ALL.

The results of this reveal that shorter sentences tend to cluster around the subject side of the diagram, longer ones towards the prepositional or object side.

In order to prove this statistically the following correlation can be calculated:

CORRELATION VARIABLES = RCLENGTH WITH PMSYN HSYN

For this procedure the alphanumeric variables must be recoded as numeric ones; here the positions towards the beginning of the sentence are allocated high numbers and those towards the end (and single words) lower ones. The two variables RCLENGTH and HSYN can be proved to correlate negatively (significantly at the 0.01 level!), demonstrating, in very simple terms, that shorter relative clauses tend to appear after subjects than after objects or prepositional phrases. This can be referred to as the 'light subject constraint'. This empirical result can be supported by arguments from various fields: there is psycholinguistic evidence that it easier to process information if the 'normal' English S(ubject)-V(erb)-O(bject) principle is not distorted too much; in language learning these structures are learnt first; in discourse the principle of theme and rheme suggests that there is less need to modify the noun phrase in subject position, which tends to contain given information, than in later noun phrases, which tend to present new information.

5. Textual aspects

Finally, the fact must be taken into consideration that *whose* relative clauses, like all relative clauses, are unevenly distributed across the text categories: they rarely occur in the press reportage (A) text type or in the miscellaneous category of government publications, etc. (H), which is quite plausible; but they occur very frequently in learned and scientific writings (J), and also

in the press review texts (C). As the Kolhapur Corpus consists of written texts only, there are more occurrences than in spoken corpora, e.g. the Survey of English Usage. More detailed sociocultural interpretations must be left to the Indian specialists.

Notes

1 I usually call the relative particle 'relativizer', because it can be not only a pronoun but also a determiner (*whose*) or an adverbial, the noun in the matrix clause 'relative head', the subordinated clause after and including the relativizer 'relative clause', and the head noun phrase and the relative clause together 'relative construction'; in my abbreviations (in variable names, etc.) I usually use P for the relativizer, because I have R for relative, alongside J for junction, H for head, etc.

2 The technical and methodological aspects of this research project were presented in greater detail at a symposium in Saarbrücken, West Germany, and published in *Workshop 'Computer und Sprache'*, Saarbrücken 1988, 40-47.

References

Quirk, Randolph 1957. Relative clauses in educated spoken English. *English Studies* 38: 97-109.

Sankoff, David 1988. Sociolinguistics and syntactic variation. In F.J. Newmeyer (ed.), *Linguistics: The Cambridge survey*. Vol.IV. Cambridge: CUP. 140-160.

Shastri, S.V. 1988. The Kolhapur Corpus of Indian English and work done on its basis so far. *ICAME Journal* 12: 15-26.

The COMMUNAL RAP: A realistic approach to probabilistic parsing

Clive Souter and Tim F. O'Donoghue
University of Leeds

Parsing techniques for relatively unrestricted English using corpus-based systemic functional grammars are considered. Traditional rule-based approaches are foregone for a more robust probabilistic alternative, the realistic annealing parser. We propose a constrained simulated annealing search method, with evaluation of potential trees according to a probabilistic recursive transition network extracted from analysed corpora.

Polysemy and vagueness of meaning descriptions as found in LDOCE

Piek Vossen
Amsterdam University

The aim of the LINKS project is the building of a database of meaning characterizations based on a computerized dictionary (*Longman dictionary of contemporary English*, or LDOCE) such

that a computer program can have systematic access to the semantic information contained in this dictionary.

The semantic information in a dictionary is stored in the form of expressions in natural language. The meaning of those expressions depends not only on the syntactic/semantic structure but also on the meaning of the words that occur in them. Searching for single words in the meaning descriptions (MDs) that have the same function in the structure, eg. the syntactic/semantic kernel of the MDs, does not necessarily lead to coherent conceptual groups of entry words, eg.:

Entry word	**Meaning description**
corpse	a dead body, esp. of a person.
planet	a large body in space that moves round a star, esp. round the sun.

