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

Machine-readable texts in English language research

No. 6
June 1982

No.6

CONTENTS

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

THAT V. ZERO CONNECTIVE IN ENGLISH NOMINAL CLAUSES

Johan Elsness
University of Oslo, Norway

1 INTRODUCTION

It is well known that in the most common type of nominal sub-clause in English the connective alternates between *that* and zero: 'I know that he is here.' v. 'I know he is here.' In this article we shall look into the question of what linguistic factors condition the choice between these two constructions.[1]

It may seem that in the latter construction the connective is omitted or suppressed. However, several grammarians have maintained that the two types of nominal clause can be traced back to two different paratactic constructions in Old English, one with, the other without the demonstrative ðæt, and that it is therefore misleading to describe the relationship between them in present-day English as one of omission or suppression of the connective in the latter type (see e.g. Jespersen 1928:32; Poutsma 1929:614; and also *OED*).

The question of the respective historical origins and developments of the two constructions will not be taken up in the present article, which will be concerned with the relationship between them in present-day English, from a strictly synchronic point of view. Nominal clauses introduced by the connective *that* will be referred to as 'THATs', those without any connective as 'ZEROs'. The term 'NOCL' will be used to refer to nominal clauses of either type.

There has not usually been assumed to be any hard and fast line of division between the syntactic or semantic functions of THATs on the one hand and those of ZEROs on the other. In many contexts they can seemingly be used interchangeably, with no apparent difference in meaning. When investigating the distribution between THATs and ZEROs one does not, therefore, expect so much to find conditioning factors making one or the other construction obligatory as factors which tend to favour one construction over the other even though either would be acceptable.

2 CONDITIONING FACTORS

The factors which have been recognised as important for the choice of connective in NOCLs can be grouped into five main types:

(A) *Style*. This is the conditioning factor most often referred to. Although it is not always made clear in what sense the term 'style' is used, it will for our present purposes suffice to say that it refers to linguistic distinctions along the formal/informal scale. THATs are said to predominate in formal style, the proportion of ZEROs increasing as the style becomes less formal. Ellinger (1933:107) and Storms (1966:264) have drawn particular attention to scientific writings, where they claim that THATs are preferred to the virtual exclusion of ZEROs, because of the greater clarity (Ellinger) or objectivity (Storms) characterising the former.

(B) *Syntactic function*. The THAT/ZERO distribution has often been claimed to depend on the syntactic function of the NOCL. It has been asserted, for example, that NOCLs are invariably or predominantly THATs when preceded by co-ordinate clauses (Ellinger 1933:106-7; Jespersen 1928:37-8; Poutsma 1929:621), and, especially, when complementing nouns (Hornby 1954:132-3; Jespersen 1928:36; Kruisinga 1932: 369, 374; Poutsma 1929:617; Storms 1966:262; Zandvoort 1957:222).

(C) *Potential ambiguity*. THATs are often said to be preferred in cases where ZEROs might lead to ambiguity. The reason may be, for instance, that with zero connective it would be difficult or impossible to determine whether an adverbial belongs in the matrix clause or in the NOCL.

(D) *The matrix verb*. In the case of object NOCLs the matrix verb is frequently referred to as a conditioning factor. Although these verbs are sometimes classified according to style and/or meaning, one also finds mere listings of verbs said to favour either THAT or ZERO objects, or to vacillate (Ellinger 1933, passim; Fowler 1965:624; Jespersen 1928:33; Poutsma 1929:615).

(E) *Semantic contrast*. Most of the linguists and grammarians who have concerned themselves with the THAT/ZERO opposition have assumed that there is no semantic difference between the two types of NOCL. A notable exception is Dwight Bolinger, who has argued that all uses of the word *that* in present-day English, including its use as a nomi-

nal connective, retain some of the anaphoric force they had at an earlier stage in the development of the language, still apparent in its use as a demonstrative. In respect of NOCLs this is claimed to have the consequence that, other things being equal, THATs tend to be selected in cases where the connective points directly to the preceding context (sometimes extralinguistic), whereas ZEROs are more likely to be preferred in other cases (see Bolinger 1972 and 1977).

