ICAME NEWS

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

Published by: The Norwegian Computing Centre for the Humanities, Bergen
The Norwegian Research Council for Science and the Humanities

NAVF

Machine-readable
texts in
English language
research

No. 8

Part I

May 1984

PART I

CONTENTS

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

INTRODUCTION

Stig Johansson
University of Oslo

Jostein Hauge
Norwegian Computing Centre
for the Humanities

The bulk of the present issue has been guest-edited by Jan Aarts (University of Nijmegen), who submitted a camera-ready copy of pp. 9-83. The contributions vary from accounts of 'tools' for corpus analysis (Aarts, van Halteren, van der Steen) to corpus-based studies of aspects of English grammar (de Haan, Akkerman, van den Hurk *et al.*). The last two papers were written by graduate students in Amsterdam. We are grateful to Jan Aarts for presenting these samples of current research in Holland and invite others to guest-edit future issues of *ICAME News*.

The demand for the material distributed from Bergen is now greater than ever before. At the time of writing, material has been distributed to over 90 individual researchers or research institutions all over the world (see Table 1). The number of researchers or research institutions receiving *ICAME News* is close to 400 (see Table 2).

The material currently available is specified on pp. 84-85. Later this year it will be possible to order a version of the LOB Corpus with grammatical tagging (cf. the description of the LOB Corpus tagging project in *ICAME News* 7, 1983). There are plans to produce lemmatized word lists and concordances sorted by word and tag as well as a dictionary of collocations.

The organization of ICAME will be informal, as it has been since the start in 1977, with Stig Johansson acting as co-ordinating secretary and the Norwegian Computing Centre for the Humanities being in charge of the distribution of material (secretary in Bergen:

Torill Revheim). Matters relating to ICAME are discussed at conferences on computers and English language research, the last one held at the University of Nijmegen in 1983 (see the report on pp. 9-24). The next conference will be arranged in Windermere, May 21-23, 1984, under the direction of Geoffrey Leech, University of Lancaster, and John Sinclair, University of Birmingham.

Table 1 ICAME material distributed from Bergen

Country	Institutions or individual researchers
Australia	5
Belgium	5
Bulgaria	1
Canada	2
Denmark	5
England, Scotland, Wales	19
Finland	9
France	1
Israel	2
Italy	3
Japan	3
Netherlands	4
New Zealand	1
Norway	6
Sweden	8
Switzerland	1
USA	9
West Germany	9
Total:	93

Table 2 Circulation of *ICAME News*

Country	Institutions or individual researchers
Australia	6
Belgium	7
Bulgaria	1
Canada	19
Czechoslovakia	3
Denmark	11
England, Scotland, Wales	52
Finland	15
France	11
German Democratic Republic	1
India	3
Israel	4
Italy	4
Japan	4
Netherlands	11
New Zealand	3
N-Ireland	1
Norway	38
Poland	1
Portugal	1
Quatar, Arabian Gulf	1
South Africa	1
Sweden	28
Switzerland	4
USA	132
USSR	1
West Germany	19
Yugoslavia	1
Total:	383

One of the objects in starting ICAME was to further cooperation and prevent duplication of research. We know that the material distributed is widely used but find it difficult to survey the research that has been completed or is currently being conducted. Some information on recent publications and ongoing research is given below. While the list is not exhaustive, it should give an idea of the range of studies using computer corpora.

Publications:

Aarts, Jan and Willem Meys, eds., forthcoming. *Recent Advances in the Use of Computer Corpora in English Language Research.* Amsterdam: Rodopi Publishers.

Aijmer, Karin. 1983. 'Emotional Adjectives in English'. In F. Karlson, ed. *Papers from the Seventh Scandinavian Conference of Linguistics.* Publications No. 9. Department of General Linguistics, University of Helsinki. 199-219.

Aijmer, Karin. 1984. '*Go to* and *Will* in Spoken English'. In Ringbom & Rissanen (1984), 141-57.

Altenberg, Bengt. 1984. 'Lexical and Sex-Related Differences in Spoken and Written English: Some Results of Undergraduate Research at Lund University'. In Ringbom & Rissanen (1984), 279-98.

Altenberg, Bengt, forthcoming. 'Causal Linking in Spoken and Written English'. To appear in *Studia Linguistica.*

Altenberg, Bengt and Gunnel Tottie. 1984. 'Will There be Texts in This Class? Writing Term Papers within a Research Project'. In Ringbom & Rissanen (1984), 265-77.

Coates, Jennifer. 1983. *The Semantics of the Modal Auxiliaries.* London & Canberra: Croom Helm.

Elsness, Johan, forthcoming. '*That* or Zero? A Look at the Choice of Object Clause Connective in a Corpus of American English'. To appear in *English Studies.*

Enkvist, Nils Erik, ed. 1982. *Impromptu Speech: A Symposium.* Publications of the Research Institute of the Åbo Akademi Foundation. Åbo: Åbo Akademi.

Fjelkestam-Nilsson, Brita. 1983. *ALSO and TOO: A Corpus-Based Study of Their Frequency and Use in Modern English.* Stockholm Studies in English 58. Stockholm: Almqvist & Wiksell.

Flognfeldt, Mona E. 1984. 'The Semantics and Pragmatics of Deverbal Nouns Ending in *-ee*: A Report on Work in Progress'. In Ringbom & Rissanen (1984), 57-67.

Forsheden, Oscar. 1983. Studies on Contraction in the London-Lund Corpus of Spoken English. ETOS Report 2. Department of English, Lund University.

Francis, W. Nelson and Henry Kučera. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin Company.

Garnham, Alan, Richard C. Shillcock, Gordon D.A. Brown, Andrew I.D. Mill and Anne Cutler. 1981. 'Slips of the Tongue in the London-Lund Corpus of Spontaneous Conversation', *Linguistics* 19, 805-17.

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

Jacobson, Sven. 1982. 'Modality Nouns and the Choice between *to*+Infinitive and *of+ing*', *Studia Anglica Posnaniensia* 15, 61-71.

Johansson, Stig, ed. 1982. *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities. (Distributed by Studia Universitetsbokhandel, Parkveien 1, N-5014 Bergen-Universitetet, Norway.)

Kjellmer, Göran. 1982. '*Each Other* and *One Another*: On the Use of the English Reciprocal Pronouns', *English Studies* 63, 231-54.

Kjellmer, Göran. 1983. '"He is one of the few men who plays jazz on a violin." On Number Concord in Certain Relative Clauses', *Anglia* 101, 299-314.

Kjellmer, Göran, forthcoming. 'Why _great:greatly_ but not _big:*bigly_? On the Formation of English Adverbs in *-ly*'. To appear in *Studia Linguistica*.

Krogvig, Inger and Stig Johansson, forthcoming. '*SHALL* and *WILL* in British and American English: A Frequency Study'. To appear in *Studia Linguistica*.

Leech, Geoffrey, Roger Garside and Eric Atwell. 1983. 'Recent Developments in the Use of Computer Corpora in English Language Research'. *Transactions of the Philological Society 1983*. Oxford: Basil Blackwell. 23-40.

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

Oreström, Bengt. 1982. 'When is it my Turn to Speak?' In Enkvist (1982), 267-76.

Oreström, Bengt. 1983. *Turn-Taking in English Conversation*. Lund Studies in English 66. Lund: CWK Gleerup.

Ringbom, Håkan and Matti Rissanen, eds. 1984. *Proceedings from the Second Nordic Conference for English Studies.* Publications of the Research Institute of the Åbo Akademi Foundation. Åbo: Åbo Akademi.

Rissanen, Matti. 1979. 'On the Position of *Only* in Present-Day Written English'. In S. Jacobson, ed., *Papers from the Scandinavian Symposium on Syntactic Variation.* Stockholm Studies in English 52. 63-76.

Stenström, Anna-Brita. 1982. 'Feedback'. In Enkvist (1982), 319-40.

Stenström, Anna-Brita. 1984. *Questions and Responses in English Conversation.* Lund Studies in English 68. Lund: CWK Gleerup.

Svartvik, Jan. 1982. 'The Segmentation of Impromptu Speech'. In Enkvist (1982), 131-45.

Svartvik, Jan. 1982. 'Information Processing in Speech'. In S. Allén, ed., *Text Processing. Text Analysis and Generation; Text Typology and Attribution. Proceedings of Nobel Symposium 51.* Data Linguistica 16. Stockholm: Almqvist & Wiksell.

Svartvik, Jan, Mats Eeg-Olofsson, Oscar Forsheden, Bengt Oreström and Cecilia Thavenius. 1982. *Survey of Spoken English. Report on Research 1975-81.* Lund Studies in English 63. Lund: CWK Gleerup.

Thavenius, Cecilia. 1982. 'Exophora in English Conversation'. In Enkvist (1982), 307-17.

Thavenius, Cecilia. 1983. *Referential Pronouns in English Conversation.* Lund Studies in English 64. Lund: CWK Gleerup.

Thavenius, Cecilia. 1984. 'Pronominal Chains in English Conversation'. In Ringbom & Rissanen (1984), 209-19.