(Examples from LDOCE, 1978)

To enable a linguist to apply an automatic semantic analysis to a corpus using the LINKS-database, it is a necessity that this problem of polysemy (words having more than one sense) is taken care of. This problem can be viewed in two ways:

- theoretical point of view: there is one (or more) basic sense; other senses are derived from that basic sense as metaphoric (eg. *fox* referring to a 'person') or metonymic (eg. *hand* referring to a 'worker') extensions and can be described by rules. (Aarts and Calbert 1979)
- practical point of view: not all distinctions between senses in the practice of dictionaries are equally defensible. It is a matter of opinion whether a new context in which a word occurs also yields a new sense or meaning of that word. In a number of cases the distinctions are vague, unclear and only the examples given make them acceptable for a human reader. (Pustojevsky 1988)

In LDOCE there are about 24,000 entry words with the part of speech code 'noun'. In all they are represented by about 37,500 MDs. The average number of MDs for each entry word

is therefore 1.5. The distribution of these MDs, however, is rather disproportionate:

c. 16,000 nouns or 67 % have 1 MD
c. 5,000 nouns or 20 % have 2 MDs
c. 1,500 nouns or 6 % have 3 MDs
c. 1,200 nouns or 5 % have 4 upto 27 MDs (average of 7)

The problem seems to be restricted to a comparatively small group of words with a high frequency (in the MDs of the dictionary as well as in ordinary language use). The following strategies are proposed to deal with the problem:

i. Basic senses are likely to be given as the first sense of a word.

ii. The semantic coding system of LDOCE (on tape, not in the book) can be used to group senses that are more related than other senses.

iii. The Longman boxcodes that refer to the usage of the words, eg. codes for 'slang', 'informal', 'technical', 'obsolete' language use, can be used to rule out candidates for the basic sense

iv. Particular forms in MDs such as *this, these* refer to previous senses in the dictionary and in most cases indicate a metonymic or metaphoric relation between the senses. It is possible to distinguish a number of very systematic classes of such relations which could be described rule-wise as in Aarts and Calbert (1979).

v. It is possible to make a distinction between three levels of words in the dictionary. Firstly, words that do not play a role in the meaning descriptions of other words, secondly words that most often are used as kernels of the meaning descriptions and finally the words to which all other meaning descriptions are necessarily related via the meaning descriptions of the previous level. This third level is necessarily circular and represents the most abstract level in the dictionary (containing words such as *thing, place, time*). This is also the level of words which are most polysemous and

whose meaning descriptions are most doubtful, vague and difficult to distinguish. By isolating these words and by separately describing them as atomic concepts the problem of polysemy can be reduced to a large extent.

References

Aarts, J.M.G. & J.P. Calbert 1979. *Metaphor and non-metaphor.* Tübingen: Max Niemeyer Verlag.

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

Pustojevsky, J. & P.G. Anick 1988. On the semantic interpretation of nominals. In *Proceedings of the 12th Conference on Computational Linguistics, 1988, Budapest.* 518-523.

The prosodic structuring of texts in the Lancaster/IBM Spoken English Corpus

Anne Wichmann
Lancaster University

This paper reported further work on prosodic variation, based on the Lancaster/IBM Spoken English Corpus. This particular study was of the distribution and function of the up-arrow in the prosodic transcription.

The theoretical motivation for the study was the theory of declination. Its aim was to investigate the domain of declination by examining instances of apparent 'declination reset', and since the only representation in the corpus of a marked step up in

pitch is the up-arrow (related to but not identical with Crystal's booster), this was examined more closely. The study considers first the phonological distribution of up-arrows, secondly their apparent textual function, and finally the way in which each of these functions is realised phonologically. Some of the findings point towards prosodic indicators of textual organisation, such as the beginnings and ends of paragraphs. Others reveal interesting prosodic phenomena at sentence level, such as the treatment of parenthesis and citations.

Reviews

Bengt Altenberg. *Prosodic patterns in spoken English: Studies in the correlation between prosody and grammar for text-to-speech conversion.* Lund Studies in English 76. Lund University Press, 1987. 229pp. Reviewed by **Gerry Knowles**, University of Lancaster.

This book, arising from the Lund TESS project ('TExt Segmentation for Speech'), provides a very welcome statistical analysis of some of the material in the London-Lund Corpus of Spoken English. The material is presented as a matter-of-fact description based on uncontroversial theoretical positions in both grammar and prosody, and where appropriate, tentative rules are set up. The writer is not attempting to put across his own views, nor to persuade the reader of a particular theoretical approach, but rather to account for the evidence in the text.

Following a brief introductory chapter, chapter 2 attempts to define speech rate in terms of tone units: the mean tone unit contains 4.5 words with a duration of 1.9 seconds at a rate of 2.4 words per second. The key section here (pp25-6) discusses the distribution of different tone unit lengths in a selected text. The modal value is three words per unit, longer units being progressively rarer, with only 2% containing 10 words or more. Earlier in the chapter, the tone unit is discussed rather differently, in terms of 'chunks of information', 'idea units', and real time planning. These two views of the tone unit are incompatible. Are we to assume that the brain imposes a time or word limit on the formulation of ideas? It is safer to assume that the tone unit belongs not to the composition of text, but to the expression of the text, so that it is simply a segment of text that accommodates what information it can within the limit. If so, an 'information chunk' is simply the content of a tone unit, not some pre-existing unit that can be used to explain tone unit division.

