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 II

May 1984

PART II

CONTENTS

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

ON THE UNIFICATION OF MATCHING, PARSING AND RETRIEVING IN TEXT CORPORA

Gert van der Steen
University of Amsterdam

INTRODUCTION

In a previous article "A Treatment of Queries in Large Text Corpora" (van der Steen, 1982) a description was given of the so-called Query-system. The system allows for

- the use of context-free grammars for the transformation of a corpus into a two-level tree-structure;
- the use of patterns containing "Booleans" and "Don't cares" for the location of interesting linguistic phenomena to be found in the tree-structure.

We reported on the experiences of the users of the system and announced some extensions which should make it possible to deal with tree-structures at any level and with context-sensitivity. We also suggested that pattern-matching and parsing should be integrated.

In this paper we describe the results achieved so far and place them in the broader context of information-retrieval. Finally, we discuss the consequences of the unification of pattern-matching, parsing and querying of data-bases in relation to the potentials for the study of tagged corpora.

It has been suggested that a hierarchical structure is preferable in order to deal with the tagging of texts. Most texts that are used for scientific research in the humanities are structured in some way. We may think of the simple structure of a book that has been subdivided into chapters, paragraphs, sentences, words and characters. It is also possible for a text to contain some deliberately placed markers. This may be the case in an interview where the markers are used to distinguish between different speakers or between various stretches of speech uttered by one and the same speaker. These markers are typical of what we might call an imposed structure.

Another type of structure is inherent in formalized prose when someone restricts himself to sentences which can be described by a formal language. While reading a text of this kind it will be fairly easy to divide it into meaningful parts and label these by means of a limited number of well-defined markers. Thus a text is transformed mentally into a structure from which different relations may be inferred which hold between the various bits of information stored under the different labels.

A third type of structure can be found in texts on which some kind of syntactic structure has been imposed, either by hand or (partly) by machine. These texts are often employed as corpora by linguists studying grammatical phenomena.

In the simulation of a transformation process by means of a computer both computational linguists and artificial intelligence people participate, each with their own insights. They share a common resource of language description: the grammar, written in various formalisms. The simulation of the inference of stored knowledge by data-base people reflects a more pragmatic approach: these sys-

tems are mostly used in a business environment and should be efficient. Compared to natural languages, query-languages for data-bases are relatively simple.

In this paper we will show how the grammar-notation for the construction of a data-base containing formalized prose (see also van der Steen, 1982) may be used as a query-language. The file-structure is the same for each type of tagging mentioned above. The structure is hierarchical and the order of the elements remains significant on the horizontal as well as the vertical level.

SYNTACTIC PATTERN-RECOGNITION AS A DATA-BASE TOOL

Earlier (see van der Steen 1982), we observed that, generally speaking, it is rather difficult to foresee what kind of queries the user of a data-base system will come up with. It proves to be impossible to know beforehand what relations will be needed. Consequently (not knowing what relations to store), it is impossible to use a relational data-base system. Therefore, a different system will have to be selected, like for instance a hierarchical data-base system (Date, 1981).

The major drawback of a hierachical data-base system is that it allows only for queries with limited complexity. In order to be able to handle more intelligent queries we will have to introduce some extensions on the vertical as well as on the horizontal level. Given the link between syntactic recognition and retrieval in a hierarchically structured data-base, we may formulate queries for which the sequence of the vertical components in a particular tree is important. These vertical sequences are usually of a simple nature and are formulated in terms of a procedural query language allowing one to navigate horizontally as well as vertically through the tree-structure. In order to increase the expressive power of the queries we should allow the grammatical notation to function to its full extent on the vertical level. Thus, we will have, at the same time, the advantage of a concise notation and the simple semantics of formal language description where the grammar describes a set of sentences.

By allowing grammar-notations on the horizontal lines, we introduce the possibility of linguistic descriptions on free formatted text. As a consequence of the increased complexity as far as retrieval is concerned, the file-structure of the data-base should be reconsidered.[1] Therefore we propose a system which, on the one hand, acts as a hierarchical data-base system and, on the other hand, as a grammatical parser. Such a system also lends itself for the organization and querying of information written in natural language and structured according to the user's wishes. This structure takes the form of a labelled tree in which each sub-tree may differ in structure from another sub-tree.

Our main objective is to provide the users of the system with a query-language in which relations they are interested in can be represented in a simple and straightforward way. Apart from this we attempt to minimize the retrieval time. At the moment, the strategy that is being followed consists of searching the entire data-base (as if it was a large piece of text) and to skip, where possible, parts of it. On the basis of experiences gained in generating code for the answering of queries, we intend to design a more elaborate index-structure. The efficiency of recognition is especially important when it comes to searching a large text sequentially. We use the general botttom-up techniques developed for grammatical analysis (extensions of LR-parsing) to compile a query into an efficient piece of machine-code or into a code for an interpreter (which may run on a separate processor).

CONSIDERATIONS FOR THE RECOGNITION OF PATTERN-GRAMMARS

In order to allow for the possibility to skip parts of the input we prefer an on-line recognition strategy (no back-tracking). Since we may come across ambiguities in the process of recognition, the recognition-strategy should be designed explicitly to cope with these ambiguities in an efficient (and online) way. When the grammar-notation is used for the transduction of raw natural language input to a tree-structure, errors in the input should be reported at the correct point. Error-correction should provide for the collection of as many errors as possible in one run. The recognition strategy should provide for *space and time* trade-offs in order to be able to tune it to small as well as large computer-configurations. It will be possible to use optimisation techniques that are already available.

CONSIDERATIONS FOR THE IMPLEMENTATION OF A RECOGNIZER FOR UNIFORMED FORMALISMS

We propose a system that is portable to a high extent. Therefore we try to compile code for hypothetical machines which may be interpreted or assembled for real machines. Our most general hypothetical machine is a two-stack automaton in which each stack takes the form of a network in order to deal with ambiguities in an efficient way. Reports and free variables find a place on these stacks. When there is no context-sensitivity it is often possible to leave out one stack. Usually there is no recursivity in data-base applications in which case the first stack may be left out as well, generating just the code for a finite-state machine. We try to minimize the stack-instructions by having the compiler generate code for the stacks only when necessary.

FAST RECOGNITION OF PATTERN GRAMMARS IN LABELLED TREES

The existing formalisms of pattern-grammars may be characterized by the following eight factors:

1. Strength of recognition:
   - only terminal symbols (pattern-matching)
   - Chomsky type 3 (regular grammars)
   - Chomsky type 2 (context-free grammars)
   - Chomsky type 1 and 0 (context-sensitive grammars)
2. Meta-notation:
   - *none (only straight-forward notation of strings)*
   - BNF
   - BNF with extensions for regular expressions
   - Automaton-notation (like ATN-grammars)
3. Allowance for ambiguity:
   - none
   - by back-tracking
   - by parallel (and cooperative) computation

4. Allowance for patterns:
   - none
   - specifying "Don't cares" and "Arbs"
5. Allowance for tree-structures in grammar and text
6. Allowance for Boolean operators
7. Allowance for variables:
   - none
   - assignment from text to variables
   - global free variables (like ATNs)
   - local free variables
   - parameter variables (added to non-terminal symbols, as in Attribute and Affix-grammars)
8. Driving of semantic routines
   - triggering and reporting with text-elements
   - transducing

THE PARSPAT SYSTEM

In our design of the so-called Parspat system we implemented (up till September 1983) the following aspects:[2]

ad.1 Type Chomsky-2;

ad.2 Meta-notation: BNF with regular expressions;

ad.3 Treatment of ambiguity by parallel parsing (complexity in time: O(nxx3));

ad.4 "Don't cares" and "Arbs" (complexity in time: O(n));

ad.5 Labelled trees;

ad.6 Booleans: "or" and "and" (complexity in time: O(n));

ad.8 Reporting with the text-element last read.

We are currently implementing the factors:

ad.1 Type Chomsky-0;

ad.6 Boolean: "not" (complexity in time: O(n));

ad.7 Assignment from text or from grammar to free variables, which may be global, local and parameter;

ad.8 Transducing by specifying instructions in the grammar for building a tree. We have already shown the report-function. So far we have used it when transforming raw input (usually free text) to a tree-structure: in general, the report drives the semantic routines which are to be supported by the user. In the construction of trees these routines are quite simple. Therefore, we are now adding, in the same way as the "report"-function, a "build"-function in which one defines how the output-tree may be transformed: in this way the system is used as a transducer to transform the input into a tree-structure.

The system works interactively: searching may be interrupted and (part of) queries (= pattern-grammars) may be stored in libraries.

UNIFICATIONS

So far we have described the evolution of a pattern-matching system to a system which combines data-base manipulations and grammatical analysis in a uniformed way. It takes time to understand the way in which these two approaches may be unified, and it takes still more time to redesign programs to accord with this view of unification. The whole field of grammatical analysis is a turbulent one.[3] In our system we try to unify the various approaches as much as possible. We are currently working on an implementation in which it will be possible to use the notations found in transformational grammars, ATN-grammars and Attribute-grammars.

AVAILABILITY

In October 1983 the Parspat-system was rewritten in standard PASCAL.

ACKNOWLEDGEMENTS

I want to thank two people who assisted (and are assisting) with great enthousiasm in the implementation of the Parspat-system: Marijke Elstrodt and Menno Ytsma. I hope that in the Dutch national work group for corpus linguistics it will be possible to share some of the painstaking work on such systems.

NOTES

1. Here one might for example think of indices on different levels.
2. The numbers relate to the eight factors distinguished above.
3. See Winograd (1983).

REFERENCES

- Date, C.J. (1981), *An Introduction to Database Systems*, Addison-Wesley, 3rd ed.
- Van der Steen, G.J. (1982), "A Treatment of Queries in Large Text Corpora", in: S. Johansson, ed., (1982), *Computer Corpora in English Language Research*, Bergen.
- Winograd, T. (1983), *Language as a Cognitive Processor Vol. 1: Syntax*, Addison-Wesley.

RELATIVE CLAUSES COMPARED

Pieter de Haan
University of Nijmegen

The purpose of this article is threefold.[1] In the first place it compares the results obtained from the analysis of a corpus of written English with those obtained from two corpora of spoken English, in order to see whether written English differs from spoken English in a relatively small area of grammar. Secondly, some attention will be paid to the differences that can be observed in a comparison of the various subcorpora of which the written corpus is composed. Thirdly, the relation between the use of the relative pronoun in certain types of relative clauses and the (syntactic) environments in which these clauses occur, will be dealt with.