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

Tottie, Gunnel. 1984. 'Is There an Adverbial in This Text? (And if so, What is it Doing There?)' In Ringbom & Rissanen (1984), 299-315.

Tottie, Gunnel and Carita Paradis. 1982. 'From Function to Structure. Some Pragmatic Determinants of Syntactic Frequencies in Impromptu Speech'. In Enkvist (1982), 307-17.

Wikberg, Kay. 1984. 'Some Critical Observations on Present-Day English Lexicology'. In Ringbom & Rissanen (1984), 103-16.

Unpublished reports/work in progress:

G.G. Corbett, Department of Linguistics and International Studies, University of Surrey (in collaboration with K. Ahmad, Computing Unit, University of Surrey): 1. We are setting up a small corpus of Australian English for comparative purposes; 2. Dr. Ahmad has written SEARCHSTRING, a simple concordance package which we intend to make available for student use. Our work is mainly in Russian.

E. Anne Cutler, University of Sussex, reports on: Andrew I.D. Mill, An Investigation of the Syntactic Contexts of Pause Fillers (M. Phil. thesis, 1982), G.D.A. Brown, frequency count on certain word classes.

Nina Devons, The Hebrew University of Jerusalem: A Study of Common Multiple-Meaning Words in the Brown Corpus. A frequency, semantic, contextual and usage dictionary of approximately 300 common English words and the compounds in which they occur (in progress).

Louis Goossens, University of Antwerp, reports that they make use of a word-in-context program to provide data for syntactic and semantic investigations. Computer corpora used to provide students with material in seminars in English linguistics. Thesis in progress: H. Mens, A Corpus-Based Study of the Verbs of Donation.

W. Hullen, University of Essen: a detailed study of 'Time and Tense in Everyday Dialogue' is in preparation. Some limited projects are envisaged, including an investigation of the use of the word *actually*.

Ossi Ihalainen, University of Helsinki: Noun Modification by Participles in American English (in progress); Ilona Rinta-Filppula, Aspects of Dependency Relations between some Verbs and Constructions with the Preposition *to*, Master's Thesis, 1982. (both based on the tagged version of the Brown Corpus)

M.F. Lynch, University of Sheffield: Extensive work has been done by us on the Brown Corpus. We have the intention of repeating these analyses, e.g. in regard to text compression, dictionary structure and access, with other corpora.

Matti Rissanen, University of Helsinki, reports on: Mervi Lundmark, Phrasal Verbs in American English (unpubl. thesis); Terttu Nevalainen, Focusing Adjuncts in Present-Day Spoken English (article in progress); Matti Rissanen, Periphrastic *Do* in Present-Day Spoken English (article in progress).

Edgar W. Schneider, University of Bamberg: The Brown Corpus and the LOB Corpus will be used as textual basis in a study of the semantics and collocability of some 200 verbs which express thinking procedures (part of a larger study which will not be finished within the next three years).

Johannes Söderlind, University of Uppsala, reports on the following theses in progress: Ingegerd Bäcklund, Abbreviated Adverbial Clauses in English; Ann-Mari Fåhraeus, Supercharging: One Aspect of the Pregnant Use of Words.

University of Nottingham: Investigation of semantics/syntax/pragmatics of modal verbs and adverbs used epistemically, by Mrs. K. Twiewicz, University of Lódz, Poland (recently in England on British Council grant).

Jan-Ola Östman, Åbo Akademi, reports on: Ann Westerlund, Topicalization of Valency Adverbials in English (unpubl. thesis, 1980); Gun Leppiniemi, Textual Parameters in the Placement of Definite Time Adverbials in English (unpubl. thesis, 1982).

For further information on publications and current work, see the bibliography in Johansson (1982) and pp. 9-24 in this issue.

CONFERENCE ON THE USE OF COMPUTERS IN ENGLISH LANGUAGE RESEARCH

30 May - 1 June 1983, University of Nijmegen

This conference was organized by the English Department of the University of Nijmegen. Papers were read by linguists engaged in research on English computer corpora. Abstracts of the papers are given below. All the abstracts are author's abstracts, with the exception of the papers read by Sinclair and Geens, which were made by Nelleke Oostdijk of the University of Nijmegen. The full texts of the papers will appear in a volume entitled *Corpus Linguistics. Recent Advances in the Use of Computer Corpora in English Language Research* (eds. J. Aarts and W. Meys), to be published by Rodopi Publishers, Amsterdam, in May 1984. In addition to the papers read at the conference, the book will also contain contributions by R. Quirk and S. Greenbaum.

(1) Jan Svartvik, University of Lund

"Four-level Tagging of Spoken English"

AIM

The primary aim is to produce a semi-automatic device for grammatical analysis of spoken English.

INPUT

The input, "the text", is authentic spoken English from the London-Lund corpus in orthographic transcription with prosodic analysis.

STRATEGY

At stage one, the analysis is carried out on four levels within the boundaries of prosodic chunks called 'tone units', one at a time. This approach is based on the assumption that tone units are valid communicative units in spoken English.

At stage two, adjacent tone units will be analysed in terms of the grammatical structure assigned to them at stage one.

The reason for adopting this order of analysis is largely practical, but there are also psycholinguistic reasons, since it may well be that we as speaking animals, process speech in this fashion.

The approach may be described as level-oriented but also as "mixed" since we believe that our computer model is also process-oriented and "a human language understander operates on many levels simultaneously" (Winograd).

LEVELS

We have implemented the following four levels for the analysis of the text (text is included here since the output will have this form):

0 Text
1 Word level
2 Phrase level
3 Clause level
4 Discourse level

TEXT LEVEL

At the moment only word and tone unit boundaries are used in the text for the grammatical analysis. However, prosodic information will be helpful in certain cases, e.g. disambiguation of *that*.

WORD LEVEL

Word-class tags (totalling about 100) are assigned to the words in the text. The process is interactive and is carried out by a statistical algorithm which proposes tags for each word in the tone units, making use of a high-frequency lexicon and a list of suffixes.

PHRASE LEVEL

There are five sets of phrase level rules, which are ordered and cyclical. The phrase rules operate on the word-class tags, from left to right, in each tone unit:

VPH	Verb phrase
APH	Adverb phrase
JPH	Adjective phrase
NPH	Noun phrase
PPH	Prepositional phrase

CLAUSE LEVEL

A set of clause level rules operate on the grammatical phrases to which are assigned tags denoting elements of clause structure. At the moment there are algorithms for the following five major types of clause elements:

V	Verb
S	Subject
C	Complement
A	Adverbial
X	Noun phrases as element with no assigned clause function

DISCOURSE LEVEL

A number of items which are typical of or restricted to spoken discourse are difficult to account for in terms of a grammatical apparatus, for example

apologies:	*I'm sorry, excuse me, pardon, ...*
smooth-overs:	*never mind, don't worry, ...*
expletives:	*fuck off, bright spark, ...*
responses:	*really, that's right, I see ...*

Such items are assigned discourse tags at word-class level and not analysed in terms of phrase and clause levels but instead at their own discourse level.

(2) Mats Eeg-Olofsson, University of Lund

"Word-class Tagging of Spoken English"

TAG SYSTEM

The system of word-class tags is a refinement of the traditional system of parts of speech. The main dividing criterion is surface syntactic function. The tags are built up hierarchically of partly mnemonic elements. A basic tag consists of two letters, the first of which denotes the main category:

A	adverb
C	conjunction
D	discourse
E	predeterminer
G	relative pronoun
J	adjective
N	noun
P	preposition
R	pronoun (other than relative)
T	determiner
V	verb
X	miscellaneous

whereas the second indicates a subdivision of it: VA main verb, VM modal auxiliary ... Morphological information can be added to the basic tag after a plus sign to distinguish, say, *will* (VM+8) from *would* (VM+9). Contractions (*don't you, d'you ...*) split up into two separate grammatical units for the purposes of phrase level analysis receive compound tags consisting of the tags of the component words separated by an asterisk. Word-class tagging is used mainly as a stepping-stone to phrase level tagging. Consequently, the system of word class tags is continually being revised to suit the needs of higher-level tagging.

CHANGES TO THE TAG SYSTEM

Recent changes to the system comprise the inclusion of the D category for units with special discourse function, and the facility to assign a single tag to several consecutive words.

Such multi-word tags are formed by adding a digit, denoting the number of words, to the basic tag. Expressions thus treated as unanalysable lexical units include adverbs (*as well*: AC2), conjunctions (*as though*: CC2), pronouns (*each other*: RO2), and more or less complex prepositions (*from the point of view of*: PA6). Other examples are adjectives (*up to date*: JA3) and proper name phrases (*New Guinea*: NP2). Multi-word tags are especially important in the D category, including greetings such as *how do you do* (DG4) and softeners such as *I mean, you know* (DS2).