Chapter 3 investigates the distribution of the prosodic patterns marked in the transcription, including prominence, the position of onset and nucleus, and nuclear tones. The relation of tone to tone unit length is interesting. The modal value for rises and falls is three words per tone unit, i.e. the same as the overall mode. For level tones it is one, and 69% of level tones occur on groups of one or two words. The status of level nuclei has always been problematical: are groups ending in a level really of the same status as those ending in a rise or a fall? The modal value for compound types (fall+rise, fall+fall, rise+rise) is five: are these really single units?

Chapter 4, by far the longest chapter and about 80 pages long, investigates the way in which grammatical constituents are divided into tone units, using Crystal's (1975:15ff) model. Statistical information is provided on a number of constructions where previously there was an educated guess. Quirk et al. (1985:918), for example, claim that 'conjoins are generally separated by a tone unit boundary in speech': this is confirmed in 98% of cases of clausal coordination (p55), but not in the 'old men and women' case (p108) in which coordinated phrases are modified together. Again, according to Quirk et al. (1985:1258) non-restrictive relative clauses are "usually" marked off with a tone unit boundary, whereas restrictive ones "usually" are not. Altenberg finds (p69) that in his material non-restricted clauses are always (100%) marked off, and so, more often than not (69%), are restrictive clauses. Much of the information in this chapter has a direct and useful application in the development of grammars which attempt to deal with the spoken language.

Chapters 5-7 deal with the distribution of prominence: stress, onset, booster and nucleus (p125). An important point to emerge here is that the traditional distinction between lexical and grammatical words (or open and closed classes) – which is generally believed to be relevant to prosody – does not actually explain much at all. Word classes have to be subdivided not only more finely, but in different ways in different positions. Prominence in general (chapter 5) is related to a subdivision of the major word classes. Onset normally (88%) comes on

the first or second word (chapter 6), but the probability that a word class will take the onset is different from its overall potential for prominence. To predict the position of the nucleus (chapter 7), the familiar 'last lexical item' principle is actually less accurate (78%) than 'last word' (88%); the best results are obtained by dividing word classes into three groups (pp164-5) according to their nucleus potential.

It is entirely to be expected that the close study of a corpus of natural data will lead to the revision of accepted views, and this is the case in chapter 4. Elsewhere the findings cut deeper, and challenge the theoretical basis of the version of the corpus under study. It is therefore worth examining some of the theoretical assumptions on which the book is based.

A general problem of corpus linguistics is to estimate to what extent the analysis of the corpus applies to the language as a whole, or the part of the language which it represents. Using a transcribed spoken corpus, we have the problem twice over: we also need to know to what extent the transcription represents the spoken original. A transcription is not a direct representation of phonetic facts, but a transcriber's interpretation of those facts in the light of a particular theory and using a particular set of categories and symbols. In many cases there are two or more possible interpretations. Tone unit boundary in particular is not always the clearly defined event it might appear from the solid black blocks in transcriptions. A different interpretation of tone units could have a major effect on the figures for tone unit length and the rules for division. A closer look at boundaries before restrictive and non-restrictive phrases and clauses might reveal not only a different probability, but the use of a different type of boundary. A more precise definition of stress – the equation of stress with loudness (p125) is after all not very convincing – would clarify the notion of potential for prominence. Altenberg had of course no practical option but to treat the transcription as a true representation, but this does mean that his figures are strictly an analysis of the transcription, and it is only by a leap of faith that they are deemed to apply to English speech.

Secondly, the argument is based on the assumption that categories set up for syntax are also relevant to prosody. This assumption is even built into the title of the book. The algorithm for tone unit division given in Appendix B inserts tone unit boundaries at constituent boundaries; but in the examples throughout the book only a subset of tone units correspond to constituents. Some are pseudo-constituents, which look like constituents but which in the total sentence context are not, and others do not even look like constituents. In chapter 4, clausal and phrasal coordination are treated separately on syntactic grounds; but judging by the examples, the syntactic level does not appear to be relevant, and the prosody seems to depend on whether the total list can be accommodated in one tone unit. Relative clauses, appositions and postmodifiers are also treated separately; but it is quite possible that they could in some cases come into the same prosodic category. For example, non-restrictive examples could be treated prosodically as special cases of parenthesis, while restrictive examples could follow a general rule for long noun phrases. Chapter 5 is based on the assumption that prominence can be predicted from grammatical categories, if only we can find the right ones. But this is surely not true: *of* and *along* can both be prepositions, but differ in their potential for stress and accent. Prosodic word classes cut across syntactic ones.