The distinction between these five types of conditioning factors is not always straightforward. In particular it may be difficult to distinguish between (A) and (E), style and semantic contrast. Indeed, it is possible to hold the view that no such distinction can be made at all, if one takes the monist stance that meaning and style are ultimately inseparable.[2]

Provided one operates with a stylistic component at all, it will probably be agreed that stylistic differences can be expressed by an unspecifiable number of linguistic variables. This is certainly the case if one takes as one's starting point the view that style is a reflection of the attitude of the speaker/writer 'to the hearer (or reader), to the subject matter, or to the purpose of [the] communication', as Quirk *et al.* put it (1972:23), or some similar formulation. In the style of a text will therefore be incorporated an unknown number of contextual features, only some of which one can hope to identify within the immediate contexts of the various realisations of any given linguistic variable. Considered as a linguistic conditioning factor style is thus not on a par with the various specific contextual conditioning factors which can be identified: style is more fundamental, possibly itself functioning as a conditioning factor of (some of) the other conditioning factors one has identified (see Ulvestad 1956:48, note). (Various interdependences may also hold among specific contextual conditioning factors, but none of these seem capable of playing the same fundamental role as style.)

Hence style, in the broad sense of the total linguistic expression of the attitude of the speaker/writer, will often have the advantage of subsuming many, perhaps even all, of the more specific conditioning factors one operates with, and can in addition be expected to incorporate some specific conditioning factors which the investigator has failed to recognise. The difficulty of making use of style in

this sense as a conditioning factor is that there is usually no obvious way of ascertaining the pivotal attitude of the speaker/writer independently of the text itself, something which is necessary in order to avoid circular reasoning. In the present article it will simply be assumed that writers of different genres can be expected to adopt different attitudes of the kind relevant to style.

3 METHOD

It follows from the above that in a study of the distribution between THATs and ZEROs distinctions between what is more and what is less likely can be expected to be more interesting than distinctions between what is possible and what is impossible, although the difference between the two kinds of distinction is clearly one of degree, the impossible being merely the limiting case of the unlikely.

Different linguistic methods lend themselves to different kinds of investigation. A corpus study is particularly well suited to examining frequencies of occurrence, and hence to ascertaining what is more and what is less likely, and could therefore be expected to shed interesting light on the THAT/ZERO distribution. The present study of English NOCLs is based on an investigation of the Syntax Data Corpus, which consists of 64 of the 500 texts making up the Brown University Corpus of American English, 16 texts from each of four of the text categories that the Brown Corpus is divided into. In a project at the University of Gothenburg directed by Alvar Ellegård these texts were supplied with a detailed system of grammatical tagging in machine-readable form (see Ellegård 1978). The four text categories (henceforth 'TCs') represented in the Syntax Data Corpus are A: 'Press: Reportage', G: 'Belles Lettres, Biography, etc.', J: 'Learned and Scientific Writings', and N: 'Fiction: Adventure and Western'. As each text consists of about 2,000 running words, the whole of the Syntax Data Corpus contains approximately 128,000 words.

Because of the wide spread over different text types, this corpus provides a good basis for the study of linguistic variation in terms of genre or, by implication, style. At the same time its limitations should be obvious: it represents only written, American English, published in printed form (in 1961). Comparison of British and American English would have been of great interest for a study of the THAT/ZERO distribution, as would comparison with spoken English, and

also with unpublished written English, such as business and personal letters.

The Syntax Data Corpus does not purport to be representative of present-day English as such, or even of written American English published in printed form in 1961, since it comprises texts from only four different TCs, and each TC is represented by the same number of texts (cf. the Brown Corpus, whose 15 TCs are weighted in accordance with their assumed importance in the language as a whole). The most one can hope for in the case of the Syntax Data Corpus is therefore that the various TCs should be representative of their respective genres. Overall occurrences in the corpus as a whole are less interesting, and will not be considered in the present article.

From the Syntax Data Corpus all potential NOCLs were extracted computationally, and subsequently checked individually for inclusion in or omission from the investigation. Among the clauses which were omitted were some whose function could be described as adverbial: clauses of result or purpose introduced by *so* + (Adj +) *that*, *such* + (Adj + N +) *that*, (*in order*) *that*.