TAGGING PROCEDURE

Word-class tags are assigned to the text in an interactive run. Tags produced automatically by a heuristic algorithm are displayed on a terminal, to be okayed or corrected by a linguist. The suggested tags are computed by a statistical decision algorithm as a Maximum A Posteriori estimate of the tag sequence corresponding to the tone unit text to be tagged. Statistics on the frequencies of tags of certain words (or endings) are used to compute the conditional probability of tag sequences, given the tone unit text.

PROBLEMS AND FUTURE DEVELOPMENT

The introduction of multi-word tags causes certain technical difficulties. For instance, the number of tags assigned to the word combination *sort of* must be two in the context *that sort of life*, but only one in *they sort of agreed*. Another technical problem is to find principled ways of estimating the probabilities of certain rare events (low-frequency tag transitions, words being used in metalanguage or in foreign language quotations).

At present word-class tagging is performed by a separate program. In the future, the tagging at all four levels will be done in a single interactive run. The ultimate goal is to create a truly integrated system, where knowledge from any level of analysis could be used to direct the tagging at any other level.

(3) Anna-Brita Stenström, University of Lund

"Discourse Tags"

The Discourse level takes care of certain speech-specific items whose functions are typically bound to the communicative situation and cannot be adequately accounted for at the word-class, phrase, and clause levels, where the pragmatic aspect is not covered. At the Discourse level such items are analysed in terms of organizational, planning, interactive, and communicative devices.

D-tags are assigned to:

1. items that occur almost exclusively in spoken interaction, eg: *yes, pardon, shut up,* and *please*;
2. items that have acquired a particular function in speech, eg: *look, sort of, you know,* and *well*.

At the moment the list of D-items includes:

CATEGORY	EXAMPLE	TAG
APOLOGIES	I beg your pardon	DA4
SMOOTH-OVERS	never mind	DB2
HEDGES	sort of	DC2
EXPLETIVES	fuck off	DE3
GREETINGS	hello	DG
INITIATORS	now	DI
NO	no	DN
ORDERS	shut up	DO2
TAGS	isn't it	DQ2
RESPONSES	sure	DR
SOFTENERS	you know	DS2
THANKS	thank you	DT2
WELL	well	DW
EXEMPLIFIERS	say	DE
YES	yes	DY

ORGANIZATIONAL devices are eg:

DI Initiators - the 'frame' *now* in '*now*, what are we going to do this weekend' indicates the transition from one stage in the discourse to the next.

INTERACTIVE devices are eg:

DG Greetings - *hello* expects another *hello* in return

PLANNING devices are eg:

DC Hedges - *sort of* in 'it's *sort of* rather pointless' gives the speaker more time to figure out how to go on (filler function) but at the same time it is partly possible to describe in syntactic terms, since it points forward to a head, *pointless* (planner function)

COMMUNICATIVE devices are eg:

DS Softeners - *you know* in utterance-final position acting as an intimacy signal and appealing for feedback

D-ITEMS IN MONOLOGUE VS DIALOGUE

I compared the occurrence of D-items in one monologue and one dialogue from the London-Lund corpus and found some notable differences. The D-items listed were distributed as follows in the two texts:

	DX	DC	DW	DI	DS	DE	DY	DN	DR	DQ	DA	DO	DP	DT	TOTAL
DIALOGUE	1	7	37	3	33	13	97	28	20	16	4	4	2	1	226
MONOLOGUE	-	-	10	7	4	1	--	--	--	--	-	-	-	-	22

The dialogue contained more than ten times as many D-items as the monologue. All categories were represented in the dialogue. Interactive items were completely lacking in the monologue, where the dominating category was DI Initiators (7 of the 10 instances of *well* are to be referred to that category), ie a matter of discourse organization. Interestingly, *well* in the dialogue was not used as an initiator in the first place, but as an R prefix. *You know*, which belongs to category DS Softeners, was realized by *as you know* in the monologue which is indicative of the different communicative situation. The monologue was not only planned, but the speaker had an audience in front of him; the dialogue was unplanned and unsurreptitiously recorded.

WORK TO BE DONE

Work at the D-level has only just begun and the list of D-items is far from complete. Additions to the list will be made continuously on the basis of more data, and certain modifications may be necessary. It has already become obvious, for instance, that the DR category, Responses, will have to be split up. At present it contains far too many heterogeneous elements.

(4) Jan Aarts, University of Nijmegen

"Report on Work in Progress"

The TOSCA (TOols for Syntactic Corpus Analysis) project has recently entered its fourth and last year. Its aim is to develop an interactive system for the automatic syntactic analysis of computer corpora, which allows the linguist to intervene at various points in the analyzing process, but also permits him to decide to let the analysis run its own course until a failure occurs.

The various components of the system were mentioned and commented on, as well as the stage of their development: the type of grammar used (Extended Affix Grammar, see below), the parser generator (a conversion program is used until a parser generator fully answering to our requirements is available), and the system's software (the Linguist's Workbench, see below). In May 1983 a new project was started in the TOSCA environment, in which a linguistic database (LDB) is constructed to accommodate the analyzed sentences resulting from sytems like TOSCA. The LDB will be built in such a way that it is easy and convenient to use for linguists without any computer experience. More recently (March 1984) a project (ASCOT) was started at the University of Amsterdam which aims to develop an English computer lexicon that can be 'plugged' into a system like TOSCA. At the same time a project was begun at the University of Nijmegen in which the TOSCA system is used for the analysis of a corpus of Modern Arabic.

(5) Nelleke Oostdijk, University of Nijmegen

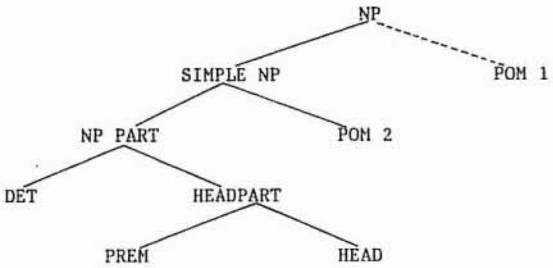
"An Extended Affix Grammar for the English NP"

Part of the work being done in the TOSCA project concerns the writing of an extended affix grammar (EAG) for contemporary English. So far subgrammars have been written for the noun phrase, the adjective phrase, the adverb phrase and the verb phrase. Experiences in writing - and recently also testing - the grammars have made it clear that EAGs, although from extra-linguistic origin, are quite suitable for linguistic purposes. One of the basic aspects of an EAG is that it is a two-level grammar (ie the grammar consists of a context-free basis which is supplemented with parameter-like affixes) so that the writer of the grammar has to decide what part of his description should be contained in the CF level of the grammar, and what information is to be handled in the affixes. The extent to which the linguistic analysis is influenced by the formalism is negligible; the form of the grammar is, however, affected by the fact that the grammar is to be used for the purpose of corpus analysis. Thus, keeping to the idea of the analysis of a structure as a hierarchical constituent structure, it was found that there was a need to distinguish not only functional and categorial constituents, but also constituents with which no function or category can be associated.

Our experiences in the writing and testing of the EAG for the English NP can be looked upon as representative of our experiences with EAGs in linguistics so far. Here it should be noted that the actual writing of the EAG for the NP was a purely theoretical affair, since the writer of the grammar had to do without any feedback that could have been obtained by testing the grammar against a test corpus. Only recently has a provisional computer program been employed which enables us to test our grammars. This is a conversion program by means of which EAGs can be transformed into CFGs and which we intend to use until the parser generator for EAGs that is being developed at the Department of Computer Science at the University of Nijmegen becomes available.

One of the problems we encountered while putting the initial version of the NP grammar to the test was that the conversion from EAG to CFG led to an enormous increase in the number of rules, so that the memory space was very soon exhausted. Therefore it was decided - for the time being - to write a new version in which only those affixes were retained which play a role in a strict NP environment, i.e. any affix yielding outgoing information was left out. The CF structure assigned to the NP remained the same.

Basically the NP structure looks as follows:



This structure allows for coordination at various levels. Apart from the usual coordination of NPs and HEADs also coordination of NP PARTs and HEADPARTs is possible. Thus it is possible to have coordination at four levels. Note that correlative coordination is only possible on the level of SIMPLE NP and NP PART. POM 1 (Postmodifier) will only be present in case there is coordination or apposition on SIMPLE NP level

In order to test the NP grammar we compiled a small testcorpus containing a variety of NPs. The analyses resulting from the test are quite satisfactory. The analysis of ambiguous NPs is in itself in no way problematic. The major problem consists in restricting apposition, especially when we have coordination as well as apposition in one and the same NP. Thus while the ambiguous analysis of an NP like

both the sugar and the milk
1.both (the sugar) and (the milk)
2.(both) (the sugar and the milk): ie two appositives

is found to be acceptable, the ambiguous analysis of a string like

three men
1. (three men)
2. (three) (men): ie two appositives

is found to be undesirable. Since it appears not to be easy to determine what restrictions should be placed on apposition, we have, for the time being, decided not to describe apposition so that the analysis of NPs like *both the sugar and the milk* and *three men* will be unambiguous. This means that in order to obtain an analysis of an appositive NP intervention is needed from the linguist.