The stated aim of this book (p11) is to provide algorithms for text-to-speech research, but in fact, it has a much more general application and is of wider interest than this. Some of the findings, particularly in chapter 4, are of direct value in themselves; others raise interesting theoretical questions with major consequences for the study of spoken corpora. The success of the book is partly due to the fact that it provides useful statistics on non-deterministic areas of speech: these statistics are much more useful to the linguist than the vague feelings of other linguists about what is normal in language. But it is also due to the sensitivity with which Altenberg has consistently dealt with the prosody of English. So much for the conventional wisdom that prosody is so difficult that only a native speaker can handle it!

Finally, there is one respect in which this book is a disappointment: it marks the end of the TESS project. Let us hope that at some time in the not too distant future it will again be possible for the group at Lund to continue their work on English prosody.

References

Crystal, D. (1975) *The English tone of voice. Essays in intonation, prosody and paralanguage.* London: Edward Arnold.

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

Roger Garside, Geoffrey Leech and **Geoffrey Sampson** (eds.). *The computational analysis of English: A corpus-based approach.* London: Longman 1987. Reviewed by **Gunnel Källgren**, University of Stockholm.

The passionate love affair between (some) linguists and computers has matured and developed into a steady relationship, going by the name of computational linguistics. In that process, a major demarcation line from old, pre-computerized linguistics seems to have been carried over; the armchair linguists, with their handful of weird examples, have developed into Lisp machine linguists and get more variation by showing the same few examples in different windows, while those who used to gather excerpts on little cards now have learnt to feed them into computers, whereby they can classify them and count them forwards and backwards in new – and for the most part totally uninteresting – ways.

This somewhat mocking picture of the extreme ends of computational linguistics should, however, not be interpreted as a criticism of the field as such. Computational linguistics is here to stay, not only as a subfield but as an integral part of linguistics in general. Every theoretical linguist must be prepared

to test his/her models on a computer, psycholinguists must use computers for tests and for simulation of hypothesized models, and linguists working with corpus-based research cannot make do with the examples they may find by browsing a couple of books. What would be disturbing, though, is if the dynamic development only continued and corroborated the differing approaches to linguistic research parodied above. Seen in this perspective, the book by Garside, Leech, and Sampson comes as a relief. It demonstrates that large-scale corpus-based linguistics can be a basis for deriving theoretically interesting insights, which in turn generate hypotheses that can best be tested by more large-scale corpus work.

The book describes work that has been carried out by the Unit for Computer Research on the English Language at the University of Lancaster (UCREL). The origins of the work can be traced back to the Brown Corpus of 'present-day American English' that was compiled and published (as a magnetic tape) at Brown University in the early sixties. To serve as a basis for comparison, the Lancaster/Oslo-Bergen (LOB) Corpus of British English was compiled. The sheer existence of these two large corpora, and the openness and generosity with which they have been made available to the rest of the scientific society, have made them invaluable sources for research in (English) linguistics already in their 'raw' format, as running text with no analysis. However, it was soon clear that the value of the corpora would increase considerably if they also contained some analysis, and at both universities, work began with the goal of adding some morpho-syntactic information to the corpora.

To enter analysis into a text corpus is known as 'tagging', i.e. the information – the tag – is connected directly to the word or other unit that it concerns. When it comes to large corpora, this tagging is not something that should be done with a lexicon and a group of hard-working students, it is a challenge that poses many relevant linguistic questions. For the Brown Corpus, the program TAGGIT was developed during the seventies. It needed much human pre- and postediting to reach an acceptable level of accuracy, but that program along with the information that could be retrieved from the tagged Brown

Corpus itself formed the basis for the considerably more sophisticated system CLAWS (Constituent-Likelihood Automatic Word-Tagging System) that was constructed in cooperation between the universities in Lancaster, Oslo, and Bergen, and that reaches an impressive level of 96-97% correctness in its output. The strongest parts of the present book are those that deal with the systems for probabilistic tagging and parsing that have been or are being developed in connection with this.

Chapter 3, written by Roger Garside, describes the methods by which each word is assigned a set of one or more possible tags, depending on its degree of homography. This is done by means of a lexicon of about 7,200 words, where each entry contains all possible tags of that word. Words not found in the lexicon are checked against a list of word-endings, both true derivational suffixes and endings that more haphazardly have come to characterize a specific word-class. Some extra mechanisms for the non-words (such as numerals etc.) that occur so frequently in natural texts, and for handling the suffix *-s*, are also described. If a word is assigned a single tag, that tag is taken to be the correct one, otherwise a procedure for disambiguating homographs is necessary.