Constructions with so-called 'comment clauses' are characterised by the matrix ('comment') clause being less important and the formally subordinated, although invariably connectiveless, clause being more important than is normally the case in NOCL constructions. These can be clearly distinguished only if the comment clause is non-initial ('John, he argued, would need all the support he could get.'), in which case they were excluded.

On the other hand, all *that*-clauses and similar clauses without connective complementing either nouns ('the *notion* (that) he was guilty') or adjectives ('I'm *sure* (that) he was guilty.') were included irrespective of the underlying syntactic and semantic relationship between the noun/adjective and the clause, even though one's view of whether these clauses should be classified as nominal will in some cases depend on one's model of description.

Co-ordinate constructions were counted as separate clauses provided they each contained both a subject and a predicate. Contrary to what has often been reported before (see above, 2 (B)), ZEROs preceded by co-ordinate NOCLs were found to be quite common. In such cases the second NOCL sometimes approximates to a main clause, co-ordinate with the matrix clause, as in the following construction, where the final

clause accounts for one of the two ZEROs recorded in TC J:

It must be remembered, however, that there are many agents for which there is no solid immunity and a partial or low-grade immunity may be broken by an appropriate dose of agent. (J08:133)[3]

In most cases our classification of such constructions followed that adopted for the Syntax Data Corpus.

4 SURVEY OF RESULTS

The screening process left a total of 1,017 NOCLs, on which the investigation was based.[4] The distribution of the recorded NOCLs within each of the four TCs can be studied in Table 1a, where the various THATs and NOCLs are classified according to their syntactic functions. It will be seen that the number of NOCLs varies considerably among the TCs: they are most numerous in the newspaper texts - TC A - and least numerous in the scientific texts - TC J.

As could be expected, NOCLs functioning as non-extraposed, non-complementary objects, with active matrix constructions, predominate in all four TCs, although in TC J there is also a substantial proportion of subject clauses. Elsness (1981) contains a detailed analysis of the distribution of the recorded NOCLs over the various syntactic functions.

Table 1b gives the percentage of ZEROs among the NOCLs recorded in each TC for the most numerous syntactic function, where the figures are more reliable than in the case of the other functions, and for all syntactic functions combined. It is interesting that the order in which the TCs come if arranged according to decreasing proportions of ZEROs, N-A-G-J, is the same that Ellegård found for several features taken to be related to stylistic differences, ranked in order of increasing formality (Ellegård 1978, esp. p. 77). However, the distribution between THATs and ZEROs shows much greater polarisation than many of the features Ellegård recorded, in respect of which TC A and TC G tended to cluster round a medium value.

It should be borne in mind, however, that the figures for each TC are based on 16 different texts. It goes without saying that any conclusions drawn about differences between the TCs depend for their validity on the distinctions between the TCs having been made consistently during the compilation of the corpus (i.e. the Brown Corpus),

Table 1a. Distribution THAT/ZERO in Syntax Data Corpus: number of occurrences according to constituent status, incl. extraposition (X), complementation and voice of matrix verb. THAT+ZERO=NOCL

Const. status	Compl.	Matrix voice	TC A			TC G			TC J			TC N		
			THAT	ZERO	NOCL	THAT	ZERO	NOCL	THAT	ZERO	NOCL	THAT	ZERO	NOCL
Non-X subj.	Non-compl.	Active	0+	0=	0	2+	0=	2	1+	0=	1	1+	0=	1
	Compl. noun	Active	6+	1=	7	10+	0=	10	11+	0=	11	9+	1=	10
		Passive	7+	0=	7	1+	0=	1	4+	0=	4	1+	0=	1
X subj.	Non-compl.	Active	5+	2=	7	18+	0=	18	29+	0=	29	15+	5=	20
		Passive	2+	4=	6	5+	0=	5	25+	1=	26	0+	0=	0
	Compl. noun	Active	0+	0=	0	1+	0=	1	0+	0=	0	2+	0=	2
Non-X obj.	Non-compl.	Active	138+	150=	288	123+	21=	144	78+	1=	79	67+	93=	160
		Passive	5+	2=	7	3+	0=	3	0+	0=	0	0+	1=	1
	Compl. noun	Active	16+	3=	19	10+	1=	11	7+	0=	7	6+	0=	6
X obj.	Non-compl.	Active	0+	0=	0	1+	0=	1	2+	0=	2	2+	2=	4
Subj. com-ple-ment	Non-compl.	Active	7+	0=	7	16+	0=	16	6+	0=	6	5+	0=	5
	Compl. noun	Active	1+	0=	1	0+	0=	0	2+	0=	2	2+	0=	2
	Compl. adj.	Active	2+	2=	4	4+	0=	4	0+	0=	0	11+	10=	21
Other	Non-compl.	Passive	1+	0=	1	0+	0=	0	0+	0=	0	0+	0=	0
	Compl. noun	Active	5+	0=	5	19+	0=	19	7+	0=	7	4+	0=	4
		Passive	2+	0=	2	2+	0=	2	7+	0=	7	0+	0=	0
	Compl. adj.	Active	1+	0=	1	0+	0=	0	0+	0=	0	0+	0=	0
S U M			198+	164=	362	215+	22=	237	179+	2=	181	125+	112=	237