Another point where we feel some improvements can be made concerns the representation of the analyses. Working with the display of the analyses on a computer terminal in the form of tree diagrams as we do, we find that the trees should only contain information that is of direct interest from a linguistic point of view. Therefore work is now being done on the design of filters for the output trees.

(6) Theo van den Heuvel, University of Nijmegen

"The Linguist's Workbench"

The central role in the software produced by the TOSCA project is played by the Linguist's Workbench (LWB). The LWB is a program regulating the interactive syntactic analysis of corpus material. The program is designed to be used by linguists inexperienced with respect to computers and their use. The comprehensive set of instructions available to the user includes a help-command and commands to fetch an utterance from the corpus to analyse it syntactically or morphologically, to store the resulting tree in a database, etc. There is a possibility to add lexical items to the lexicon for the duration of a terminal session. The correction of typing errors and other changes in the current utterance are allowed.

The LWB program is written in CDL2(see Koster, 1976); it is in rough working order. This first version can only be operated from a 3270-compatible display terminal and runs in a CMS environment. However, the programs are designed so as to enable transportation to different types of computers, including micro-computers, and the use of various kinds of terminals, with the least possible effort.

The LWB was demonstrated on video.

REFERENCES

- Koster, C.H.A. (1976), "Using the CDL Compiler Compiler" in: F. Bauer & I. Eickel (ed.), *Compiler Construction. An Advanced Course*. Springer Verlag.

(7) Geoffrey Leech, University of Lancaster

"Work on the LOB Corpus: Progress Report"

The Automatic Grammatical Tagging of the LOB Corpus, described in the last issue of *ICAME News*, is due to be completed by the end of September 1983. The Corpus has been divided into two halves: one half being processed in Norway (by Knut Hofland, Mette-Cathrine Jahr, and Stig Johansson), and the other half at Lancaster.

In October 1983 two new three-year projects on the LOB Corpus and related text processing will begin.

One project, funded by the Science and Engineering Research Council, will have four objectives:

1. the improvement and generalization of the Automatic Tagging System so that it will accept any input text and will attain a higher degree of success than the present 96.5%;

2. the development of a probabilistic syntactic analysis system for automatic parsing of the LOB Corpus;

3. the production of a syntactically-analysed version of the Corpus;

4. the development of a prototype computer lexicon for application to intelligent computer systems.

The other project will have as its objective the development of a context-sensitive textual error detector. This project will be carried out by Eric Atwell, and will use programs similar in principle to those already developed for the grammatical tagging of the LOB Corpus.

(8) Gunnel Tottie, University of Uppsala
Mats Eeg-Olofsson, University of Lund

"Tagging Negative Sentences in LOB and LLC"

Negative sentences with an indefinite expression after the verb can be of two types in English, SYNTHETIC as in *He saw nothing*, or ANALYTIC, as in *He did not see anything*. It has been argued, by Jespersen (1917) and Poldauf (1964: 370), that the syntactic variant is favoured in formal language 'because it yields a more elegant expression'. In order to ascertain in the distribution of the two types in conversational and written English as well as the factors determining the choice of variant, some 500 instances of relevant negative sentences were culled from each of the London-Lund Corpus (LLC) and the Lancaster-Oslo/Bergen Corpus (LOB), and a program was designed for interactive tagging of these sentences at a computer terminal. The tagging program was based on the findings of a heuristic study of synthetic and analytic negation in the Brown Corpus (cf Tottie 1983).

The program covers 34 factor groups, of which the following seem to be the most important: type of main verb, type of indefinite element (NP, pronoun, or adverb), syntactic function of the indefinite element, and the abstractness, premodification, and occurrence in a prepositional phrase of the indefinite element. Within each factor group the list of factors is exhaustive, i.e. it should cover every possible construction that might occur. For each factor group, the KWIC concordance line in which the relevant negative occurs is displayed on the screen. (If the context is insufficient, the linguist may consult a printout of the full corpus, but this is rarely the case.) The list of factors is displayed simultaneously with the concordance line, with the most probable factor listed first and serving as a default value, so that the linguist is only required to press the RETURN key to enter that value. (Probability estimates are based on Tottie 1983 and hold for written language.) The program is ordered so that only relevant factors are tested for each sentence; thus for example, if the negative occurs in a non-finite clause, the program will bypass the factor group TENSE (of the finite verb), and factor groups concerning countability, premodification or the distinction abstract/concrete will be bypassed as irrelevant if the indefinite element is not a noun. Similarly, information focus as manifested by intonation will only be displayed if the medium has been specified as conversation. Finally, the program allows the linguist to enter specific comments on each tested sentence.

REFERENCES

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

- Poldauf, Ivan (1964), "Some points of Negation in Colloquial English", in: J. Vachek, ed. (1964).

- Tottie, Gunnel (1983),*Much about Not and Nothing*. A study of the variation between analytic and synthetic negation in contemporary American English. Scripta minora: Kungl. Humanistiska Vetenskapssamfundet i Lund, 1983-1984: 1.

- Vachek, J., ed. (1964), *A Prague School Reader in Linguistics*, Indiana University Press, Bloomington.

(9) Willem Meys, University of Amsterdam

"You can do so if you want to - some elliptic structures in Brown and LOB and their syntactic description"

The theoretical importance in syntactic discussion of a particular construction may be quite unrelated to the actual spread of the phenomenon in statistical terms. The two constructions that are discussed here are a case in point: while both *do so* and elliptic (or "stranded") *to* have figured prominently in theoretical debates within the generative paradigm, there were only relatively few instances to be found in Brown and LOB, while stranded *to* occurred 48 times in Brown and 47 times in LOB.

Both constructions derive their theoretical importance from the fact that they can arguably be used as VP-tests; as such they can be used to establish what is, and what is not, supposed to be part of the VP constituent - assuming that replacement processes such as these will involve natural constituents rather than just old combinations of items. Most of the theoretical discussion has been based on a limited number of artificial, thought-up instances (with the commendable exception of Zwicky 1983). Analysis of corpus-instances may help to provide an empirical basis for the judgments involved, as well as providing welcome corrections and refinements where appropriate.

Traditionally, transformational descriptions of English have recognized a constituent labelled Aux as in (1)

(1) Aux → (not) Tns (M) (have en) (be ing)

By and large the *do so* instances in Brown and LOB suggest that these Aux elements are indeed outside of the VP. Cf. for instance:

(2) ... the Western powers have not acquiesced and should not do so.
(Brown B 02 40)

Clearly here *do so* replaces the stem *acquiesce* only; *not, have* and *en* being part of Aux rather than VP, are not replaced. A sentence like (3) presents an interesting question:

(3) For the only time in the opera, words are not set according to their natural inflection; to do so would have spoiled the dramatic point of the scene.
(Brown N 09 1310)

Notice that *do so* here apparently replaces *set words to their natural inflection*. This means that we either stick to rule (1) and assume that *do so* replacement is (or can be) ordered before a passive transformation, or, in a framework which does not make use of a passive transformation, adapt our PS-rule as in (4):

(4) Aux → (not) Tns (M)(have en)(be ing)(be en)

Another aspect that has not received much (if any) attention in theoretical discussions, is the fact that *do so* can also be used in a cataphoric (forward-pointing) way, as witness corpus-example (5):

(5) Though I can only do so as a layman, it is going to be necessary to *look at some of the scientific findings*.
(LOB J 52 26)

It would seem that in some cases there is a free choice between replacing either a "higher" or a "lower" VP, cf. corpus-example (6):

(6) All members who desire to do so may extend their remarks ...
(Brown H 03 140)

Here one could also have *do so* as the "higher" VP:

(7) All members who desire to extend their remarks may do so ...

This raises the question of what would be a proper formulation of the *do so* rule, and what, if any, are the constraints that may limit its application. In some corpus-cases the VP that is apparently replaced by *do so* is quite far from it. (There may even be a few sentences in between). Close examination of the data also reveals that neither c-command nor government can be invoked as a structural guiding-principle. Other problematic aspects of *do so* emerging from the data concern conjoined structures and - a well-known problem - the status (inside or outside of the VP) of certain adverbials.

Although both stranded *to* and *do so* can be used to replace VPs, (8) and (9) may go to show that they cannot always be used in the same contexts:

(8) I'll write what you tell me to. (Brown P19 610)
(9) *I'll write what you tell me to do so.

By and large, though, stranded *to* presents similar problems to those discussed here in connection with *do so*. For a fuller account see the article in Aarts & Meys (eds.) forthcoming.

(10) Dirk Geens, Université Libre de Bruxelles

"Semantic Analysis Automated for Large Computer Corpora and Their Exploitation"

Corpus linguists are frequently confronted with problems concerning the malleability of their corpus. In order to increase this malleability it is desirable to work with enriched data, ie a corpus containing supplementary information. Preferably the tagging of a corpus should be done automatically. Unfortunately, however, the reliability of automatic procedures is to some extent limited due to lack of semantic information. At the Free University of Brussels work has been done to develop a semantic analyser to make this information available.