Researchers in the field of language analysis know that homography poses large and severe problems; the scale of the problems can be seen from the eight-word sentence

Norman forced her to cut down on smoking

where the ambiguities of the words make it theoretically possible to have about 3,500 different combinations of word-class labels (p.9). Chapter 4, by Ian Marshall, presents the probabilistic methods used – very successfully – for this task. They are based on a transition matrix where the probability of a specific tag following another tag is calculated. It can handle chains of ambiguous words, where the over-all most likely sequence of tags is picked. In this step, all tags except one are eliminated in situations where the context is sufficient to give full certainty. Otherwise, as many tags as possible are deleted and the remaining ones are ranked in order of likelihood. If a manual post-editing

is performed, the editor can choose among the remaining tags, else, the statistically most likely tag is delivered as output.

Chapters 5, by Eric Atwell, and 6, by Roger Garside and Fanny Leech, show how the same principles can be generalized to the analysis of sentence structure, i.e. to parsing. While the assignment of grammatical word-class to single words can be a fairly (though not fully) uncontroversial business, syntactic parsing involves more of theoretical decisions and should preferably involve some underlying model with explicit principles. Syntactic parsing also runs the risk of a combinatorial explosion similar to the one exemplified in the foregoing paragraph, a risk which a superordinate guiding principle for the parsing may help to decrease. A model for large-scale probabilistic parsing called Constituent-Likelihood Grammar is presented by Atwell. It may not look like hard-core theoretical linguistic models usually do, but it still is a model and deserves to be regarded as such. Garside and Leech show its application to the output from the CLAWS system.

A specification of what constructions can be parsed by a certain parser and what structures they are assigned can be called a parsing scheme. In chapter 7, Geoffrey Sampson discusses this important concept and how the parsing scheme used by the UCREL group was reached. It is entirely empirically based and, as Sampson states:

> Thus the scheme as it now stands does represent a consensus rather than one linguist's private view of English grammar, but it is much more detailed than any scheme which is tacitly taken for granted by the profession in general. Indeed, so far as I know our parsing scheme represents the only extant attempt to lay down explicit rules defining a specific grammatical analysis for whatever occurs in modern written English. (p.85)

The scheme itself is, of course, too voluminous to be given in the article, but will hopefully be published elsewhere.

Chapters 8 and 9 deal with aspects of the surface of language that become very important in dealing with unrestricted text. In chapter 8, Barbara Booth describes ways of avoiding manual

pre-processing of computer readable texts. The interpretation of punctuation, case and type shift in letters etc. can be quite as complex as the interpretation of words. The aim of this work is a version of CLAWS, CLAWS2, that can be run on an arbitrary, unrestricted English text with almost no manual pre- or postprocessing and with about as good results as the CLAWS1 system. If this goal could be reached, it would facilitate the work with building large tagged corpora of different types of text and thereby increase the potentials and the standards of both general linguistic and literary research.

Susan Blackwell, in chapter 9, deals with the parsing of idioms (in a wide sense of the word). The idiom-tagging is an important part of CLAWS and actually occurs between the tag assignment and the disambiguation process described above. Idioms and other collocations are receiving increasing interest, not only as obstacles to a smooth parsing but also because of the interesting properties they seem to have in themselves. This is however not treated in the chapter, which describes the technical solutions to the problems caused by idioms in CLAWS.

The last three chapters are not so closely connected with the main work presented in the book, but rather show some practical applications of probabilistic methods. Chapter 10 deals with error detection and analysis in text and chapter 11 with speech synthesis, where a probabilistic parsing of the intended output can give a better assignment of, for example, accent and stress patterns. Chapter 12 describes work on a distributional lexicon, i.e. a kind of 'super-concordance' where information about frequency and cooccurrence of both words and tags can be found. This chapter also deals with the central question of lemmatization, which ought to be considerably simplified in a tagged text. This is actually a peculiar change of direction as compared to manual analysis: the grammatical form and function of a word is decided before its lemma is identified!

Chapter 1 is a general introduction by Geoffrey Leech, while chapter 2, by Geoffrey Sampson, is a vivid defence of the probabilistic 'paradigm' (my choice of word) in contrast to "computational linguists of the cognitive persuasion", as Sampson puts it (p.23). The fierce and well-formulated attack on the

generative linguistic tradition is very refreshing reading, whether one agrees or not. (Personally, I agree quite far.) It is only to be hoped that Sampson's clear joy over the battle will not hide all the truly important points he has to make. It is remarkable that he – apparently without noticing – also brings up the field where 'probabilists' and 'cognitivists' have a chance to meet and join their forces; Sampson gives several references to, and Eric Atwell a short description in chapter 5 of, the technique called simulated annealing, which is a part of the new fashion in linguistics – connectivism, neural nets, parallel distributed processing, stochastics. If this fashion is to become anything more than a handful of catchy words, we need exactly the kind of large-scale, statistics-based and statistics-generating, work that Sampson is such an eager proponent of.