Table 1b. Percentage of ZEROs among NOCLs functioning as non-extraposed, non-complementary objects in active matrix constructions, and among NOCLs in all syntactic functions combined.

Syntactic function	TC A	TC G	TC J	TC N
Non-X, non-compl. obj., active matrix clause	52.1%	14.6%	1.3%	58.1%
All functions combined	45.3%	9.3%	1.1%	47.3%

so that each text in a given TC is similar in respect of genre to the other texts in that TC, and different from any text in the other TCs. We can only assume that this is the case with the texts included in the Syntax Data Corpus.[5]

The THAT/ZERO distribution showed considerable variation from one text to another within each TC. It could not, therefore, be taken for granted that the quite small overall differences in the proportion of ZEROs between TC A and TC N, or even between TC G and TC J, were statistically significant. To find out, I used Student-Fisher's *t* test, which takes into account the variation within each sample, i.e. within each TC. The test showed that the difference in the THAT/ZERO distribution between TC A and TC N is *not* statistically significant in respect of either object NOCLs (t=0.07; d.f.=30) or NOCLs in all syntactic functions combined (t=0.18; d.f.=30). It further showed that the difference in the THAT/ZERO distribution between TC G and TC J *is* statistically significant, at 5% level, as regards both object NOCLs (t=2.61; d.f.=28) and NOCLs in all syntactic functions combined (t=2.73; d.f.=30). In other words: the recorded differences in the THAT/ZERO distribution between TC A and TC N may be due to chance, and hence do not warrant any claims about real underlying differences between the kind of English used in newspaper reportage and the kind used in Adventure and Western Fiction; the recorded differences between the other TCs, however, can be assumed to reflect real differences between the respective genres.

Of previous frequency studies of the alternation between THAT and ZERO irrespective of syntactic function the most comprehensive one is probably McDavid (1964), which is based on a corpus estimated at 100,000 words, made up of books and periodicals taken to be 'samples of well-edited written English' (McDavid 1964:103). Stylistically, McDavid's corpus appears to have been much more homogeneous than mine, and thus less suitable for a study of linguistic variation. It was found to contain 'about 650' NOCLs, of which only 7.5 per cent were ZEROs.

When one considers that from a stylistic point of view McDavid's corpus apparently belongs somewhere between TC G and TC J in the Syntax Data Corpus, my findings can be seen to agree very well with hers. The wide spread in the proportion of ZEROs in my material under-

lines the importance of basing frequency studies of this kind on as diverse a corpus as possible. Thus the fact that ZEROs can be roughly as frequent as THATs even in written English published in printed form appears to have escaped most previous investigators.

4.1 Complementation

In Table 1c the NOCLs recorded in each TC are distinguished only according to their complementary status, although to consider overall occurrences according to just one conditioning factor is of limited value, since any interdependences between this and other conditioning factors, such as constituent status, are ignored.

Table 1c. Distribution THAT/ZERO according to complementary status. THAT+ZERO=NOCL, with percentages of ZERO in brackets.