The corpora in question are Quirk's (1957) corpus of spoken English, Schwartz & Shine's (1972) corpus of spoken English,[2] and the written part of the Nijmegen corpus (see, eg Aarts & Van den Heuvel, 1980). It has not been possible to compare the three corpora with each other on all the aspects that were investigated.[3] Thus, whereas Schwartz & Shine lay considerable stress on the fact that their corpus consists of two clearly distinguishable subcorpora, and therefore break their figures down for the two subcorpora and the various fragments of which each of them is composed, Quirk finds no differences between the three distinct bodies of material of which his corpus is composed. They are

1. "impromptu conversation recorded under friendly and informal conditions";
2. "also impromptu, but in this case the groups of participants knew that the proceedings were being recorded and that it was the intention that a shortened and tidied version should ultimately be broadcast";
3. "impromptu discussion on a platform in front of an audience and simultaneously broadcast" (cf. Quirk, 1957: 98-99).

The total material amounts to about sixteen hours of continuous, impromptu talk (± 150,000 words). About this material, Quirk writes:

> Each body of material was analysed separately to observe the extentto which the different conditions produced different grammatical phenomena, and to see whether or not there was a continued emergence of new configurations which would invalidate the size of each sample. The three revealed a striking uniformity: no one group of examples added significantly to the variety of structures present in either of the others, nor (still more importantly) did any reveal a different pattern of distribution for these structures. In short, whatever difference these different conditions evoke in speech (and some differences there certainly appear to be), they do not lie in the formation of relative clauses: the patterns here presented would seem to be distributionally complete and accurate for any body of educated spoken English of comparable size.
>
> (1957: 99)

Schwartz & Shine comment on this as follows:

> What aroused our suspicion in the first place was that Quirk did not allow for any sort of deviation from the "complete and accurate distribution that he found *in the full body* of his material." Intuitively we were of opinion that there would be distributional differences, even though we were not aware of the nature of these differences
>
> (1972: 18)

The material analysed by Schwartz & Shine consists of the following broadcasts from the BBC: eleven half-hour programmes from the series *Talking about Antiques*, which will be referred to as S/SAnt, and eleven half-hour programmes from the series *The Living World*, henceforth S/SLW, the total material covering approximately 10 1/2 hours of recorded speech, which is about 108,000 words.

Quirk (1957), having concluded that there are no significant differences between the three different bodies of texts making up his material, devotes the rest of his article to a description of the relation between syntactic environment and the occurrence of relative pronouns based on the examination of his *entire* corpus.

The written corpus that has been used for the present study comprises six 20,000 word samples of running text, taken from the following books: *The Mind Readers* (MR: crime fiction), *The Bloody Wood* (BW: crime fiction), *Eight Modern Writers* (EMW: literary criticism), *Techniques of Persuasion* (TP: a popularised sociological study), *Cell Biology* (CB: scientific writing), and *Stop It* (DR: drama).[4] Although BW and EMW belong to different registers, they represent the same idiolect, since they were both written by the same author. The written corpus comprises five different registers and thus lacks the unity of topic that can be observed in the two subcorpora in Schwartz & Shine. The only possible major subdivision that can be made in the written material as a whole would be, perhaps, that between fiction and non-fiction, the latter being represented by EMW, TP and CB.

From the Nijmegen corpus all noun phrases with postmodifying clauses were extracted, and were given an elaborate code for a more extensive study of postmodifying clauses in the English noun phrase (see de Haan, forthcoming, a). This body of data contains information about relative, appositive and other types of postmodifying clauses. It describes not only finite but also non-finite clauses. For a comparison with both Quirk's and Schwartz & Shine's material, a number of selections from this body of material had to be made for the present article, for the following reasons.

First of all, the distinction between WH-, THAT and ZERO in both Quirk and Schwartz & Shine makes it clear that only *finite* clauses are considered in those studies.[5] Secondly, neither Quirk nor Schwartz & Shine mention relative *adverbs*. Finally, in neither Quirk nor Schwartz & Shine is any mention made of the occurrence of *coordination* of relative clauses, whereas in the Nijmegen corpus 86 of the finite relative clauses were found to be coordinated with other postmodifying clauses or phrases, which in many cases involved ellipsis of the relative pronoun. This meant that the number of ZERO's in the written corpus was excessively high, as compared with that found in the two spoken corpora. Disregarding these cases brought the number of ZERO's down drastically (from 254 to 191). In one of the subcorpora (TP), the number was even halved: from 60 to 28. Seeing that coordination necessarily involves greater complexity, it can be appreciated that Quirk and Schwartz & Shine not only do not mention the phenomenon, but may not even have come across it in their corpora.[6] Cases of coordination have, therefore, also been disregarded. These considerations have led to a selection from the Nijmegen corpus of finite relative clauses which are *not* introduced by a relative adverb and which are *not* coordinated with other clauses or phrases.

Whereas Quirk distinguishes between *restrictive* and *non-restrictive* clauses, Schwartz & Shine do not. It is especially differences like these that made a comparison between the various corpora so difficult. Therefore we shall first dis-

cuss the occurrence of WH-, THAT and ZERO in relation to the texts in which they occur. This means a comparison of the various subcorpora in the Nijmegen corpus with those in Schwartz & Shine. Since Quirk does not break his figures down, his subcorpora will not be included in this comparison. After the comparison of the Nijmegen corpus with Schwartz & Shine's corpus we shall compare the Nijmegen corpus with Quirk's corpus, in order to discuss the relation between the occurrence of WH-, THAT and ZERO and the syntactic environment in which they occur. Since Schwartz & Shine do not break their figures down for those features, they will be absent there.

The written part of the Nijmegen corpus comprises roughly 120,000 words, which is approximately the average of Quirk's corpus (150,000 words) and Schwartz & Shine's (108,000 words). The overall percentages for WH-, THAT and ZERO found in the three corpora compared, are listed in table 1.

	SPOKEN		WRITTEN
	Quirk	S/S	Nijm
WH-	54	66	76
THAT	29	25	8
ZERO	17	9	16
Total	100	100	100

Table 1: Comparison between overall figures Quirk, S/S and Nijm, percentagewise

In the light of what grammarians have observed about the distribution of THAT, ZERO and WH- forms in written and spoken English,[7] it is not surprising that there is a relatively low number of THAT's in the written corpus. However, the fact that the number of ZERO's in that same corpus is almost as great as that found in Quirk's is unexpected. This will be gone into in the course of this article. It will also be noted that the figures found for WH- in the written corpus are higher than those found in either of the spoken ones.

A more detailed picture is given in table 2, where the figures obtained from the two different bodies of material in S/S are compared with the overall figures of Quirk and S/S, and with the figures of the subcorpora and the he overall figures of the Nijmegen corpus.

	S P O K E N				W R I T T E N						
	Quirk	S/SAnt	S/SLW	S/S overall	MR	BW	EMW	TP	CB	DR	Nijm overall
WH-	54	78	55	66	72	55	83	89	96	31	76
THAT	29	17	33	25	3	21	10	2	1	13	8
ZERO	17	5	12	9	25	24	7	9	3	56	16
Total	100	100	100	100	100	100	100	100	100	100	100

Table 2: Comparison between Quirk, S/SAnt, S/SLW, S/S overall, MR, BW, EMW, TP, CB, DR and Nijm overall, percentagewise

As in the material discussed in Schwartz & Shine, a great deal of variation can be observed in the written material. Thus the percentages for WH- vary from 31 to 96, those for THAT vary from 1 to 21 and those for ZERO from 3 to 56.

It is interesting to observe that EMW, TP and CB (all of them non-fiction) score very high on WH-, whereas the different percentages of THAT appear to be randomly distributed among the various corpus texts. A remarkably high score for ZERO is found in DR, a score which is matched in none of the other columns. That it is higher than the scores in any of the other written texts, is almost certainly due to the fact that this is a drama text, and as such, an attempt at representation of speech. That it is significantly higher even than that of either of the two spoken corpora, however, is probably explained by the fact that the drama text is intended to represent colloquial, uneducated English, which is in sharp contrast with the material collected by Quirk, for instance, who specifically calls his material "educated spoken English", and who writes:

> The speakers were English men and women, educated to university standard, and mainly between the ages of 25 and 50.
>
> (1957: 98)

Also noticeable is the position of the two other fiction texts (MR and BW). They occupy a position midway between the non-fiction fragments and the drama text, in that both contain fewer WH-'s than the non-fiction texts, but more than the drama text, more ZERO's than the non-fiction texts and fewer than the drama text.

Before we turn to the third table, which is still more detailed than table 2, we have to account for a few changes in the figures given by Schwartz & Shine (for Quirk's corpus similar figures are not available). In the first place, they distinguish between *who* and *whom*. The latter form, however, occurs only twice in their entire corpus, in which they have found 1097 finite relative clauses. Since the use of *whom*, rather than *who*, is rather predictable, and especially so in written English, the two have been treated as a single group, *who(m)*. In the larger body of data collected from the Nijmegen corpus, which was referred to earlier, the two are distinguished by means of another set of codes. For this comparison, the two occurrences of *whom* found in Schwartz & Shine have been added to the total number of *who*'s. A different attitude has been adopted towards *whose*, the use of which necessarily involves a greater complexity of the relative clause, and which, therefore, might reveal interesting differences between the various texts.

	S P O K E N		W R I T T E N					
	S/SAnt	S/SLW	MR	BW	EMW	TP	CB	DR
WH-								
which	333	220	48	49	134	193	178	13
who(m)	92	75	50	31	83	83	4	21
whose	4	5	6	-	13	3	2	3
WH- total	429	300	104	80	230	279	184	37
THAT	94	181	5	31	27	7	3	16
ZERO	27	66	36	35	20	28	5	67
Total	550	547	145	146	277	314	192	120

Table 3: Actual occurrences of relative pronouns

A second change concerns the occurrences of *which*. In the tables in Schwartz & Shine, a separate group is formed by "prep + *which*" as distinct from *which* not accompanied by a preposition. However, apart from merely being listed, the figures obtained for them are not used in any way. It might have been interesting to look at the relation of the various subcorpora to cases in which the preposition is put in final position and those where the preposition is kept in initial position, preceding the relative pronoun (see, eg de Haan, forthcoming, a). Since this was not done, their figures for *all* occurrences of *which* have been added up in this paper. Allowances have to be made for the fact that in table 3 only the totals of the two bodies of text in Schwartz & Shine are listed (each approximately 54,000 words) and the six subcorpora of the written corpus (each approximately 20,000 words).[8]

We notice that the lowest total number is found in the drama fragment (DR), apparently pointing to a relatively low number of embedded clauses generally and a large number of simple sentences. (The Nijmegen corpus is currently being used in two other projects, one of the aims of which is to run frequency counts on the occurrence of various syntactic structures. It will, no doubt, be possible to compare the various subcorpora with respect to complexity of the structures found.) The figures also show that the large number of ZERO's in DR is not matched in the (much larger) bodies of text in Schwartz & Shine. It can also be observed that apparently the use of THAT depends more heavily on the individual author than on a specific register.[9] The comparison is made easier when the numbers are reduced to percentages, as is done in table 4.