It may thus be that old and sometimes ill-reputed statistical methods can give remarkable results if they are put to use with skill. A renewed interest from the side of automatic translation can already be seen. It may also be that theoretical linguistics could gain new methodological insights from the results of large-scale work with unrestricted language. In general, I would regard this work as a milestone in its field, and it is my hope that other scholars as well will find it useful in evoking new thoughts and ideas.

Pieter de Haan. *Postmodifying clauses in the English noun-phrase. A corpus-based study.* Amsterdam: Rodopi, 1989. Reviewed by **Josef Schmied**, University of Bayreuth.

Although relative constructions have attracted the attention of linguists for a long time and many assumptions about frequency and occurrences can be found in the relevant literature, this study breaks new ground in several respects:

1) It confirms most previously held assumptions empirically, because it takes data from a full analysis of long passages of running text.

2) It uses a framework of relevant variables systematically with the help of the well-known Statistical Package for the Social Sciences (SPSS).

3) It is more comprehensive than previous descriptions in including other postmodifying clauses than the usual finite relative ones: appositive and discontinuous modification, restrictive and non-restrictive clauses and even postmodifying clauses introduced by subordinators (such as *as*), be they finite, non-finite or verbless.

Although the book has nine chapters it can basically be divided into two parts: a general qualitative descriptive part is complemented by a quantitative statistical part in which almost all the 55 tables are to be found. This division into a qualitative and quantitative section allows readers who are not quantitatively minded to use only the first, descriptive part, but obtaining a comprehensive qualitative and quantitative picture of the results of a particular topic proves somewhat difficult, especially as the corresponding sections of the two parts are not easy to find. This difficulty is overcome, however, by a detailed and informative subject index as well as an author index.

The study is based on the Nijmegen Corpus of approximately 130,000 words of running text, about 120,000 of which are printed. This yielded 2,430 examples of postmodifying clauses. These were all classified according to 22 variables which were coded into an SPSS system file and then analysed as to their frequency in absolute terms and in relative ones depending on other parameters.

This methodology yields a wealth of information. In most cases the author lets the figures speak for themselves and only comments in detail on the most significant results. He takes great care in calculating and proving statistically which cell values are significant on the 0.05 level (p.51f). Thus indirect objects are not realized by noun phrases with postmodifying clauses at all, while subjects are significantly infrequently realized by these structures (p.117). He interprets these results carefully and offers psycholinguistic reasons for the first and the possibility of prepositional alternatives for the latter finding.

Another interesting result is, for instance, that relative appositive clauses (in other descriptions subsumed under content clauses) have a significantly high share of infinitive clauses, but a significantly low one of the finite clauses analysed. The author pays special attention to determiners and predeterminers of the head of the noun phrase and thus discovers that restrictive relative clauses are found in indefinite noun phrases in a fair number of cases. Whereas this result (partly) refutes the general view on postmodifying clauses that their function is to identify the antecedent, most other views are confirmed, such as the hypothesis that postmodifying clauses are significantly longer and more frequent in non-fiction than in fiction texts. These examples are only a brief indication of the attention to detail which makes this contribution to a central chapter of English grammar so valuable.

The study is primarily a description of the corpus material and is not intended as a theoretical contribution to any linguistic school. This is reflected in the bibliography and the author index, which shows that contributions from the traditional philological as well as from the tranformational school are used more to discuss previous approches (mainly in Section 1.3.) than to interpret the quantitative research results, the unique detail provided in this study making comparison with previous results difficult.

De Haan's study with its clear methodological framework, its extensive tables, its detailed interpretation of the results, its subject index and its comprehensive list of variable labels and values, syntactic tests and texts in appendices can serve as a model for future quantitative corpus-based analyses.

Shorter notices

The International Corpus of English

Sidney Greenbaum
University College London

It is less than eighteen months since I published a proposal for an international computerized corpus of English in *World English* (1988, 7.315). Evidently the time was ripe, since my call has been eagerly answered from all parts of the globe. Plans are under way for the establishment of 12 regional corpora (covering 14 countries) and of three specialized corpora. In all, scholars in 16 countries have expressed their willingness to participate in this international project.