Complementary status	TC A	TC G	TC J	TC N
Non-complementary	158+158=316 (50.0%)	168+ 21=189 (11.1%)	141+ 2=143 (1.4%)	90+101=191 (52.9%)
Complementing nouns	37+ 4= 41 (9.8%)	43+ 1= 44 (2.3%)	38+ 0= 38 (0.0%)	24+ 1= 25 (4.0%)
Complementing adj.	3+ 2= 5 (40.0%)	4+ 0= 4 (0.0%)	0+ 0= 0 (-)	11+ 10= 21 (47.6%)
S U M	198+164=362 (45.3%)	215+ 22=237 (9.3%)	179+ 2=181 (1.1%)	125+112=237 (47.3%)

It will be seen that the proportion of ZEROs is highest among non-complementary NOCLs in all four TCs. The quite considerable number of NOCLs complementing nouns are predominantly THATs, although it is noteworthy that ZEROs complementing nouns were also recorded in all TCs except J. This is contrary to what several grammarians have claimed before (see above, 2 (B)).

Some of these complementary ZEROs are in a non-appositive relationship with their respective head nouns. This is true of both the ZEROs complementing nouns acting as non-extraposed subjects (see Table 1a), where underlying prepositions can be posited between head noun and NOCL (besides, these subject NOCLs are non-initial because embedded in existential *there* constructions):[6]

Should there be *evidence* they are shirking, he has said, the state police will step into the situation. (A05:72)

Montero's shot had caught him high in the chest; there was no *doubt* he was dying. (N04:132)

However, constructions where a complementary ZERO can be looked upon as an appositive of the head noun also occur:

He expressed the *opinion* the city could hire a CD director for about 3500 dollars a year and would only have to put up half that amount on a matching fund basis to defray the salary costs. (A05:13)

They would like to convey the *notion* something is being done, even though it is something they know to be ineffectual. (A08:148)

All but one of the adjectives recorded with complementary NOCLs function as subject complements. They are numerous only in TC N, where they can be seen to be distributed almost equally between THATs and ZEROs. The adjectives occurring as subject complements with complementary NOCLs are:

in TC A with THATs: *agreed, certain*;

in TC A with ZEROs: *glad, sure*;

in TC G with THATs: *aware, glad, surprised, worried*;

in TC N with THATs: *aware* (5x), *certain* (2x), *proud, sure, thankful, worried*;

in TC N with ZEROs: *certain* (2x), *glad, lucky, sure* (6x).

It will be seen that in TC A and TC N there is a tendency for monosyllabic adjectives to take ZEROs and other adjectives to take THATs, although the total number of recorded instances is too small to warrant any firm conclusions.

4.2 Matrix voice

In Table 1d overall occurrences in each TC are broken down according to the voice of the matrix clause. These distributions, too, should be treated with caution, since any interdependences between conditioning factors are again concealed. It may nevertheless be of some interest to note that, although rare, ZEROs embedded in passive matrix constructions were recorded in all TCs except G, albeit with just one occurrence in TC J and TC N. The small number of passive constructions with embedded ZEROs in TC G, TC J and TC N was only to be expected, since ZEROs are rare irrespective of matrix voice in the former two TCs and passive constructions are rare in the latter; clearly, there is no question of the differences being statistically significant in any of those TCs. In the one TC where both ZEROs and

Table 1d. Distribution THAT/ZERO according to matrix voice. THAT+ZERO= NOCL, with percentages of ZERO in brackets.

Matrix voice	TC A	TC G	TC J	TC N
Active	181+158=339 (46.6%)	204+ 22=226 (9.7%)	143+ 1=144 (0.7%)	124+111=235 (47.2%)
Passive	17+ 6= 23 (26.1%)	11+ 0= 11 (0.0%)	36+ 1= 37 (2.7%)	1+ 1= 2 (50.0%)
S U M	198+164=362 (45.3%)	215+ 22=237 (9.3%)	179+ 2=181 (1.1%)	125+112=237 (47.3%)
Chi-square[7]	$X^2 = 2.88$; $0.05<p<0.10$	-	-	-

passive constructions are numerous, TC A, combinations of the two can be seen to be common, although the overall proportion of ZEROs is lower in passive than in active constructions, a difference which is not, however, statistically significant at 5% level, according to the Chi-square test.

5 OBJECT NOCLs

We shall henceforth concentrate on the NOCLs acting as non-extraposed, non-complementary objects in active matrix clauses, which we have seen is the most common syntactic function in all four TCs. It accounts for 79.6% of all the NOCLs recorded in TC A, 60.8% in TC G, 43.6% in TC J, and 67.5% in TC N.