	S P O K E N		W R I T T E N					
	S/SAnt	S/SLW	MR	BW	EMW	TP	CB	DR
WH-								
which	60	40	33	34	48	62	93	11
who(m)	17	14	35	21	30	26	2	17
whose	1	1	4	-	5	1	1	3
WH- total	78	55	72	55	83	89	96	31
THAT	17	33	3	21	10	2	1	13
ZERO	5	12	25	24	7	9	3	56
Total	100	100	100	100	100	100	100	100

Table 4: Occurrences of relative pronouns, percentagewise

Table 4 clearly shows that DR is the only fragment in which the number of *who(m)*'s is considerably higher than the number of *which*'s, MR being the only text in which the number of *who(m)*'s is only slightly greater than the number of *which*'s. In all the other subcorpora the number of *which*'s is much larger. The extreme difference between the percentages of *which* and *who(m)* in CB is obviously due to the fact that this text contains scientific writing, figuring a large number of non-personal noun phrases.[10] In table 5, which is, in fact, a more detailed version of table 1, the differences between the average scores (percentagewise) of S/S and Nijm are given. It should be borne in mind that the averages for the Nijmegen corpus cannot directly be derived from table 4, but that the actual number of occurrences also have to be taken into consideration.

A comparison between tables 4 and 5 shows the relevance of breaking down figures for register. Especially the figures found for *which* are telling in this respect, there being considerable differences between the figures found in table 4, but hardly any between the two found in table 5.

	SPOKEN	WRITTEN
	S/S	Nijm
	%	%
WH-		
which	50.0	50.9
who(m)	15.5	23.1
whose	1.0	2.3
WH- total	66.5	76.3
THAT	25.0	7.4
ZERO	8.5	16.3

Table 5: Differences between the averages found for S/S and Nijm, percentagewise

It has been shown, not only by Schwartz & Shine, but also in the analysis of the written data provided above, that a distinction between various bodies of text is relevant and is supported by the frequencies found.

The rest of this paper is devoted to a comparison of the written corpus as a whole to Quirk's spoken corpus as a whole, in order to see how the written corpus differs from it as far as the syntactic phenomena investigated are concerned.

First of all, Quirk distinguishes between *restrictive* and *non-restrictive* clauses, as follows:

> Restrictive clauses ... are linked to their antecedents by close syntactic juncture, by unity of intonation contour, and by a continuity of the degree of loudness. In contrast, non-restrictive clauses are characterised by open juncture (recognised, together with the following features, by a comma in written materials), a fresh intonation contour, and a change (especially a diminution) in the degree of loudness.
>
> (1957: 101)

Since Quirk does not break down WH-figures for *which*, *who(m)* and *whose*, they can only be compared with the overall figures as they are presented in tables 1 and 2 in this paper. However, in the discussion of his table Quirk mentions a number of features that seem worth comparing to the results of the analysis of the written data.

First of all, the percentage of non-restrictive clauses is far higher in the latter corpus: 28,1%, as compared to Quirk's 13.4%. Since Quirk, in his subsequent analyses compares WH-, THAT and ZERO, the latter of which does not occur at all in non-restrictive clauses, it has been assumed that the non-restrictive clauses have been disregarded in those analyses, and, consequently, they have also been disregarded in the written corpus.

Quirk finds that in less than 1% is the relative pronoun ZERO if it is the subject of the relative clause and if the antecedent is personal, whereas the THAT's make up almost 9% and the WH-'s 91%. A comparison with the written corpus shows that there is an even greater tendency to use WH- forms in written language: 0% ZERO, 5% THAT and 95% WH-. Tables 6 - 9 give a comparison of the spoken corpus to the written.

	SPOKEN		WRITTEN	
	Quirk		Nijm	
	n	%	n	%
ZERO	1	-	-	-
THAT	19	9	8	5
WH-	200	91	151	95

Table 6: Distribution of relative pronouns in restrictive relative clauses, in which the relative pronoun functions as the subject and the antecedent is personal

What is immediately striking is the fact that though Quirk remarks that the number of WH-'s in his material is far larger than that of THAT's, notwithstanding "the persistent observation in grammars since the time of Sweet that in spoken English *that* and zero constructions are preferred to *wh*-forms" (p. 106), the number of THAT's (both in absolute terms and percentagewise), is greater in the spoken material than in the written, in all of the four tables given here. The differences are particularly striking in table 7, where, in an almost identical total number, there are almost one hundred more THAT's in the spoken corpus than in the written, and in table 8, where there are no THAT's at all in the written corpus.

	SPOKEN		WRITTEN	
	Quirk		Nijm	
	n	%	n	%
ZERO	1	0.3	1	0.3
THAT	152	52.1	59	20.1
WH-	139	47.6	233	79.6

Table 7: Distribution of relative pronouns in restrictive relative clauses, in which the relative pronoun functions as the subject and the antecedent is non-personal

	SPOKEN		WRITTEN	
	Quirk		Nijm	
	n	%	n	%
ZERO	11	34	13	65
THAT	10	32	-	-
WH-	11	34	7	35

Table 8: Distribution of relative pronouns in restrictive relative clauses, in which the relative pronoun functions as the object and the antecedent is personal

A comparison of tables 6 and 8 reveals the curious fact that if the antecedent is personal, no THAT's occur in the written material when the pronoun is the *object* of the relative clause, but that THAT *is* sometimes used when it is the *subject*. Given the three alternatives for 'object' (see table 8), THAT is apparently "skipped" in favour of ZERO.[11] For 'subject' there are in reality only two alternatives: THAT and WH-.

A look at table 9 reveals a relative absence of THAT's and WH-'s in the written corpus if the relative pronoun is the object in the clause and if the antecedent is non-personal, for, whereas the actual number of ZERO's is almost identical in the spoken and written corpus, those of WH- and THAT are far lower in the written corpus.

	SPOKEN		WRITTEN	
	Quirk		Nijm	
	n	%	n	%
ZERO	89	36	87	69
THAT	94	38	14	11
WH-	65	26	25	20

Table 9: Distribution of relative pronouns in restrictive relative clauses, in which the relative pronoun functions as the object and the antecedent is non-personal

Another feature that Quirk draws attention to is the number of restrictive clauses that are "distanced from" their antecedents. He points out that normally the relative clause immediately follows its antecedent, but that there are cases in which this may be inconvenient, and a certain distancing of the parts is required. The figures found are given in table 10:

	S P O K E N		W R I T T E N	
	Quirk (total: 1124)		Nijm (total: 859)	
	n	%	n	%
ZERO	6	0.5	2	0.2
THAT	48	4.8	2	0.2
WH-	87	7.7	12	1.3

Table 10: "Distancing" of relative clauses from their antecedents, related to the total number of restrictive relative clauses

If we look at the number of relative clauses that have been moved away from a position immediately following the antecedent, we see that the numbers are higher in the spoken corpus, especially for THAT and WH-. In table 11, which lists the total number of restrictive relative clauses as well as the number of distanced clauses, this becomes even more clearly visible:

	S P O K E N			W R I T T E N		
	Quirk			Nijm		
	total	dist.	%	total	dist.	%
ZERO	228	6	2.6	191	2	1
THAT	372	48	12.9	89	2	2.2
WH-	524	87	16.6	579	12	2

Tabble 11: "Distancing" of relative clauses from their antecedents, related to the number of ZERO's, THAT's and WH-'s, respectively

It may be concluded from this comparison that in spoken English the relative clause is less apt to follow its antecedent immediately than in written English. This would seem to indicate a preference in spoken English for a final position of the relative clause in its superordinate sentence (or clause).[12] However, this is not borne out by a comparison of the number of clauses with end position in the sentence with those that have a non-final position. For the comparison in table 12, relative clauses with non-personal antecedents have been chosen in which the relative pronoun functions as the object.[13]

	SPOKEN					WRITTEN				
	Quirk					Nijm				
	total	n-fin		final		total	n-fin		final	
		n	%	n	%		n	%	n	%
ZERO	137	48	35	89	65	121	34	28	87	72
THAT	135	41	30	94	70	16	2	13	14	87
WH-	72	7	10	65	90	28	3	11	25	89

Table 12: A comparison of the number of final and non-final relative clauses, in which the relative pronouns function as objects, and with non-personal antecedents

If we look at the percentages, we see that the tendency in both the written and the spoken corpus seems to be the same (most non-finals with ZERO, most finals with WH-). It can also be seen that the percentages for final positions are not higher in the spoken corpus (with ZERO and THAT they are even lower) than in the written corpus. This means that apparently in the written corpus there is a tendency to use noun phrases with clausal postmodification in final positions generally, rather than having to revert to either using medial positions for relative clauses, which in certain cases would be awkward, or to moving the clauses away from the antecedents, by which operation the unity of the head and its postmodifier would be lost. This can onlybe explained by the fact that in spoken language a speaker, once he has started a sentence has to finish it as best he can, whereas in writing, a sentence can be rewritten so that a noun phrase postmodified by a relative clause gets a final position in the sentence. It is also remarkable that in speech non-final relative clauses often result in anacolutha, as Quirk has found (1957: p.104).