The regional corpora include countries where English is a first language or a second language functioning for internal communication. Research teams are being organized for the regional corpora in Australia, Canada, East Africa (Kenya, Tanzania, Zambia), Hong Kong, India, Jamaica, New Zealand, Nigeria, Philiphines, Singapore, UK, and USA. The three specialized corpora will consist of samples of written translations into English, spoken interchanges in English between speakers from different countries, and texts used in the teaching of English as a foreign language.

Each regional corpus will contain at least a core corpus of 1 million words that can be used for comparative studies of the varieties of English in different countries. The core corpora

will therefore be compiled along parallel lines. In addition, some teams may compile special-purpose corpora (e.g. newspapers, business letters) and open-ended monitor corpora.

Each of the research teams will be responsible for obtaining funding for its own component of the international project and for assembling the material, computerizing it, and analyzing it. Access to adequate funding may delay the processing of the data in some instances, but I expect that all teams will be able to undertake the initial task of assembling the language samples for the specified period. The various stages in the processing of the data can be implemented at a later time as funds become available.

We have agreed that all the texts in the core corpora should originate in the three years 1990-1993, though some teams will start collecting in 1991. Each corpus will consist of 500 texts, and each text will contain approximately 2,000 words. The texts will be drawn from spoken English and manuscript English as well as printed English. In general, the population represented in the corpus will be adults of 18+ who have received formal education through the medium of English to the completion of secondary level of schooling. Account will be taken, for later analysis, of sociological variation in the participants, such as age, sex, region, occupation, ethnic affiliation, and educational attainment.

We envisage that the computerized material will be concordanced, using an interactive concordance program such as WordCruncher or KayeIBM. Further stages of processing would involve the application of word-tagging and parsing programs and (for the ambitious and well-funded) digitization of recorded speech. We expect that the Lancaster University CLAWS2 program will be used for word-tagging. I am currently investigating the relative merits of using the Lancaster and Nijmegen programs for parsing. Software devised by one team will be made available to all the others.

Much of the planning for the ICE project has been conducted by correspondence (including e-mail), but some participants have been in contact by phone or through personal visits. In addition, the participants and the International Advisory Board will come

together at the annual ICAME conference. We did so for the first time at the 1989 Bergen venue.

Besides hosting ICE meetings and providing an opportunity for papers from ICE contributors, ICAME will be the conduit for distribution of the ICE corpora and recordings.

The supplement to the London-Lund Corpus

Sidney Greenbaum
University College London

The London-Lund Corpus (LLC) as presently constituted consists of 87 texts of spoken English from the Survey of English Usage (SEU) at University College London that were converted into machine-readable form at the Survey of Spoken English at Lund University under the direction of Jan Svartvik. For the computerized corpus the Lund Survey devised a reduced form of the SEU prosodic transcription. The full prosodic and paralinguistic transcription is available at the SEU premises in the original non-computerized version.

When LLC was computerized, only 87 of the targeted 100 SEU spoken texts were available. Since then, the remaining 13 texts have been transcribed at the Survey of English Usage in the full transcription and they have also been computerized with the reduced transcription. Like all the SEU texts, these new spoken texts consist of about 5,000 words each. Disks containing this supplement (about 65,000 words) to the original LLC (about 435,000 words) have been sent to the Norwegian Computing Centre for the Humanities at Bergen. The supplementary texts should soon be obtainable from the Bergen Centre. Some of the

new texts are drawn from situations not represented in the original London-Lund Corpus. Only one (S.3.7) was surreptitiously recorded. In the list of texts below, I give their SEU and LLC numbers and the dates of the recordings:

S.3.7	(1984)	conversation between an architect and clients
S.5.12	(1985)	choir committee meeting
S.5.13	(1986)	meeting of a university academic council
S.6.7	(1971)	radio interview with an elder statesman
S.6.8	(1977)	psychiatrist's discussion group
S.6.9	(1985,1987)	computer lessons
S.9.4	(1985)	radio phone-in
S.9.5	(1985)	dictaphone recordings
S.10.9	(1976)	science demonstrations
S.10.10	(1984)	tennis commentaries
S.10.11	(1986)	cookery demonstrations
S.11.6	(1986)	House of Lords debate
S.12.7	(1983)	university oration on Foundation Day

The 100 SEU written texts have also been computerized. These include 17 scripted texts that were read aloud, and in their computerized form they are also provided with a reduced prosodic transcription. The written half of the SEU computerized corpus is available only at the SEU premises, where it is provided with an interactive concordance browser.