(11) Pieter de Haan, University of Nijmegen

"Postmodifying Clauses in the English NP"

In this paper a system of manual tagging of corpus data was presented. The system uses numerical codes to represent syntactic and semantic features of postmodifying clauses in English NP's. These codes can be processed by means of a computer for statistical analyses.

The variables of the tagging system were discussed, together with the values that they can assume. Specific problems that were encountered during the tagging and the way in which these were solved, were also dealt with.

Finally, some examples of statistical analyses of the tagged corpus data were presented and discussed.

(12) Göran Kjellmer, University of Gothenburg

"Some Thoughts on Collocational Distinctiveness"

Some English collocations, recurring word sequences that possess grammatical structure, have a higher degree of lexical identity or distinctiveness than others. *Free enterprise* is somehow more distinctive than *free job*, *for all that* is more distinctive than *with all that*, etc. The paper addresses itself to the question whether collocational distinctiveness can be measured, and, if so, how it can best be done.

It is suggested that collocational distinctiveness is discernible in many linguistic dimensions, and some of the dimensions relevant to corpus-based study are discussed. They are

- Absolute frequency of occurrence
- Relative frequency of occurrence
- Length of sequence
- Distribution of sequences over texts
- Distribution of sequence over text categories
- Structure of sequence

Collocations behave differently in these dimensions - some are more frequent than others, etc. - but while it is clear that the performance of a sequence in any one of the dimensions does not supply enough evidence for it to be placed with any degree of certainty on a final scale of distinctiveness, it is claimed that the combined results in all the dimensions will place the sequence on such a scale in a way that agrees well with speakers' intuitive assessment of its distinctiveness.

(13) Antoinette Renouf, University of Birmingham

"A New Specialized Corpus: EFL Materials"

Since the last conference in Stockholm, Birmingham University has embarked on the production of a new series of specialised corpora of English. The first of these is informally referred to as the 'TEFL Side Corpus', and consists of approximately one million words taken from the leading EFL course books on the international market, including such well-known English courses as the Longman *Kernel* series, the *Access to English* series from O.U.P., and *Encounters* and *Exchanges* from Heinemann. Titles have been selected on the basis elicited from sixty-one British Council Offices throughout the world; the 26 works so selected comprise lesson material across the range of competence from beginner to upper intermediate level.

The purpose of the new corpus is to provide information on the kind of language which is likely to be accessible to learners of English. It will also be a unique resource for the analysis of the instructional (meta)language commonly used in

EFL books, and for a comparison of the features of constructed and natural language.

With such purposes in mind, the selected books have been edited, by means of a simple coding system, so that five different language types are identified and are retrievable for analysis. These consist of instructional language, non-authentic spoken, non-authentic written, and authentic spoken and written language.

Once coded, the books were keyed in their entirety onto computer tape. Whilst it would normally have been more efficient to carry out this process by means of the Kurzweil optical scanner, the non-linear nature of the text layout led to the decision to make use of a keyboarding agency for the task. The more recent EFL publications have been adventurous in incorporating speech balloons, cursive script and a wide range of type-faces in their design layout, in an effort to achieve visual appeal. Indeed, these features caused certain problems even for the keyboarders, in the matter of verification. Further work in error reduction has been done in-house, and overall, the degree of accuracy is now acceptably high.

After keying, the corpus was concordanced and word-lists were produced over a weekend on the university ICL 1906 mainframe computer. This data now exists in basic KWIC format on microfiches.

The corpus still remains to be analysed, but it is already clear from an initial glance at the statistical data that the language represented differs in many ways from the language of other corpora comparable in size, such as the BROWN, LOB and LEUVEN collections. This is to be expected, since it is a single genre corpus, reflecting variations within that genre rather than across the broader range covered by the other corpora.

Enhancements to the EFL corpus are currently being considered. One of these would involve the recording of classes based on certain parts of the course material, particularly at the beginner level. The aim here would be to identify the degree and type of language reinforcement which the learner is exposed to in the classroom in addition to that which he/she meets in the text book. Another supplement to the corpus would involve the transcription and inclusion of language available on cassette tapes relating to the various course series. These activities would provide a broader framework for interpreting the raw statistics from the corpus.

There is already considerable interest being shown, both within and outside Birmingham University, in the potential for EFL exploitation which this corpus offers, and it is expected that it will spawn a number of significant pieces of research.

(14) John Sinclair, University of Birmingham

"Naturalness in Language"

There are a very large number of well-formed sentences which do not seem natural to a sensitive native speaker. Therefore it seems logical to assume that these sentences violate some restrictons which are not among the criteria for well-formedness. Whereas well-formedness judgements are typically statements about the structure of a sentence in isolation, *naturalness* can best be defined as the concept of well-formedness of sentences in text. There is no reason to believe that the restrictions on the naturalness of a sentence are any less central in language structure than those for well-formedness. Moreover, the concept of naturalness may be particularly useful to the learner of a language.

It might be supposed that naturalness will always be probabilistic, and therefore distinct from well-formedness, which is absolute. When, however, sentences are described in their textual environment (or *co-text*) there may well be absolute or nearly absolute statements to be made about their privileges of occurrence.

So far evidence has been found for the existence of three sets of choices which indicate text dependency of various kinds; these are:

- *allowables*, i.e. requirements of text which do not need to be realised in any particular sentence, and do not affect well-formedness;
- *rangefinders*: somewhere in the co-text or context will be found the item signalled by the dependency choice - or the text is problematic;
- *supporters*, many lexical and/or syntactic choices have a tendency to occur with each other, and so the presence of one is valuable evidence for the existence of another.

The analysis of a sentence will be in terms of allowables, rangefinders and supporters in the first instance. Allowances will be made for features of register and other types of systematic variety and shared knowledge and experience. Then some observations will be made about the naturalness of a sentence. The form of naturalness statements is currently in terms of three parameters:

- neutrality
- isolation
- idiomaticity

The study of structure below the sentence has suggested certain assumptions that can be made for the study of text structure at sentence level. These assumptions must now be tested through an extended study of texts, which will establish the precise conditions for naturalness.

The general concept of the well-formedness of text is arousing interest at present, and naturalness is offered as a useful category to describe textual well-formedness among sentences.

THE LDB: A LINGUISTIC DATA BASE

Jan Aarts
University of Nijmegen

Past and current research in the field of syntactic corpus analysis has shown the need for a system that makes syntactic analyses accessible for the linguist. Projects like the Dutch CCPP (Computer Corpus Pilot Project) and the TOSCA-project are concerned with the construction of syntactically analyzed corpora. Unlike the CCPP project-design, the TOSCA-system was designed to incorporate a database component. In the database the results of the analysis are stored. By means of a query system the database user should be able to retrieve non-trivial syntactic information from the corpus. Since it was felt that within the TOSCA-project only limited attention could be given to the development of such a database and query system, and in view of the fact that it is an essential requirement for any kind of corpus-based research, it was decided to start another project: the LDB (= Linguistic DataBase) project.

The LDB project then, can be seen as supplementary to projects like the current TOSCA-project. It will be concerned exclusively with the development of a database and further tools which will enable the linguist to retrieve syntactic information from a corpus. It is a joint project of the Departments of English and Computer Science of the University of Nijmegen. The project started in May 1983 and will last two years. It is financed by the Netherlands Research Council.

The project includes the following activities:

- the development and implementation of a query system which enables the linguist to put questions in linguistic terms to a database containing fully analyzed sentences or other linguistic units;
- the development of the database organization;
- literature study with respect to tree-pattern-matching.

The software to be developed within the project should be portable.

1. A HIERARCHY OF CORPORA

A syntactically analyzed corpus is a set of sentences in the corpus language, to each of which has been assigned a syntactic description which can be represented in the form of a tree structure. With each node in the tree a function and a category label may be associated. The function label indicates the role the constituent plays in the larger linguistic structure.

We take it that the linguist when starting out with his research on an analyzed corpus will want to investigate the structures of one particular kind at a time and not all the structures contained in the corpus simultaneously. Therefore, it appears logical to create a new (sub)corpus containing only the constituents or structures involved. For example, a linguist who is interested in postmodifying clauses in the NP will want to extract from the initial corpus (C-0) a subcorpus of just those NPs which contain a postmodifying clause. The resulting corpus (C-1) forms the starting-point for further research. This new corpus is likely to be considerably smaller than the initial one. Consequently, the time needed to search it will be less than would be the case if the entire (initial) corpus C-0 was to be searched. Whereas the creation of a subcorpus C-1 can be looked upon as a first step in the process of reducing the bulk of data originally contained in the database, a second step consists of the successive classification of subcorpora. As a result of such a classification process a hierarchy of corpora emerges (Fig.1).