The conclusion that we can draw from this is that clauses are generally considered to interfere in the structure of the superordinate clause or sentence, and that therefore they are preferred in final position. In speech this results in a relatively great number of cases in which the clause is separated from its antecedent, whereas in written language sentences are generally constructed in such a way as to avoid separation as much as possible, by giving those noun phrases with clausal postmodification final positions in the sentence.[14]

Finally, some interesting differences can be observed in connection with the length of the relative clauses investigated. The figures provided by Quirk are broken down for clauses with non-personal antecedents, in which the relative pronoun functions either as the subject or as the object of the relative clause. The same division has been made for the written data as well. From table 13 it can be concluded that the clause length in the written corpus is considerably greater than in the spoken corpus.[15]

	S P O K E N		W R I T T E N	
	Quirk		Nijm	
	spread	average	spread	average
THAT	3 - 7	4.9	3 - 7	6.4
WH-	2 -11	6.3	2 -17	9.5

Table 13: Clause lengths compared, non-personal antecedents, relative pronoun subject

	S P O K E N		W R I T T E N	
	Quirk		Nijm	
	spread	average	spread	average
ZERO	2 - 3	4	2 - 6	4
THAT	2 - 5	5	2 - 6	5
WH-	2 - 9	6	2 - 9	8

Table 14: Clause lengths compared, non-personal antecedents, relative pronoun object

There is a difference between the spoken corpus and the written corpus, although it is not so great in table 14 as in table 13. The main difference appears to lie within the WH-'s, where the clause contains two more words in the written material on an average. Although there is not much difference between either spread or averages (by spread is meant that area of clause length where most clauses are found), Quirk remarks that, for instance for WH- in the last table the peak (the *mode*) is found at a length of 5 words, whereas in the written corpus the peak is found, curiously enough, at 12 words, which is well outside the area of the thickest concentration.

Unfortunately, Quirk does not break his figures down for indefiniteness or definiteness of the noun phrase heads, which might have revealed some interesting differences. At the moment a publication is being prepared in which the figures of the written material described in this article are discussed in relation to this feature (see de Haan, forthcoming, b).

The discussion of the results of the analyses of the various fragments in both Schwartz & Shine and the written corpus has shown clearly that there are dif-

ferences between medium (spoken and written English) and register (the topics dealt with in the various fragments). It would, therefore, seem that Schwartz & Shine are justified in their scepticism as to Quirk's claims that no differences were found in the various bodies of text of which his corpus was composed.

A comparison of Quirk's material and that found in the written corpus, however, has shown that although there are certain differences between spoken and written English, especially with reference to clause length and the use of relative pronouns, in general the same syntactic restrictions apply to both spoken and written English with respect to relative clause formation. However, the same restrictions tend to result in the choice of different syntactic alternatives, as can be seen, for instance, if we look at the position of noun phrases postmodified by clauses, and at the number of cases in which the relative clause is "distanced" from its antecedent.

Schwartz & Shine remark that "an investigation into possible 'profiles' in written English would appear to be long overdue" (1972: p.28). It is hoped that this article has contributed towards a description of such 'profiles'.

NOTES

1. I am much indebted to Jan Aarts and Flor Aarts for reading an earlier version of this paper and giving many valuable comments.

2. In their article, Schwartz & Shine compare their results with Quirk's findings.

3. A fourth study of the same grammatical phenomenon (Taglicht, 1973) has been excluded from the comparison, since, although it is based on a collection of written sentences, it is not *corpus*-based.

4. This play, by Henry Livings, amounts to a total of 15,000 words. In order to complement the 20,000 words, which facilitates comparison with the other samples, the first 5,000 words of *Nil Carborundum*, another play by the same dramatist, have been added. The two fragments are here used as one sample and referred to as DR.

5. Non-finite clauses are usually *not* introduced by relative pronouns.

6. Spoken language is assumed to be less complex than written language.

7. This point is also made by Quirk (1957), who, notwithstanding this, finds that in his material the number of WH-'s is larger than that of THAT's and ZERO's put together.

8. For Quirk's corpus no similar breakdown was available.

9. This is borne out by a comparison of BW and EMW, both written by the same author, which show a remarkable similarity in the actual number of THAT's.

10. The scientific character of this text and the impersonal style generally is borne out by the fact that in 26% of all the finite postmodifying clauses in this text a passive verb phrase occurred, which is a much larger proportion than is seen in any of the other texts.

11. Taglicht (1973), who also studies the occurrence of relative pronouns in written English, obtains similar results from his data.

12. Quirk does not indicate whether *final* and *medial* mean final and medial in the sentence or in the clause; it has been assumed that the latter is the case.

13. Quirk points out the unsuitability of other antecedents and other functions of the pronoun, since they already heavily restrict the occurrences of THAT and ZERO.

14. Unfortunately, Quirk does not provide any information about the syntactic function of the antecedent. In de Haan (forthcoming, a) this feature will be taken into consideration, to see whether there is a relation between the position of the relative clause and the function of the antecedent.

15. In this and the following table the relative pronouns themselves have not been included in the count.

REFERENCES

- Aarts, J. & T. van den Heuvel (1980). 'The Dutch Computer Corpus Pilot Project', *Icame News*, 4: 1 - 6.
- Haan, P. de (forthcoming, a). *Postmodifying Clauses in the English Noun Phrase.*
- Haan, P. de (forthcoming, b). 'On the Distribution of Relative Clauses with Indefinite Noun Phrase Heads'.
- Quirk, R. (1957). 'Relative Clauses in Educated Spoken English', *English Studies*, 38: 97 - 109.
- Schwartz, H. & N. M. Shine (1972). 'Relative Clauses Revisited', *CEBAL. Publication of the Copenhagen School of Economics and Business Administration Language Department*, 2: 16 - 28.
- Taglicht, J. (1973). 'The Choice of Relative Pronouns in Written English', *Scripta Hierosolymitana*, 25: 327 - 336.

VERB AND PARTICLE COMBINATIONS:
PARTICLE FREQUENCY RATINGS AND IDIOMATICITY

Eric Akkerman
University of Amsterdam

INTRODUCTION

This article presents an outline of my graduate thesis on verb particle combinations, more specifically phrasal verbs[1] and verbs with two particles. The main source of data for my research was the 'tagged' version of the *Brown corpus*, in which each word (token) has been given a grammatical tag.

The aim of my research was threefold. First, I wanted to establish frequency ratings of particles. Secondly, I intended to analyse the idiomaticity of the verbal combinations found in the corpus. Finally, I wanted to see how phrasal verbs (idiomatic and non-idiomatic) and verbs with two particles were distributed in the varieties of language represented by the sub-corpora of the *Brown corpus* and find out which factors played a role in this distribution. For each of the three aspects of my research I made use of sub-corpora A-J[2], which together consist of 510,543 words. In order to elicit my data, I used the 'Query'-system developed by v.d. Steen (cf. v.d. Steen, 1982 and Meys, 1982).[3]

METHOD

As RP is the tag for adverb particles, and VB denotes all verbal forms, I decided to start with the following query-pattern:

(1) * VB * RP

This pattern would yield all 'simple' transitive and intransitive phrasal verbs. The output would also contain all verbs with two particles, whether active or passive. In order to obtain those cases in which the particle occurred in front position and also for more complex instances like

(a) he saw the children across

(where there is a wide range of possibilities for filling the object slot), I intended to use the query-pattern[4]

(2) * RP .EN. .NIET. * VB * RP

However, when the corpus was searched for the two patterns mentioned above, I ran up against two major problems:

1. For the entire corpus the results for the first pattern were satisfactory in that it gave a fairly large but correct output. For the sub-corpora after J, however, the output was suddenly almost nil. It appeared that in these sub-corpora the particles were tagged RB (general adverb). After checking A-J with this knowledge in mind, it appeared that there too a (relatively small) number of particles was *sometimes* tagged RB. The reason for this has never become clear to me. When later, in order to obtain (in combination with pattern (1)) a complete output of phrasal verbs, the corpus was searched for the pattern

   (3) * VB * RB

   8,168 instances were found which, unfortunately, included far too much unwanted material (e.g. *I came early, he ran fast*)

2. The second query-pattern (* RP .EN. .NIET. * VB * RP) did *not* work: contrary to the expected output the actual output *did* contain instances of * VB * RP. Cooperation with staff of the Computer Department in an attempt to come to an alternative formulation of the pattern did not yield any positive results.

In view of these problems I decided to undertake a completely different approach, taking *specific words* as a starting-point. Since the total number of combinations of any verb with any particle is enormous, a selection had to be made. It seemed logical to focus on the particle because phrasal verb particles form a closed system and are relatively few in number, whereas verbs form an open class so that their number is much larger.

Many scholars have compiled lists of the particles which in their opinion are the most important ones. Combining some of these lists resulted in a list of 36 particles that would be used in the research:

aback	about	above	across	after	ahead
along	alongside	(a)round	aside	away	back
behind	below	beneath	between	beyond	by
down	forth	forward	in	inside	off
on	opposite	out	outside	over	past
through	to	under	underneath	up	without

The particles were incorporated in query-patterns of the following type:

UP RP .OF. UP RB

where *up* can be replaced by any of the other particles. Thus all instances of a particle were located whether tagged RP or RB. This pattern can be simplified to

(5) UP [# RP, # RB]

Here the square brackets indicate internal specification of the grammatical tag. The hash-symbol indicates the code-boundary and the comma distinguishes between alternatives. (cf. Meys, 1982: p.41) Because of the possible co-occurrence of these particles with verbs, there was no need to include * VB in the pattern. Apart from the kind of instances searched for, the output will also include instances like

(b) Away he went
(c) He saw (complex) NP across

Strictly speaking it is possible to use a query-pattern which can be expected to be much more efficient since it includes more than one particle. This pattern would look more or less as follows:

(6) [#OUT#, #UP#, #OFF#, #...#, #...#, etc.] [#RP, #RB]

Notwithstanding this possiblitity I preferred the approach mentioned earlier i.e. to concentrate on the separate particles rather than the lot.[5] The advantage of this approach is of course that the separate outputs give a clear impression of the frequency of the particles as well as their distribution over the subcategories.

MAIN RESULTS AND OBSERVATIONS

The approach outlined above proved to be satisfactory. The phrasal verbs wanted were yielded with great accuracy while at the same time the percentage of unwanted material in the output was relatively low.

As far as my third aim is concerned (the distribution of phrasal verbs in different varieties of language), the results of the analysis seem to point to the existence of a relation between the formality/informality of a discourse and the percentage of (idiomatic) phrasal verbs occurring in that discourse. If a variety contains variables that point to informal use[6], the percentage of (idiomatic) phrasal verbs in that variety will be much higher than if these variables point to more formal use. As this conclusion remained rather tentative, I shall not go into this aspect any further here, but focus on the results of the other two.