LLC has been extensively used in publications. When the supplementary spoken texts are distributed, scholars are likely to want to refer to them either alone or in combination with the earlier 87 texts. To avoid misunderstanding, the texts should be referred to in this way in future publications:

LLC:o the original corpus (87 texts)

LLC:s the supplement (13 texts) to the original corpus

LLC:c the complete corpus (100 texts)

A description of the complete LLC is due to appear later this year in 'The London-Lund Corpus of Spoken English' by Sidney

Greenbaum and Jan Svartvik, a chapter in *The London-Lund Corpus: Description and Research*, edited by Jan Svartvik (Lund: University of Lund Press). An appendix to the chapter lists publications that have used material from LLC and SEU.

The Text Encoding Initiative

The Text Encoding Initiative (TEI) continues its work. This is a major international project sponsored by the Association for Computing in the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. The goal is to develop an interchange standard for text archives and collections of all kinds. The TEI project has decided to adopt the framework of the Standard Generalized Markup Language (SGML; ISO8879), which provides a syntax for descriptive markup of texts. The first recommendations are expected by the summer of 1990. For access to more information about the project, subscribe to the electronic discussion list TEI-L at the University of Chicago by sending the message SUBSCRIBE TEI-L + subscriber's full name to LISTSERV@UICVM.BITNET.

The ACL Data Collection Initiative

The Association for Computational Linguistics has started a Data Collection Initiative. The object is to assemble a large corpus to be made available to researchers at cost and without

royalties. The initial goal is at least 100 million words encoded in standard form with SGML tagging. Inquiries can be directed to: Mark Y. Liberman (MYL@RESEARCH.ATT.COM) or Donald E. Walker (WALKER@FLASH.BELLCORE.COM). A related project is the work by Mitch Marcus at the University of Pennsylvania on a "Tree Bank of Written and Spoken American English", i.e. sentences with part-of-speech markers, syntactic parsing, and other forms of tagging.

A National Center for Machine-Readable Texts in the Humanities

Rutgers and Princeton Universities have received grants to undertake the planning for a National Center for Machine-Readable Texts in the Humanities. The initial goals will be "the continuation of an ongoing inventory of machine-readable texts; the cataloging and dissemination of inventory information to the broader scholarly community; the acquisition, preservation and servicing of textual datafiles which would otherwise become generally unavailable; the distribution of such datafiles in an appropriate manner; and the establishment of a resource center/referral point for information concerning textual data" (Humanist Discussion Group, Vol. 3, No. 761, 19 Nov 1989). Inquiries can be directed to Marianne Gaunt (GAUNT@ZODIAC.RUTGERS.EDU) or Robert Hollander (BOBH@PHOENIX.PRINCETON.EDU).

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

New material

A supplement to the London-Lund Corpus is now available through ICAME; see the presentation by Sidney Greenbaum earlier in this issue. Another new addition is an example text illustrating the transcription system used in the London-Lund Corpus. The example text is available in transcription with an audio cassette (in contrast to the actual corpus texts, which can only be obtained in transcribed form). It is accompanied by a transcription guide.

Through ICAME it is now also possible to obtain the machine-readable version of the Polytechnic of Wales Corpus,

which contains orthographic transcriptions of some 65,000 words of child language data. The corpus is parsed according to Hallidayan systemic-functional grammar. There is no prosodic information. See further the presentation by Clive Souter, *ICAME Journal* 13 (1989), p. 20ff.

The tagged LOB Corpus (horizontal format) is now also available in a version indexed for WordCruncher.

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 (1987), pp. 44-47.

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

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

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

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

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

LOB Corpus, WordCruncher version (diskette): This is an indexed version of the tagged LOB Corpus (horizontal format). It can only be used with WordCruncher.

London-Lund Corpus, text, original version *(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. 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 original London-Lund Corpus (text categories 4-12).

London-Lund Corpus, supplement (diskette): The remaining 13 texts of the 100 spoken texts collected and transcribed at the Survey of English Usage, University College London. See the presentation by Sidney Greenbaum earlier in this issue.

London-Lund Corpus, example text: An example text available in transcription with an audio cassette. See above under 'New material'.

Melbourne-Surrey Corpus (tape or diskette): 100,000 words of Australian newspaper texts. See the article by Ahmad and Corbett, *ICAME Journal* 11 (1987), 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 (1988), pp. 15-26.

Lancaster/IBM 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 (1988), pp. 76-77.

Polytechnic of Wales Corpus (tape or diskette): Orthographic transcriptions of some 65,000 words of child language data. The corpus is parsed according to Hallidayan systemic-functional grammar. There is no prosodic information. See further *ICAME Journal* 13 (1989), p. 20ff.

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, the Lancaster/IBM Spoken English Corpus, and the Polytechnic of Wales 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. Note also the example text referred to above. Users of the London-Lund material are also recommended to consult J. Svartvik (ed.). *The London-Lund Corpus: Description and Research*, Lund University Press (in press).

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