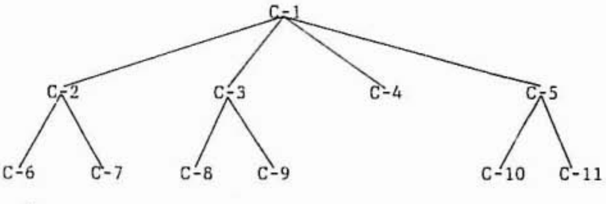


Figure 1

Each corpus in this hierarchy of corpora can serve as a source of examples of a particular syntactic phenomenon. The phenomenon characteristic of a given corpus is more specific than the phenomenon typical of the corpus immediately dominating it. The number of elements within each corpus can serve as a basis for quantitative statements. On the basis of a comparison between a node and all its daughters, statements can be made about the frequency of a particular phenomenon in relation to the frequency of a more general phenomenon.

The basic procedure in the classification process is the creation, on the basis of a corpus C-m, of a new corpus C-n, which contains just those elements of C-m that fit a particular syntactic description.

2. *FILTERS*

This procedure is illustrated in figure 2, where C-m and C-n are the source and the target corpus respectively, and F(ilter) stands for the syntactic restriction imposed on C-m and giving rise to C-n.

```
C-m  --------/--------  C-n
             F
```

Figure 2

It will be clear that filters can serve a dual purpose:

1. they can be used as patterns to identify structures so that these can be counted , displayed on a terminal screen for inspection, etc.;

2. they can be used to create subcorpora containing just those structures defined in the filter.

In our approach syntactic descriptions are represented by tree diagrams. Filters then are *tree patterns* representing the structure of the constituents the user is interested in. If such a pattern is used to create a subcorpus, the root of the pattern will become the root of the structures contained in the subcorpus.

Although at this stage we do not want to commit ourselves as to the form the filters should have, we will illustrate the possible forms filters might have below.

SU:NP will yield all subject noun phrases in the corpus

:NP(PLU) will yield all plural noun phrases, irrespective of their function

[SU,OD]:NP gives noun phrases functioning either as subject or as direct object

```
:NP
 \
  POM:SF
```

gives all noun phrases containing as immediate constituent a postmodifying finite clause. The branch does not imply that there must be no other constituents

```
:NP
 .
  .
  NP
```

selects all noun phrases containing another noun phrase at any level of its structure, not necessarily as an immediate constituent.

As said above, the LDB software will also offer other tools than filters. There will be provisions for trees to be conveniently printed on paper or displayed on a terminal screen. In addition, certain standard statistical operations will be available.

3. FURTHER SYSTEM REQUIREMENTS

The system to be developed in the course of the project will have to be such that future users will be able to operate it without any computer-experience. Therefore, it is important that the system should be easy to handle and that, at the same time, non-trivial linguistic questions can be dealt with.

Efficiency is essential to the users of the system: the time needed to reach an answer should be in proportion to the intuitive complexity of the task.

Finally, if the system is to be of any practical use, it should be possible to transfer the databases, i.e. the software as well as the (sub)corpora, to other computers. More especially, it should be possible to use subcorpora on microcomputers. For a first extraction of the requested material a mainframe computer will be needed. The resulting subcorpus can be conveyed to a microcomputer by means of a direct connection or a floppy disk.

4. DISCRETE ACTIVITIES WITHIN THE PROJECT

A number of discrete activities can be distinguished within the project. They are mentioned and commented on below. They are given in rough chronological order, but some of them overlap in time.

4.1 CONNECTION WITH THE LINGUIST'S WORKBENCH

The LDB project is supplementary to the TOSCA project, in which an interactive system for the syntactic analysis of computer corpora is developed. The software of the TOSCA system is called the Linguist's Workbench (LWB); the products yielded by the interactive analysis are analysis trees of the sentences in the corpus with labelled nodes that give information about function, category and 'affixes'.[1] It is these trees that constitute the data for the LDB. The first task within the LDB project, therefore, is to establish the connection between the LWB and the LDB, which will at the same time provide the interface between the LDB and other systems yielding analyzed corpora. The data will be transported to the LDB by way of a file. The form of this file will be mainly determined by the question what information and what level of redundancy will be needed by the LDB to work efficiently.

The method of shipping data from one machine to another will also be considered at this stage of the project.

4.2 DEFINITION OF THE USER INTERFACE (EXCEPTING FILTERS)

The general command structure will be defined. Its form will depend on facility for the user and the possibilities of the terminals used. A selection will be made of

- one key commands
- abbreviations with clear mnemonics
- full word commands

Apart from the general command structure, the way of displaying, navigating over and printing a tree will be defined.

4.3 IMPLEMENTATION OF THE LDB WITHOUT FILTERS

Depending on the functions needed and on the possibilities of the target machines and operating systems, an initial form of the primitive access to the database will be defined. It may later be extended, when the use of the database by the filter algorithm necessitates this.

An implementation of the LDB using this first model of access will be made on a VAX 11/780. Filtering will not be possible yet at this stage.

4.4 DEFINITION OF THE FORM OF FILTERS AND THE FILTER EDITOR

After a study of the literature on tree grammars and automata, and consultation of future users of the LDB about the kind of questions that will be asked, a first definition of the form of filters will be made. The experiences with QUERY[2] (what users do or want to but cannot do with it) will form a large part of the information about user requirements. The first definition of filters may be slightly changed later, due to user evaluation or difficulty of implementation. A test set will be created for the filter algorithm.

The user interface of the filter editor will be influenced by the form of filters and the design of the general command structure.

4.5 IMPLEMENTATION OF FILTER EDITOR AND SIMPLE FILTER ALGORITHM

The filter editor will be implemented and integrated with the LDB of stage 4.2.

A simple (probably brute force) filter algorithm will be designed and implemented as a plug-in module to the LDB. Any possible insertion of a new filter algorithm causes new demands on the form of the data itself. The question of how to keep these demands from causing changes in other areas of the LDB will need some study.

4.6 TESTING THE SYSTEM

At this stage the now complete but not yet fully efficient system will be tested and evaluated by users. Feedback should cause no changes to the user interface. Two categories of feedback can be distinguished:

- discovery of bugs and other inconveniencies;
- assessment of the structure and complexity of the filters that are actually used.

This feedback will lead to expansion of the test set.

Also at this stage the LDB will be transferred to IBM. Apart from a working system on IBM, the result of this will be a desription of a procedure to transfer the LDB to other machines, encompassing the steps needed and the problems that will be encountered.

4.7 CREATION OF A MORE EFFICIENT FILTER ALGORITHM

With the help of the literature and the feedback of stage 4.6, a more efficient filter algorithm will be created and integrated with the LDB. The resulting system will be tested with the current test set and then distributed to the user LDB's for further feedback.

4.8 DOCUMENTATION

The documentation written in the course of the project will be collected into an LDB user manual.

NOTES

1. For a fuller account of the TOSCA system, see Aarts & v.d.Heuvel (1983) and Aarts & v.d.Heuvel (1984).
2. See Van der Steen (1982) and also his contribution to this issue.

REFERENCES

- Aarts, J. and T.v.d.Heuvel (1983): "Corpus-Based Syntax Studies", in: *Gramma* 7 (1983), 153-173.
- Aarts, J. and T.v.d.Heuvel (1984): "Linguistic and Computational Aspects of Corpus Research", in: Aarts & Meys, eds.(1984).

- Aarts, J. and W. Meys, eds.(1984): *Corpus Linguistics. Recent Advances in the Use of Computer Corpora in English Language Research*, Amsterdam, Rodopi.
- Johansson, S. ed. (1982): *Computer Corpora in English Language Research*, Bergen, Norwegian Computer Centre for the Humanities.
- Van der Steen, G. (1982): "A Treatment of Queries in Large Text Corpora", in: S. Johansson, ed.(1982).

USER INTERFACE FOR A LINGUISTIC DATA BASE

Hans van Halteren
University of Nijmegen

One of the results of the TOSCA project was the recognition of the need to provide linguists with means to explore syntactically analyzed corpus data. To satisfy this need the LDB project was started. A general outline of this project can be found in Aarts' contribution to this issue.

The future users of the linguistic database that is to result from the project are linguists. These should not be burdened with learning complicated program control procedures, since this would only divert from the linguistic aspects. Therefore we felt that special attention was to be given to the user interface of the LDB.[1]

This article presents our thoughts on this subject. They should not be read as final and irrevocable. The ideas presented should be seen as a snapshot of an evolutionary process.

GENERAL STRUCTURE

Our current model of the LDB is shown in Figure I. We see that the LDB consists of five major components, three of which the linguist can communicate with. The function of the components becomes clear if we examine the uses of the LDB. The two main uses are the following:

1. to enable users to examine analysis trees, which may be selected directly (Figure II) or by way of search (Figure III)
2. to extract statistical data about syntactic structures (Figure IV)

These uses necessitate:

1. a general command handler to tell the LDB what to do
2. a store of analyzed corpus data to be examined (eg the results yielded by the LWB)
3. a store of filters for search and restriction[2]
4. a tree viewer to show analysis trees to the linguist

FIGURE I

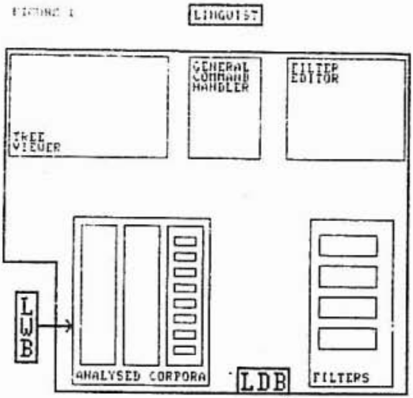
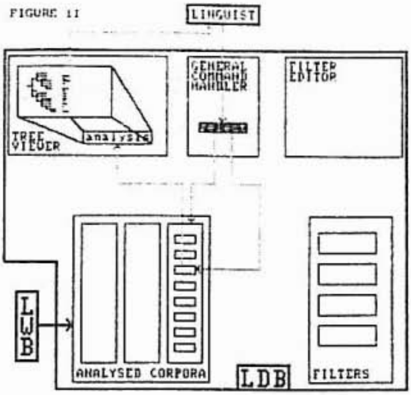


FIGURE II



Furthermore, the use of filters makes it necessary to provide tools to create them. For this a filter editor is included (Figure V).