GENERAL RESULTS

Of the original list of 36 particles (see above), the following did not occur as part of a phrasal verb in the corpus I dealt with:

aback	after	beneath	between
beyond	opposite	past	to
under	underneath	without	

For the remaining 25 particles the following results were obtained:

particle	total	as first particle in phrasal verbs (verb + 1 particle)		as first particle in verb + 2 part. combinations		overall total
		absolute	relative(%)	absolute	relative(%)	
up	474	416	23.4	58	34.1	24.3
out	450	438	24.6	12	7.1	23.1
off	131	124	7.0	7	4.1	6.7
back	130	114	6.4	16	9.4	6.7
down	129	115	6.5	14	8.2	6 6
on	127	118	6.6	9	5.3	6.5
in	124	106	6.0	18	10.6	6.4
away	86	73	4.1	13	7.6	4.4
over	81	74	4.2	7	4.1	4.2
forth	26	25	1.4	1	0.6	1.3
about	25	25	1.4	-	--	1.3
along	25	23	1.3	2	1.2	1.3
forward	25	20	1.1	5	2.9	1.3
around	24	21	1.2	3	1.8	1.2
aside	24	24	1.3	-	--	1.2
ahead	20	17	1.0	3	1.8	1.0
by	15	15	0.8	-	--	0.8
through	14	12	0.7	2	1.2	0.7
across	6	6	0.3	-	--	0.3
inside	5	5	0.3	-	--	0.3
behind	3	3	0.2	-	--	0.2
outside	3	3	0.2	-	--	0.2
alongside	2	2	0.1	-	--	0.1
above	1	1	0.1	-	--	0.1
below	1	1	0.1	-	--	0.1
TOTAL	1951	1781	100.0	170	100.0	100.0

Figure 1

Looking at the frequency distributions of particles in phrasal verbs and verb + two particle combinations, the following six frequency categories can be distinguished:

I out,up (24.3 - 23.1%)
II off, on down, back, in (6.7 - 6.4%)
III over, away (4.4 - 4.2%)
IV about, forth, aside, along, around, forward, ahead (1.3 - 1.0%)
V by, through (0.8 - 0.7%)
VI across, inside, behind, outside, alongside, above, below (0.3 - 0.1%)

It is interesting to compare these results with the work done by Makkai[7], on the basis of which a list can be made up in which particles are ranked according to their potential number of *different* verb + particle (+ particle) combinations in which they can occur. The relevant particles can be listed as in figure 2.

particle	number of possible verbal combinations for each particle (according to Makkai)		frequency class in Brown Corpus
	absolute	relative(%)	
up	143	16.0	I
out	117	13.1	I
off	74	8.3	II
down	72	8.1	II
in	70	7.9	II
over	*63*	*7.1*	*III*
on	61	6.8	II
back	59	6.6	II
through	54	6.1	V
away	42	4.7	III
across	39	4.4	VI
along	33	3.7	IV
by	24	2.7	V
about	23	2.6	IV
behind	17	1.9	VI
TOTAL	891	100.0	

Figure 2

As far as the first two categories are concerned, it seems that there is a relation between the potential number of verb + particle (particle) combinations on the one hand, and the number of combinations that is actually found in the corpus on the other. Only *over* contradicts this assumption. For the last four categories of particles, however, there is no relation at all between Makkai's figures and the actual number of occurrences in the corpus. The main reason for this will be that Makkai mentions combinations in which the particle is a prepositional adverb as part of an intransitive phrasal verb with the object deleted or understood. Compare

(d) let me through
(e) help her through

where in a specific case the object will be understood as eg *the gateway* or *the hatchway*. In the actual corpus, however, relatively few of these forms were found.

IDIOMATICITY

First the results of the phrasal verb, presenting the number of idiomatic and non-idiomatic phrasal verbs for each particle:

particle	total	non-idiomatic		idiomatic	
		absolute	relative(%)	absolute	relative(%)
out	438	91	20.8	347	79.2
up	416	111	26.7	305	73.3
off	124	55	44.4	69	55.6
on	118	45	38.1	73	61.9
down	115	34	29.6	81	70.4
back	114	86	75.4	28	24.6
in	106	37	34.9	69	65.1
over	74	9	12.2	65	87.8
away	73	42	57.5	31	42.5
about	25	5	20.0	20	80.0
forth	25	7	28.0	18	72.0
aside	24	8	33.3	16	66.7
along	23	16	69.6	7	30.4
around	21	15	71.4	6	28.6
forward	20	11	55.0	9	45.0
ahead	17	12	70.6	5	29.4
by	15	8	53.3	7	46.7
through	12	4	33.3	8	66.7
across	6	3	50.0	3	50.0
inside	5	5	100.0	-	----
behind	3	3	100.0	-	----
outside	3	3	100.0	-	----
alongside	2	2	100.0	-	----
above	1	1	100.0	-	----
below	1	1	100.0	-	----
TOTAL	1781	614	34.5	1167	65.5

Figure 3

It is interesting to see that of the total number of phrasal verbs 65.5% is idiomatic and only 34.5% non-idiomatic. This is especially striking because Makkai's work shows that the potential number of different non-idiomatic verbal compounds is much higher than that of idiomatic ones (figure 4).

particle	total	non-idiomatic		idiomatic	
		absolute	relative(%)	absolute	relative(%)
up	143	58	40.6	85	59.4
out	117	59	50.4	58	49.6
off	74	39	52.7	35	47.3
on	61	34	55.7	27	44.3
down	72	42	58.3	30	41.7
in	70	42	60.0	28	40.0
through	54	37	68.5	17	31.5
over	63	44	69.8	19	30.2
away	62	31	73.8	31	26.2
back	59	47	79.7	12	20.3
about	23	19	82.6	4	17.4
across	39	33	84.6	6	15.4
by	24	21	87.5	3	12.5
along	33	30	90.9	3	9.1
behind	17	17	100.0	-	---
TOTAL	891	553	62.1	338	37.9

Figure 4

Only with *up, back, behind, along* and *away* does the relation of possible idiomatic/non-idiomatic phrasal verbs weakly correlate with the relation of idiomatic/non-idiomatic phrasal verbs actually found in the corpus. Notable differences are found for *over, about, out, down, aside* and *in*: here the percentage of non-idiomatic phrasal verbs actually found is much lower than the percentage of possible non-idiomatic combinations. An explanation may again be that most of these particles are prepositional adverbs (see above): in many literal senses they *may* occur with an adverbial function, but in an actual corpus they are not very frequent. Perhaps a study based on a corpus of spoken English would show different results.

Finally, the particles *inside, behind, outside, alongside, above* and *below* were never part of an idiomatic phrasal verb at all. As far as the idiomaticity of verbs with two particles is concerned, I have followed the approach of Makkai (1972), who centres upon the bilexonic phrasal verb (i.e. the verb + the first following particle). If that is non-idiomatic, and the additional particle makes the combination idiomatic, the new particle is called a compulsory idiom bridge (CIB). For example, *for* is called a CIB in case of *go in for* where *go in* is non-idiomatic, whereas *go in for* is idiomatic. If the bilexonic verb *is* idiomatic, the added particle is an optional idiom bridge (OIB) as for instance in

fall out (idiom. = 'quarrel') → he fell out with his sister
(*with* = OIB)

Figure 5 shows the relation of phrasal verbs with CIB/phrasal verbs with OIB, for each first particle:

	Total	with CIB		with OIB	
		absolute	relative(%)	absolute	relative(%)
up	58	24	41.4	34	58.6
in	18	8	44.4	10	55.6
back	16	2	12.5	14	87.5
down	14	10	71.4	4	28.6
away	13	7	53.8	6	46.2
out	12	2	16.7	10	83.3
on	9	3	33.3	6	66.7
off	7	-	----	7	100.0
over	7	2	28.6	5	74.4
forward	5	-	----	5	100.0
ahead	3	-	----	3	100.0
around	3	3	100.0	-	----
along	2	-	----	2	100.0
through	2	-	----	2	100.0
forth	1	-	----	1	100.0
TOTAL	170	61	35.9	109	64.1

Figure 5. Verbs with two particles; the relation CIB/OIB, for each first particle

On the whole there are far more verb + two particle combinations with an OIB than a CIB. Only for combinations with *around, down,* and *away* (as first particle) there are more CIBs than OIBs. The most frequent verb + two particle combinations with a CIB are those with *up* and *down* as the first particle (together

with 55.7% of total) and to a smaller extent also those with *in* and *away* (together 24.6% of total). The most frequent combinations with an OIB are those with *up* as first particle (31.2% of total). Most frequent combinations with an OIB are those with *up* as first particle (31.2% of total) and to a smaller extent also those with *in, back* and *out* (together 22.0% of total).

It is also interesting to take a closer look at the *second* particles, which function as the idiom bridges in these combinations (fig.6).

	Total		functioning as CIB		functioning as OIB	
	abs.	rel.(%)	absolute	relative(%)	absolute	relative(%)
to	63	37	18	29	45	71
with	36	21	21	58	15	42
on	18	11	6	33	12	67
off	17	10	1	6	16	94
for	8	5	6	75	2	25
at	2	1	1	50	1	50
about	1	1	-	--	1	100
under	1	1	-	--	1	100
others	24	14	8	33	16	67
TOTAL	170	100	61	36	109	64

Figure 6. Percentages of particles functioning as idiom bridges (i.e. the second particle of the verb + two particle combinations). There are 63 verb + two particle combinations with *to* as second particle (37% of total); of these, 29% functions as a CIB, and 71% functions as an OIB.

	CIBs	OIBs
to	29%	41%
with	34%	14%
on	10%	11%
off	2%	15%
for	10%	2%
at	2%	1%
about	--	1%
under	--	1%
others	13%	15%
	100%	100%

Figure 7. Contribution of each particle to the total number of CIBs and OIBs.

It is clear that *to* is the most frequent second particle, followed at a distance by *with, on* and *off*. The most important particle functioning as a CIB is *with*, followed closely by *to*. The most important particle functioning as an OIB is *to*, followed at a great distance by *off, with* and *on*.

FINAL REMARKS

I think that most of the results reported in this article speak for themselves, and on the whole are a fairly reliable source for further study of the phrasal verb. As I have based my research on an American corpus only, it would be interesting to see the results of parallel research based on a British English corpus (e.g. LOB or CCPP).

As far as particle frequencies and idiomaticity of phrasal verbs and verb + two particle combinations are concerned, it would be nice to see if my results would be confirmed. It would, however, especially be interesting from the viewpoint of the study of varieties of English (which has been almost ignored in this article), because it might show possible differences between the distribution of (idiomatic) phrasal verbs in American and British English.

NOTES

1. I consider a phrasal verb to be a grammatical construction consisting of a verbal part plus an adverb particle (i.e. a particle with an adverbial function); as such it is defined by contrasting it with prepositional verbs (or verbs + prepositional phrase); cf. Cowie and Mackin (1979).

2. The sub-corpora are the following:

```
A = Press/reportage          E = Skills and hobbies
B = Press/editorial          H = Miscellaneous (mainly government
C = Press/reviews                documents)
D = Religion                 J = 'Learned'
```

3. Explanation of 'Query' symbols used in the match programs:

```
.EN.    input operator meaning "and"
.NIET.  input operator meaning "not"
.OF.    input operator meaning "or"

*  indicates 'any word'; e.g. VB denotes 'any word with tag VB'
#  indicates word boundary and sentence boundary
%  indicates "any code'; e.g. UP % yields up as RB, as RP,
   as IN, as JJ (adjective), etc.
-  indicates 'any sequence of words'
```

   As for the Brown tagging system see Francis and Kucera, 1979. Relevant Brown tags will be explained as we go along.

4. This match program seemed much simpler, but probably better than the separate inputs:

* VB - * N - * RP .EN. .NIET. ^ * VB - * VB - * N - * VB - * RP

(indicating that some kind of NP must occur between the verb and the particle) and:

* VB * P * RP

(indicating that a pronoun must fill the object slot). At the same time it makes the following pattern

* % * RP - * VB

(where the first * corresponds with the code number of the sentence) to account for cases with front position of the particle superfluous.

5. I used lists from Bolinger (1971), Close 1975), Fraser (1976), Hill (1968), Kennedy (1920), Live (1965), Makkai (1972) and Sroka (1972).

6. Random tests showed that, apart from the confusing use of the tags RP and RB, very few mistakes were made in the tagging. When TROUGH IN (*through* as a preposition) was checked, for example, it was found that of the 415 instances in subcategories A-J, only one had the wrong tag.

7. Makkai (1972). It should be noted that Makkai's approach differs from mine on three points:

   a. in my reseach one specific phrasal verb is counted as many times as it occurs in the corpus; Makkai counts e.g. *to set up* only once, as one *possible* combination with *up*;

   b. Makkai includes prepositional verbs in his research (he ignores any distinction between adverbs and prepositions). These, however, constitute a minority;

   c. Makkai not only works with a fixed set of 25 formants (23 of which were incorporated in my list) but with a fixed set of (100) verbs as well.

8. I regard an indiom as a structure of two or more free components, whose meaning cannot be understood (solely) on the basis of knowledge of the individual parts.

REFERENCES

- Bolinger, D. (1971), *The Phrasal Verb in English*, Cambridge.
- Cowie, A.P. and R. Mackin (1979), *Oxford Dictionary of Current Idiomatic English*, Vol.1. Oxford.
- Close, R.A. (1975), *A Reference Grammar for Students of English*. London.

- Francis, W.N. and Kučera, H, (1979), *Manual of Information* to accompany Standard Corpus of Present-Day Edited American English, for use with Digital Computers; Revised and Amplified Edition (original edition 1964). Providence, Rhode Island, Brown University Press.
- Fraser, B. (1976), *The Verb-Particle Combination in English*, New York.
- Hill, L.A. (1968), *Prepositions and Adverbial Particles*, London.
- Johansson, S. ed. (1982), *Computer Corpora in English Language Research*, Bergen, Norwegian Computer Centre for the Humanities.
- Kennedy, A.G. (1920), *The Modern English Verb-Adverb Combination*. Stanford University Publication in Language and Literature, Vol.1, No.1. Stanford California.
- Live, A.H. (1965), 'The Discontinuous Verb in English'. in: *Word*, Vol. 21, No. 3, 428-451.
- Makkai, A. (1972), *Idiom Structure in English*, The Hague.
- Meys, W. (1982), 'Exploring BROWN with QUERY'. in: S. Johansson, ed., (1982)
- Sroka, K.A. (1972), *The Syntax of English Phrasal Verbs*.The Hague.
- Steen, G.J. van der (1982), 'A Treatment of Queries in Large Text Corpora'. in: S. Johansson, ed., (1982).

TO STRAND OR NOT TO

Inge van den Hurk
Linda Kager
Lois Kemp
Marjolein Masereeuw

University of Amsterdam

This paper, which tries to shed some light on stranded-*to* cases, is based on findings in LOB. For information concerning stranded *to* we have focussed mainly on Zwicky's article *Stranded to and the Phonological Phrasing in English* and a comparison of our findings with Zwicky's will be presented here. In our analysis of stranded *to* we have looked at the following points:

- the nature of the deleted material;
- the context in which stranded *to* occurred and the material to which it is attached.

As a great deal of data in this paper refers back to Zwicky's article a short summary of this article will be given first.

The English infinitive marker *to* usually forms a phonological phrase with its following verb phrase, but sometimes it is stranded from its VP and phrases instead with preceding material, as in "I'd hate to". Within the general area of phonological phrasing the analysis of words which are prosodically subordinate to neighbouring material is dealt with in this article. These words are called "leaners" by Zwicky: they form a rhythmic unit with the neighbouring material, are normally unstressed with respect to this material and do not bear the intonational peak of the unit. The infinitive marker *to* is such a leaner. If it is stranded it *must* have material to attach to, because it cannot stand alone:

Do you want to leave or not to?
Not to.
* To.

However, *to* will not attach to just anything and Zwicky's main concern is to define the conditions under which it can be stranded.

There are two means by which *to* can be stranded:

a. Parenthetical material

(1) "I made a decision to - and he made a decision, too - have him in the group."

b. Verb Phrase Deletion

(2) "I wanted to, but was afraid to."

Zwicky states that VPD with *to* is strictly conversational, its natural habitat being informal two-person discourse. From his research he draws the following conclusions:

i. NOT seems to be the only preceding adverb to which *to* attaches with any frequency;

ii. *to* attaches freely to preceding V + Pro combinations, but very infrequently to V + full NP object;

iii. impersonal constructions are much less receptive to stranded *to* than personal constructions;

iv. stranded *to* only occurs in a small number of the total of constructions possible with non-stranded *to*;

v. missing are absolutes, exclamations, idioms, subject complements in subject position.

Conditions that Zwicky mentions for stranded *to* include:

i. *To* reattachment or 2R: when it does not form a VP constituent with an immediately following NP, the English infinitive marker *to* attaches to the constituent immediately to its left, to form a phonological phrase with it;

ii. (The Own-S Condition or OSC) ... except that it cannot move out of its surface structure; ... and except that *to* will not attach to a form of *be*;

iii. Stranded Filter or SF: structures containing VP exhaustively dominating the infinitive marker *to* are unacceptable;

iv. Any surface structure in which some node other than VP exhaustively dominates *to* is unacceptable;

v. The constituent preceding *to* must be a VP or a predicator in a VP;

vi. The more a construction describes an action or attitude directed towards some goal, the more acceptable it is when followed by stranded *to*;

vii. An exception clause to v: readjustment is always possible if the word preceding *to* is a monosyllabic non-lexical item;

ii. Revised: readjustment rules cannot move an item so that it would no longer be dominated by its S.

In his article Zwicky devotes a lot of attention to phonological phrasing of stranded *to* which we will not touch on in this paper.

We have pointed out that our methods differ from Zwicky's in one essential aspect: while Zwicky bases his findings on informants' reactions to artificial data and the examination of instances found in texts, we restricted ourselves to the latter type of data, i.e. the occurrences of stranded *to* in the LOB corpus. It is difficult to say whether and to what degree our respective methods account for any discrepancies between our findings and Zwicky's.

We can by no means claim that our list of instances of VPD is complete. The process of elicitation was impeded by the absence of grammatical tagging in the LOB corpus. In order to avoid having to sift the elicited materials by hand we restricted the match patterns to: *to* plus delimiter, *to* plus conjunctions *yet*,

but, which meant that our data would only consist of examples of clause final VPD with *to*. Match patterns that produced no results were:

```
to % : %
to % ? %
to % yet %
to % [*-] %
to % [(] %
```

Successful match patterns were:

to % but %	1 instance
to % [)] %	1 instance
to % [.] %	29 instances
to % [,] %	16 instances
to % ! %	1 instance

This means that we elicited 48 examples of clause final VPD with *to* out of a corpus containing 500 printed texts of about 2,000 words each.

The distribution of these instances throughout the corpus was not very even. Our match programs extracted no cases of VPD with *to* from text categories B - Press: editorials, D - Religion, E - Skills, trades and hobbies and J - Learned and scientific writings. The following list indicates the number of VPD with *to* instances found in the other text categories.

Text category	Subcategory	No. of instances
A - Press: reportage	National Daily, Sports	1
C - Press: reviews	National Sunday	1
F - Popular lore	Popular politics, psychology, sociology	4
G - Belles lettres, biography, essays	Biography, memoirs	3
	Literary essays and criticism	3
	General essays	2
H - Miscellaneous	Government documents: Proceedings, debates	2
K - General fiction	Novels	3
L - Mystery and detective fiction	Novels	6
	Short stories	1
M - Science fiction	Short stories	1
N - Adventure and western fiction	Novels	9
	Short Stories	1
P - Romance and Love story	Novels	7
	Short stories	2
R - Humour	Articles from humorous books other than novels	2
	Total	48

As Zwicky points out in condition v, stranded *to* must attach to a VP or a predicator in a VP. Verbal patterns with clause final (unstranded) *to*-infinitive found

in Hornby (1976) are listed below. In each case, *to*-infinitive (phrase) is dominated by VP within the same S as the rest of the verbal pattern. (See condition ii (OSC)).

1. Subject + intransitive verb/*it* + *be* adj./noun + preposition + (pro)noun + to-inf.(phrase)
2. Subject + intrans. verb + (*to* + pro/noun) + to-inf.(phrase)
3. Subject + trans. verb/ *have/ought* + (not) + *to*-infin.(phrase)
4. Subject + trans. verb + interr. pro-noun/adverb + *to*-infin.(phrase)
5. Subject + trans. verb + noun/pronoun + (not) + *to*-infin.(phr.)

Our data contains no examples of pattern 1 or 4. There are four cases of pattern 2 with the verbs *endeavour, seem, aim* and the quasi-modal *be going to*, and 35 instances fit pattern 3. There are four instances where *to* is preceded by verb + pronoun as in pattern 5 and four instances where *to* is preceded by verb plus *not* as in pattern 3 or verb plus noun plus *not* as in pattern 5. Our examples of pronoun *to* and *not to* support Zwicky's exception clause which states that to can attach itself to monosyllabic non-lexical items (see vii).