FIGURE III

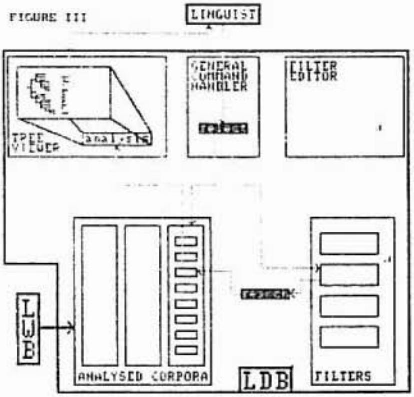
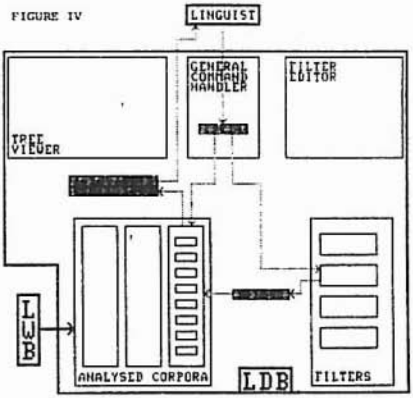


FIGURE IV



Sometimes we want to have information, not about the trees matching just one filter, but about trees containing two filters in some relation to each other.

FIGURE V

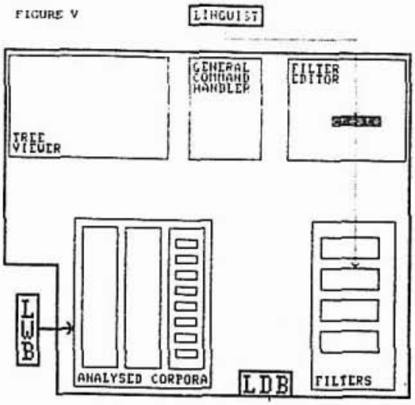
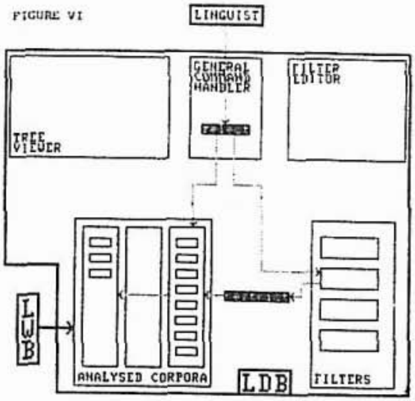


FIGURE VI



This could be done by constructing a more complicated filter incorporating just the kind of relation between the original filters we are interested in. Usually it is easier if we are able to constrict our view to just a part of an analysed corpus by creating a new one that only consists of those trees that match one of the original filters (Figure VI). After that we examine the new subcorpus with the second filter.

We will now describe in detail the way in which the user communicates with the LDB. It will be described as if it were it complete working system. In fact, a partial implementation has been created for testing the user interface and stimulating discussion about it. The filter editor will not be described, as its construction is planned at a later stage in the LDB project.

THE DATA AVAILABLE TO THE USER

In the LDB, the user is provided with several collections of data. First of course there are the corpora (including location codes, and possibly wordclass tags). Apart from this, in the initial state, he finds analysis tree structures for all utterances in the corpus. The nodes of these trees contain function and category labels as well as supplementary information, in the form of so-called affixes (eg SUBJECT:NOUN PHRASE(3RD PERSON,PLURAL)). The names of the analysed corpora incorporate the name of the corpus they belong to. The user can then create several additional items by constructing filters, which can be edited and used for selection. He can also create subsets of analysed corpora. To maintain simplicity, we will refer to these subcorpora with the term corpus.

THE GENERAL COMMAND HANDLER

The user is given a menu choice of the functions provided. We have chosen the menu form of input rather than the free typing of commands, because free typing implicates complete knowledge about possible commands, which ought not to be necessary in such a system. Also it is easier to choose with one or two keystrokes than to have to type long commands. For reasons of functional difference, there are two menus, one providing navigation within one database file and another for other functions.

MAIN MENU

The general function menu offers the following functions:

administrative commands

1. leave system

   RESULT: The session is terminated.

2. set options

   RESULT: The user is allowed to specify certain characteristics of the LDB (with respect to terminal communications).

general access

1. give a list of analysed corpora

   RESULT: a list is presented containing pairs (analysis name, corpus name) of the analysed corpora in the database.

2. examine analysed corpus

   RESULT: the user chooses an analysed corpus, it is selected for examination and the examination menu is presented to the user.

3. delete analysed corpus

   RESULT: the user chooses an analysed corpus, which is then removed from the system.

4. filter analysed corpus

   RESULT: the user chooses a filter and an original analysed corpus, of which a subanalysis is created containing only the trees matching the filter. It is given a name also provided by the user.

5. create database

   RESULT: the user chooses an analysed corpus and a new database is created consisting of a subcorpus containing those records of the original corpus that are referred to by the analysed corpus specified, and, of course, the analysed corpus itself. A possible use of this command is to create a small database for working on a smaller system (possibly a microcomputer).

filters

1. give a list of filters

   RESULT: a list is presented of the names of all the filters in the database.

2. edit filter

   RESULT: the filter editor is put into action and starts working on a filter specified by the user. If the filter does not exist, this function creates a new one. The editor can also be used to just look at a filter.

3. delete filter

   RESULT: the filter named by the user is removed from the database.

4. print filter

   RESULT: the filter specified is printed on a lineprinter.

EXAMINATION MENU

For navigation and information gathering within an analysed corpus another menu is provided. The choices in this menu represent the functions described below.

1. count trees

   RESULT: the total number of trees in the current analysed corpus is given.

2. count trees matching a filter

   RESULT: same as above, but only the trees that match the filter specified by the user (who is prompted for it) are counted.

3. select first tree

   RESULT: the first tree of the analysed corpus is taken as current tree.

4. select next tree

   RESULT: the next tree of the analysed corpus is taken as current tree.

5. select previous tree

   RESULT: the previous tree of the analysed corpus is taken as current tree.

6. select tree by number

   RESULT: the tree with the number chosen by the user is taken as current tree.

7. select tree by location

   RESULT: the user provides a location code and the tree which refers to the part of the corpus defined by the location code is taken as current tree.

8. select tree by search

   RESULT: the analysed corpus is searched for a tree matching the filter specified by the user (after prompting) and the first tree found is made the current tree. The user is asked if the search should be forward from the current tree, backward from the current tree, or forward from the beginning of the analysed corpus.

9. examine selected tree on screen

   RESULT: the tree viewer is activated showing the current tree, along with the original utterance.

10. print trees

    RESULT: the specified trees are printed on a line printer in the mode specified by the user.

11. examine corpus

    RESULT: after the user has been asked for start and end location codes, the part of the corpus between these locations is displayed.

12. go back to main menu

RESULT: the first menu is presented to the user, but for the moment the current analysed corpus and tree remain selected.

PARAMETERS OF FUNCTIONS

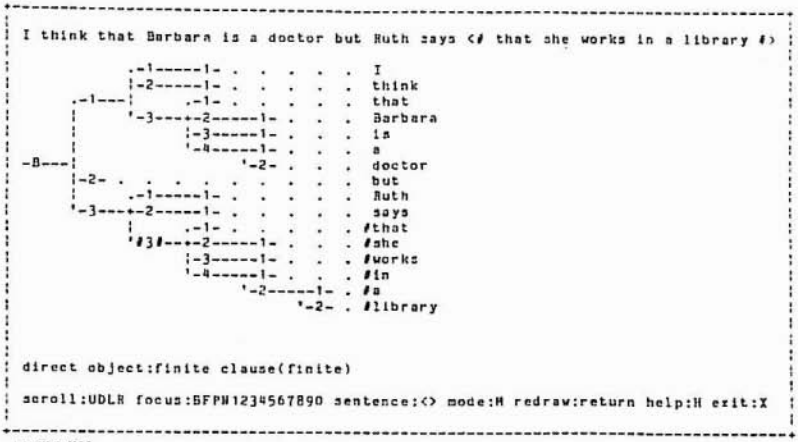
After choosing a function the user will be prompted for the information that is needed by that function. This will be done by means of a menu choice wherever possible, but in some cases (eg a location or a number) the information will have to be typed in full. Inline editing is allowed during typing of this information. An escape mechanism is provided to come back on a choice in case of a mistake, as well as a help mechanism to explain what input the LDB is waiting for.