Zwicky divides his data into five categories, depending on what type of control stranded *to* has within its predicate constructions. When treated in the same way our data produces the following results:

a) Subject controlled predicate constructions: 43
   i Frequent: want to (14)
               have to (6)
   ii Others: ought to; wish to; endeavour to;
              pretend to; expect to; seem to;
              tempt to; be willing to; love to;
              be going to; aim to; try to (2);
              like to (3); be delighted to;
              pleased to.
b) Impersonal constructions (extraposed subject complement): 1
              it is meant to
c) Indirect questions: None
d) Object controlled predicate constructions: 4
              want pro to
              expect pro to
              ask pro to
              allow pro to
e) *not to*: 4 All subject controlled predicate constructions
              see ... not to
              beseech not to
              try not to
              seem not to

Except in the case of 'indirect questions' our results are proportional to Zwicky's, which supports the conclusions he drew from his research (listed above, i-v). Two of Zwicky's three core instances, *want to* and *have to* (obligative) also appear frequently in our data, whereas the third, *used to*, is completely absent.

It is striking that in our data the majority of verbs preceding stranded *to* are stative (non-conclusive) verbs which denote a state, feeling or attitude rather than an action. These are all embedding verbs followed by embedded clauses which have been replaced by stranded *to*. Embedding verbs, even when non-stative, tend to have a low degree of activity and often express an attitude towards the activity described in the complement.[1] Our findings indeed show that while the majority of embedding verbs preceding stranded *to* are stative (or do not describe an action as such), verbs heading phrases replaced by stranded *to* are in 40 of the 47 cases non-stative; of the other 7, 1 is a modal and 6 are stative verbs. The following list shows the kind of embedding verbs preceding stranded *to* in our data.

Stative verbs:
    expressing mental state, perception:
        mean (1); aim (1); expect (2); see (1); allow (1).
    expressing emotional state:
        want (15); love (1); like (4); wish (1).
    expressing a state with *be*:
        be delighted (1); be willing (1); be pleased (1):
    others:
        seem (2).

Modals and quasi-modals (neither stative nor non-stative):
    have to (6); ought to (1); be going to (1).

Verbs in the passive (not indicating direct action):
    beseech (1); tempt (1).

Miscellaneous(verbs not describing a specific action):
    pretend (1); endeavour (3); try (3); ask (1).

VPD can replace verb phrases headed by a stative or non-stative verb, whereas *do so* and *do it* elliptic constructions as a rule only replace those headed by a non-stative verb. This means that VPD is more likely to lead to ambiguity than *do so* and *do it*, as shown below.[2]

F1585 ... even if it wanted to do so, has no power or authority to commit to,
the countries ....

Stranded *to* can either refer forward to *have* (stative) or to *commit* (non-stative) while *do so* refers forward to non-stative *commit*.

L701 And she wouldn't have given it.
Maybe she'd be alive today if she'd been willing to.

Despite the fact that stranded *to* can refer back to a stative or a non-stative verb, the verb *'be'* is not usually deleted unless there is complete subject control,[3] which is not the case in the above example. Therefore, if the complement of *be willing to* was meant to refer back to *be alive*, the text would have read:

Maybe she'd be alive today if she'd been willing to be.

In most of the examples from LOB the material that is deleted after stranded *to* can be retrieved from the sentence immediately preceding. There are, however, a few cases in which the deleted material that is understood after *to* is found further back in the text. We found three instances where the corresponding material is two sentences back (for full examples see appendix):

L 1531 deleted material: go across (some time) - from L 1529
P 1019 deleted material: join you in a little drink - from P 1017
P 1084 deleted material: have tea with you (tomorrow) (at four o'clock) - from P 1012

and one instance where it is situated as many as six sentences back:

L 1535 deleted material: go across (some time) - from L 1529

There is an important parallel between these examples: in all of them only very short sentences separate *to* from its referent (see appendix), enabling, as it were, the reader/hearer to keep in mind the material that is eligible for VPD. If then stranded *to* appears, the reader/hearer can easily add the appropriate VP in thought. Note also that all three cases occur in conversational English and that in the last two cases the conversation has been temporarily broken up by 'the narrator'.

While in the majority of our examples stranded *to* refers to material preceding it, we found some instances in which it is preposed, that is, where it refers to material to the right of it.

F 1585 The United Kingdom, even if it wanted to, has no power or authority to commit the countries of the Commonwealth to anything.

deleted material:
commit the countries of the Commonwealth to anything

G 1141 But his lying was of a special kind - it did not and could not by him have been expected to, deceive anyone who did not secretly wish to be deceived.

deleted material:
deceive anyone who did not secretly wish to be deceived

G 2219 Wind was ruffling the grass, and the corn-crakes (as I knew they would have to) sensed danger, and then scuttled into the field with the clumsy chicks tumbling over themselves as they followed as best as they could.

deleted material:
sensed danger

L 1334 He came at a pace that, without seeming to, carried him over the distance between us at a speed that left me no time to think at all.

deleted material:
carried him over the distance between us at a speed that left me no time to think at all

Zwicky does not mention this possibility, yet with 4 out of 47 instances it seems to be a feature that deserves our attention. The above-mentioned cases where *to* has forward reference are all either between brackets or commas. They function as 'asides' before the main information of the sentence is given. As far as can be judged from our examples this feature only occurs in fairly formal English. Zwicky, however, claims that the use of Verb Phrase Deletion with *to* is 'distinctly conversational', its natural environment being informal two-person discourse.

More than 50% of all the instances of stranded *to* that we found occurred in informal, conversational contexts. Cases of stranded *to* were common enough in non-conversational or even explicitly formal usage, though, suggesting that Zwicky's claim is somewhat too strong. This point is illustrated by the above sentences. We already mentioned that constructions containing stranded *to* in preposed position occur in formal usage, the more so when, as in the first sentence above, the deleted verb phrase and its identical counterpart are separated by intervening material. In cases like this, the reader/hearer must divide his/her attention between keeping the first part of the sentence stored away in memory and registering what follows, until the corresponding VP finally appears. The fact that no stylistic discord results from the combination of stranded *to* with formal lexical items (*besought, dispelled, endeavoured, pronouncements*) and with formal constructions like the passive further undermines Zwicky's claim concerning the distinctly conversational character of VPD with *to*.

Zwicky says that one of the means by which *to* can be stranded is VPD, in other words, the material that is deleted after stranded *to* consists of a VP, usually corresponding to the VP of the preceding clause. But what makes up a VP? In particular, Zwicky gives no decisive answer to the question whether adverbials should be considered part of the deleted VP or not. This is indeed a difficult issue. In our examples there are cases where the adverbial seems to be included in the deleted material:

M 644 - deleted material: start right away
N 2023 - deleted material: call you sooner
N 2186 - deleted material: go up there

There are cases where the adverbial is felt to be excluded:

G 3415 - deleted material: Gothicise his house(excl. 'at a cost of £ 20,000')
H 1171 - deleted material: develop that point (excl. 'now')
R 818 - deleted material: realise (excl. 'afresh')

And there are cases where it is ambiguous:

K 767 - deleted material: explain his failure (by his father's illness)
L 1531 - deleted material: go across (some time)
P 1267 - deleted material: see it (one day)

It seems that, in our small sample, no rule can be detected that would include some kinds of adverbials in the detected material and exclude others. All kinds of adverbials occur freely inside as well as outside the scope of the deleted material. In most cases the context will make clear how the sentence is to be interpreted. For example in sample F 1448 the adverbial, a clause which itself contains the stranded *to*, cannot be part of the deleted material, otherwise a recursive string of words would occur:

They meet each other whenever Keith's job as a collier on shift work will allow them to meet each other whenever Keith's job as a collier on shift work will allow them to meet each other whenever ...
(ad infinitum)

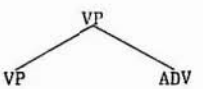
In his discussion of the Constraint on Verb Phrase Deletion, Kuno claims that only AdvP's which cannot occur in sentence-initial position are contained within the VP. Others form a right sister to the VP and then VPD is blocked. Kuno echoes Lakoff and Ross' claim (1966) that manner, duration, frequency, instrumental and several other kinds of adverbials are outside the VP; he also uses their 'do so test' to establish which adverbials can be derived from sentence-initial position, thereby claiming they cannot be constituents of the VP as in:

I drive with great care, but he does so recklessly.
He read for two hours, while I did so for five hours.

Nancy Levin(1980), however, argues that this only holds if *do so* replaces no less and no more than a VP, providing as counter-examples sentences like:

I drive with great care and I do so because I don't want to get arrested
(*do so* = drive with great care)
He read for two hours, and I did so too
(*did so* = read for two hours)[4]

If it is maintained that adverbials of this kind are outside the VP, the deletion of the adverbials in the last set of sentences poses a problem for Kuno's claim. Nancy Levin's suggestion is to analyse the problematical adverbs as outside one VP and inside another, as represented in the following tree structure:[5]



Of course this is not the end of the matter, but it would be beyond the scope of this paper to analyse this problem in great detail.

As mentioned before, VPD with *to* is particularly common in spoken language or 'two-person discourse', as Zwicky puts it.[6] In this type of conversation stranded *to* is frequently used (30% of the instances found) in answering questions, invitations and commands:

N 1020/1 "You listen to your mother."
"I have to."
N 1764/5 "That tells me a hell of a lot, doesn't it?"
"Yes," he said , "it's meant to."
N 2022/3 "... why the devil didn't you call me sooner?"
"I tried to but you were sound asleep."
P 1017/8 "Won't you two ladies join me in a little drink ..."
"We shall be pleased to ..."
P 1083/4 "After tea I will show you round the house and grounds."
"I shall be delighted to."

What Zwicky does not mention is the fact that the deleted VP is here not identical to the VP in the preceding sentence, as is usually the case. These utterances are spoken alternately by two persons addressing each other; the viewpoints change accordingly and as a consequence the implied pronoun in the ellipted answer, though still referring to the same person, is changed. For example:

N 1020/1 "You listen to *your* mother."
(Answer) "I have to."
Deleted: listen to *my* mother

In this paper we have tried to make an analysis of stranded *to* cases based on data consisting of attested examples of clause final stranded *to*. Arnold Zwicky's article 'Stranded *to* and Phonological Phrasing' proved to be a good starting point. Zwicky looks at stranded *to* in great detail using acceptable and unacceptable cases of both final stranded *to* and stranded *to* followed by other material. All our examples have been elicited from written texts while Zwicky's analysis was based on examples from texts, actual occurrences and those he had made up. Our limited data shows some of Zwicky's statements to be rather sweeping, as when he mentions VP Deletion without going into the complex structural aspects of adverbials. Many of our findings, however, do tally with Zwicky's conclusions and thus confirm them.

NOTES

1. N.Levin (1980): p.129.
2. ibid.
3. ibid., p.136.
4. OSU, *WPL* 24, March 1980, p.113.
5. ibid., p.114.

6. *Linguistics* 20, 1982: p.20.

REFERENCES

- Hornby, A.S. (1976), *Guide to Patterns and Usage in English*. O.U.P., London.
- Kuno, S. (1975), "Conditions for Verb Phrase Deletion", *Foundations of Language* 13: 161-175.
- Lakoff, G. and J.R. Ross (1966), "A Criterion for Verb Phrase Constituency", *Computation Laboratory of Harvard University Report*, No. NSF-17. Published under the title "Why you can't DO SO in the sink" in McCawley, ed., (1976).
- Levin, N. (1980), *Main Verb Ellipsis in Spoken English* OSU Working Papers in Linguistics No. 24: 65-165.
- McCawley, J. , ed. (1976), *Syntax and Semantics 7: Notes from the Linguistic Underground*, New York.
- Zwicky, A.W. (1982), "Stranded *to* and Phonological Phrasing in English", *Linguistics* 20: 3-57.

APPENDIX - EXAMPLES OF STRANDED TO EXTRACTED FROM THE LOB CORPUS

824 a07 I can't make players' contracts.
825 a07 I can't make a club pay a player so much a week.
826 a07 And, what's more, I don't want *to*.

1014 c11 PERHAPS you recognise that heavy and somewhat sullen face on the left.
1015 c11 If you are fond of being in the fashion you certainly ought *to*.

284 f03 Let the other fellow tell YOU something - if he wishes *to*!

1448 f14 They meet each other whenever Keith's job as a collier on shift work will allow them *to*.

1585 f15 The United Kingdom, even if it wanted *to*, has no power or authority to commit the countries of the Commonwealth to anything;

2144 f21 He will waste no time in vain regrets as he struggles with the laces, knowing very well that in all probability he will change his mind next May and put the great heavy things on again - and that, if it does not, it will be because he doesn't want *to*.

1127 g14 They admit they probably could not operate the engine any better themselves, while claiming as credit to themselves that at least they are not even pretending *to*.

1141 g14 But his lying was of a special kind - it did not, and could not by him have been expected *to*, deceive anyone who did not secretly wish to be deceived.

2219 g26 Wind was ruffling the grass, and the corn-crakes (as I knew they would have *to*) sensed danger, and then scuttled into the field with the clumsy chicks tumbling over themselves as they followed as best as they could.

3346 g38 They do not talk about reading and would be inarticulate if they tried *to*.

3415 g39 To please his mother, he had Gothicised his house at a cost of £20,000,though besought by Sir William not *to*.

5312 g63 Thus, there are two types of explanatory answers that one may give to the question: Do you intend?
5313 g63 One may, on the one hand, say something like "I want *to*, but I doubt if I can", in which case it is clear that the first of our conditions holds whereas there is uncertainty about the second.
5314 g63 On the other hand, one sometimes says
5315 g63 "I could go, but I don't really want *to*."
5316 g63 Here one is sure of the means but lacks the want.

6307 g75 You are in the habit of saying that the pronouncements of the other side tend to promote war and, if they seem not *to*, that is only because they are insincere and hypocritical.

1022 h15 If one read that statement in the context of the New Towns Act, it was reasonable to assume - and I think that most people assumed - that the Government were not intending to provide any new towns, either because they could not find sites or because they did not want *to*.

1171 h17 I think that the Forestry Commission have done a grand job which was never done until they were set up, but I will not develop that point now, although on another occasion I may be tempted *to*.

145 k02 But when I dispelled my irritation, or endeavoured *to*: apart from its being so irrational, I had, probably, another two or three weeks of his company ahead of me.

767 k06 They would be able, later, when time had become a little confused, to explain his failure by his father's illness, if they wanted *to*.

842 k07 Just brush them like this and this ... and this ... you'll feel the blood pulsing ... don't attack a mouth as if you're dipping a mop into a slop-bucket ... always go much slower than you want *to*, it increases desire ...

699 105 So long as Rose was alive Hilary couldn't have mortgaged everything he owned without her consent.
700 105 And she wouldn't have given it ...
701 105 Maybe she'd be alive today if she'd been willing *to*."

961 107 Don't complain."
962 107 "I wasn't going *to*, " I said.

1334 109 He came at a pace that, without seeming *to*, carried him over the distance between us at a speed that left me no time to think at all.

1529 111 "Can we go across some time?" she asked.
1530 111 "I love ferries."
1531 111 "If you'd like *to*."
1532 111 "Please."
1533 111 Rather diffidently she added,
1534 111 "I'm sorry, perhaps you'd rather ..."
1535 111 "I'd like *to*," MacLeod assured her.

2174 114 "Yes, but he didn't like to spend it.
2175 114 Not in getting rid of a wife when all he had to say was - if he wanted *to*, that is -

3464 122 So I just went like she asked me *to*, and told nobody.

643 m05 "Going to start right away, sir?"
644 m05 "We'll have *to*.

339 n03 That boy of mine seems to know how to do everything, when he wants *to*.

493 n04 "You'll get nothing out of me," said John, "but you can pay for my drinks if you want *to*, as long as you keep off Service matters."

799 n06 "You coming back?" he said when he'd got the saddle fixed.
800 n06 "I aim *to*, I said, cold as a fish.

954 n07 Asked a few questions on the way back yesterday, but nobody opened up.
955 n07 Didn't expect them *to*.

1020 n07 "You listen to your mother."
1021 n07 "I have to."

1764 n11 "That tells me a hell of a lot, doesn't it?"
1765 n11 "Yes," he said, "it's meant *to*."

2022 n13 "If you've been up since four," I said, "why the devil didn't you call me sooner?"
2023 n13 "I tried *to* but you were sound asleep."

2127 n13 Max was worrying, as he always did because he liked *to*.

2186 n14 He can go up there any time he wants *to*.

4250 n25 And he had kissed her; not wanting *to*, holding himself back as if it was a sacrilege, and yet drawn down to her.

1017 p07 "Won't you two ladies both join me in a little drink then we can all go into the dance room together."
1018 p07 Vera looked at Caroline.
1019 p07 "We shall be pleased *to*, Mr. Carson," said Caroline.

1041 p07 "Shall we dance?"
1042 p07 "I should love *to*."

1082 p07 "Come and have tea with me tomorrow afternoon at four o'clock.
1083 p07 After tea I will show you round the house and grounds."
1084 p07 "I shall be delighted *to*."

1203 p08 If you could explain it isn't that I really want to go home. I just have *to*.

1265 p08 I have a large house there beside the river.
1266 p08 You must see it one day."
1267 p08 "Yes, I should like *to*."

1669 p10 "I never get up unless I have *to*."

2262 p14 Besides, I've done what I wanted *to*.

3192 p21 "You didn't tell him - Ralph back there?"
3193 p21 "Yes, I'm afraid I did, I didn't see any reason not *to*."

790 r07 Of course, everyone forgets; but obliviscents are people who try not *to*, who worry about it.

818 r08 They appeared to look elsewhere, indeed, until I realized afresh, as you have *to*, that all birds look at you from the sides of their heads.

MATERIAL AVAILABLE FROM BERGEN

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

Brown Corpus, text format I (without grammatical tagging): A revised version of the Brown Corpus with upper- and lower-case letters and other features which reduce the need for special codes and make the material more easily readable. A number of errors found during the tagging of the corpus have been corrected. Typographical information is preserved; the same line division is used as in the original version from Brown University except that words at the end of the line are never divided.

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

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

LOB Corpus, text: The LOB Corpus is a British English counterpart of the Brown Corpus. It contains approximately a million words of printed text (500 text samples of about 2,000 words).

LOB Corpus, KWIC concordance (also on microfiche): A complete concordance for all the words in the corpus. It includes word statistics for both the LOB Corpus and the Brown Corpus, showing the distribution in text samples and genre categories for both corpora. The text of the LOB Corpus is not available on microfiche.

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

London-Lund Corpus, KWIC concordance I: A complete concordance for the 34 texts representing spontaneous, surreptitiously recorded

conversation (text categories 1-3), made available both in computerised and printed form (J. Svartvik and R. Quirk (eds.) *A Corpus of English Conversation*, Lund Studies in English 56, Lund: C.W.K. Gleerup, 1980).

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

The material has been described in greater detail in previous issues of *ICAME News*. Prices and technical specifications are given on the order forms which accompany this newsletter. *Note that the concordances are now also available on higher-density tapes at a lower price.*

A printed manual accompanies tapes of the LOB Corpus. Printed manuals for the Brown Corpus cannot be obtained from Bergen. Some information on the London-Lund Corpus is distributed together with copies of the text and the KWIC concordances for the corpus. Users of the London-Lund material are, however, recommended to order the recent book by Svartvik *et al.*, *Survey of Spoken English: Report on Research 1975-81*, Lund Studies in English 63, Lund: C.W.K. Gleerup, 1982. The grammatically tagged version of the Brown Corpus can only be obtained from: Henry Kučera, TEXT RESEARCH, 196 Bowen Street, Providence, R.I. 02906, U.S.A. The Syntax Data Corpus, which consists of part of the Brown Corpus, with detailed syntactic tagging, can only be obtained from: Alvar Ellegård, Department of English, University of Gothenburg, Lundgrensgatan 7, S-412 56 Göteborg, Sweden.

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 used for bona fide research of a non-profit mature. Holders of copies of corpora may not reproduce any texts, or parts of texts, for any purpose other than scholarly research without getting the written permission of the individual copyright holders, as listed in the manual or record sheet accompanying the corpus in question. (For material where there is no known copyright holder, the person/s/ who originally prepared the material in computerized form will be regarded as the copyright holder/s/.)

3 Commercial publishers and other non-academic organizations wishing to make use of part or all of a corpus or a print-out thereof must obtain permission from all the individual copyright holders involved.

4 The person/s/ who originally prepared the material in computerized form must be acknowledged in every subsequent use of it.

EDITORIAL NOTE

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

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