THE TREE VIEWER

The problem in showing an analysis tree on a terminal screen is that the amount of information in such a tree is too big to be shown all at once. Seeing that there are in fact two dimensions of information within the tree (ie the structure of the tree and the labelling of the nodes), we have to choose which dimension to concentrate on. If we want to see as much structure as possible a normal size terminal (24 lines of 80 characters) will accommodate about 20 by 20 nodes, but this precludes the presentation of any node information. Therefore, there has to be a way to select parts of a tree. We have chosen for a presentation that always has a node as the centre of attention, the so-called focus. The user is provided with commands to move the focus around in the tree. He also has at his disposal two ways of looking at the tree. The treemap view shows the structure of the tree, along with the full node information of the focus. The environment view shows only the structure of the direct environment of the focus, but with abbreviations of the labels of the nodes shown.

DISPLAY MODES

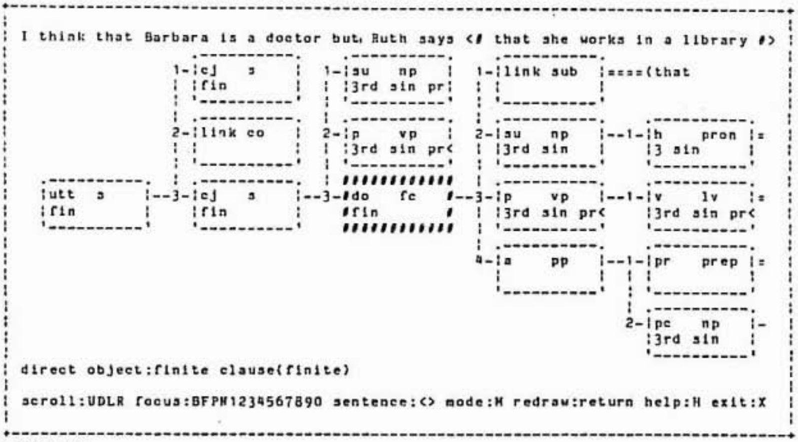
The first display mode is the tree map view (Figure VII). It shows the layout of the tree and the position of the focus. It shows no node information, except the information associated with the focus. This mode will be the one entered when the user starts viewing the tree.



```
I think that Barbara is a doctor but Ruth says <# that she works in a library #>

            .-1-----1- .  .  .  .  .  I
            |-2-----1- .  .  .  .  .  think
      .-1---|     .-1- .  .  .  .  .  that
      |     '-3---+-2-----1- .  .  .  Barbara
      |           |-3-----1- .  .  .  is
      |           '-4-----1- .  .  .  a
 -B---|                 '-2- .  .  .  doctor
      |-2- .  .  .  .  .  .  .  .  .  but
      |     .-1-----1- .  .  .  .  .  Ruth
      '-3---+-2-----1- .  .  .  .  .  says
            |     .-1- .  .  .  .  . #that
            '#3#--+-2-----1- .  .  . #she
                  |-3-----1- .  .  . #works
                  '-4-----1- .  .  . #in
                        '-2-----1- . #a
                              '-2- . #library


 direct object:finite clause(finite)

 scroll:UDLR focus:BFPN1234567890 sentence:<> mode:M redraw:return help:H exit:X
```

FIGURE VII

The second mode is the environment view (Figure VIII). The environment view shows the focus along with its grandfather, father, uncles, brothers, children and grandchildren. The information in the nodes has been abbreviated in order to be able to put it all on the display.[3]



```
I think that Barbara is a doctor but Ruth says <# that she works in a library #>
                  .----------.     .----------.     .----------.
                1-|cj    s   |   1-|su   np   |   1-|link sub  |====(that
                | |fin       |   | |3rd sin pr|   | |          |
                | '----------'   | '----------'   | '----------'
                | .----------.   | .----------.   | .----------.     .----------.
                2-|link co   |   2-|p    vp   |   2-|su   np   |--1-|h    pron |=
                | |          |   | |3rd sin pr<   | |3rd sin   |    |3 sin     |
                | '----------'   | '----------'   | '----------'    '----------'
 .----------.   | .----------.   | ############   | .----------.    .----------.
 |utt  s    |--3-|cj    s   |--3-#do   fc   #--3-|p    vp   |--1-|v    lv   |=
 |fin       |    |fin       |    #fin       #   | |3rd sin pr<    |3rd sin pr<
 '----------'    '----------'    ############   | '----------'    '----------'
                                                | .----------.    .----------.
                                                4-|a    pp   |--1-|pr   prep |=
                                                  |          |  | |          |
                                                  '----------'  | '----------'
                                                                | .----------.
                                                                2-|pc   np   |-
                                                                  |3rd sin   |
                                                                  '----------'
 direct object:finite clause(finite)

 scroll:UDLR focus:BFPN1234567890 sentence:<> mode:M redraw:return help:H exit:X
```

FIGURE VIII

The information associated with a node consists of a function name, a category name and affixes. Actually there are affixes belonging to the function and affixes belonging to the category, but because these two sets are always almost entirely the same they are merged into one set. The abbreviations used for all names are provided by the grammar writer in the grammar.

The abbreviations of function and category names must consist of five characters or less, so that the combination of these two will fit in the ten character space of the top part of each node shown. The space needed for affixes, however, may be larger than ten positions. In that case only the first ten characters are shown and a marker is placed to the right of them, signifying that there is more to see. The user can then scroll the affix part of all nodes to the left to see what is there. If he has scrolled to the left a similar marker is also placed to the left of the affix positions, and he can scroll right again. Scrolling takes place in steps of one affix name and is counted at the left of the window (ie one can start viewing from the first, then from the second, etc.).

In both display modes there is a line on the screen containing the original sentence. Since the sentence will almost always be too large for the area provided, it is possible to position this area as a window on the sentence.

COMMANDS

The commands provided inside the tree viewer can all be activated with one keystroke. They fall into several types.

General commands

1. X : leave the tree viewer
2. H : show the help screen
3. <return> : redraw screen (used when the WAIT option is activated and the screen is only partially changed with some commands, to force complete updating of the screen)

Changing view mode

1. M : flip between tree map view and environment view

Choosing the focus

1. B : set focus to the root of the tree (initial focus)
2. F : set focus to father
3. P : set focus to previous brother
4. N : set focus to next brother
5. 1 : set focus to first son
6. ...
7. 9 : set focus to ninth son
8. 0 : set focus to last son

Changing the display in environment view

1. D : scroll offspring of focus down
2. U : scroll offspring of focus up
3. L : scroll affixes (if not entirely shown) to the left
4. R : scroll affixes (if not entirely shown) to the right

Scrolling the screen in tree map view

1. D : scroll screen down
2. U : scroll screen up
3. L : scroll screen to the left
4. R : scroll screen to the right

Scrolling the sentence window

1. < : scroll sentence to the left
2. > : scroll sentence to the right

THE FILTER EDITOR

The filter language and hence the filter editor will be defined at a later stage of the project.

THE CURRENT SITUATION

At the moment we are extending the partial implementation mentioned above. At the same time we are examining the impact of ambiguous analyses on the system.

However, we have not stopped thinking about the part of the interface that is described here. It may well be that prospective users have specific wishes with respect to such a database. Therefore, any remarks on the interface are welcome. We stress once again that we consider a good user interface a very important feature, so that improvements due to new insights will be not be dismissed, however far the project has already proceeded.

NOTES

1. The term LDB will be used for the project as well as for the database itself.
2. We use the term *filter* both to indicate a certain structural pattern and the use to which it is put (ie that of creating subcorpora).
3. The abbreviations used in Figure VIII will be explained in the appendix. But the full information of the focus is presented at the bottom of the screen.

APPENDIX - ABBREVIATIONS USED

In Figure VIII abbreviations were used for the names of functions, categories and affixes. In this appendix we present the meaning of these abbreviations for the interested reader. We remark that the names and abbreviations can be chosen by the writer of the grammar and are not provided by the LDB.

FUNCTIONS

a	adverbial
cj	conjoin
do	direct object
h	head
link	linker
p	predicator
pc	prepositional complement
pr	prepositional
su	subject
utt	utterance
v	verb

CATEGORIES

co	coordinator
fc	finite clause
lv	lexical verb
np	noun phrase
pp	prepositional phrase
prep	preposition
pron	personal pronoun
s	sentence
sub	subordinator
vp	verb phrase

AFFIXES

fin	finite
pr	proper
prs	present tense
sin	singular
3rd	third person

ICAME NEWS is published by the Norwegian Computing Centre
for the Humanities (NAVFs EDB-senter for humanistisk forskning)
Address: Harald Hårfagres gate 31, P.O. Box 53, N-5014 Bergen-University, Norway