5.1 Matrix verbs

As regards the THAT/ZERO distribution, the conditioning factor most often referred to besides the general style factor is the matrix verb taking the NOCL as its object (see above, 2 (D)). In Table 2a the recorded matrix verbs are listed in order of decreasing frequency in each TC, with their respective THAT/ZERO distributions.

The material lends little, if any, support to claims to the effect that certain verbs occur exclusively with either THAT or ZERO objects. In the two TCs where THATs and ZEROs are about equally frequent, A and N, and also in TC G, nearly all the verbs recorded with object NOCLs in any substantial numbers can be seen to occur with both THATs

Table 2a. Active matrix verbs taking non-extraposed, non-complementary object NOCLs, in order of decreasing frequency in each TC, with THAT/ZERO distribution. THAT+ZERO=NOCL

TC A			
SAY	28+	86=	114
TELL	9+	8=	17
THINK	1+	15=	16
ADD	9+	0=	9
BELIEVE	2+	5=	7
FEEL	1+	6=	7
ANNOUNCE	5+	1=	6
POINT OUT	5+	1=	6
RECOMMEND	6+	0=	6
STATE	5+	0=	5
ESTIMATE	1+	3=	4
ARGUE	2+	2=	4
DECLARE	2+	2=	4
FIND	2+	2=	4
INDICATE	1+	3=	4
INSIST	3+	1=	4
MEAN	2+	2=	4
NOTE	4+	0=	4
ADMIT	2+	1=	3
PREDICT	2+	1=	3
SHOW	3+	0=	3
SUGGEST	3+	0=	3
ADVISE	2+	0=	2
ASK	2+	0=	2
CHARGE	2+	0=	2
COMPLAIN	2+	0=	2
DISCLOSE	2+	0=	2
EXPLAIN	2+	0=	2
HOPE	2+	0=	2
INFORM	2+	0=	2
REPORT	2+	0=	2
RULE	2+	0=	2
URGE	2+	0=	2
AGREE	0+	1=	1
ASSERT	1+	0=	1
ASSUME	0+	1=	1
ASSURE	1+	0=	1
BET	1+	0=	1
CONCEDE	1+	0=	1
CONCLUDE	1+	0=	1
CONFESS	1+	0=	1
CONTEND	0+	1=	1
DENY	1+	0=	1
DISCOVER	1+	0=	1
EMPHASIZE	1+	0=	1
EXPECT	1+	0=	1
FEAR	1+	0=	1
FIND OUT	1+	0=	1
INSURE	1+	0=	1
KNOW	0+	1=	1
LEARN	0+	1=	1
MAKE SURE	1+	0=	1
PLEAD	1+	0=	1
PROVE	0+	1=	1
REALIZE	0+	1=	1
REGRET	0+	1=	1
REMARK	1+	0=	1
REQUEST	1+	0=	1
SUSPECT	0+	1=	1
TESTIFY	0+	1=	1
WISH	0+	1=	1
WRITE	1+	0=	1
Sum	138+	150=	288

TC G			
SAY	17+	3=	20
BELIEVE	9+	3=	12
KNOW	5+	1=	6
THINK	3+	3=	6
FEEL	4+	1=	5
MAINTAIN	5+	0=	5
MEAN	5+	0=	5
TELL	4+	1=	5
REALIZE	2+	2=	4
RECOGNIZE	4+	0=	4
SEE	4+	0=	4
ASSURE	3+	0=	3
CONSIDER	3+	0=	3
FIND	3+	0=	3
INSIST	2+	1=	3
MENTION	3+	0=	3
SHOW	3+	0=	3
ADMIT	2+	0=	2
ASSERT	2+	0=	2
CLAIM	2+	0=	2
CONFESS	2+	0=	2
DENY	2+	0=	2
IMAGINE	2+	0=	2
REPLY	2+	0=	2
STIPULATE	2+	0=	2
SUGGEST	1+	1=	2
SUSPECT	2+	0=	2
SWEAR	1+	1=	2
UNDERSTAND	2+	0=	2
ADD	1+	0=	1
AGREE	1+	0=	1
ARGUE	1+	0=	1
AUTHENTICATE	1+	0=	1
CONCLUDE	1+	0=	1
CONTEND	0+	1=	1
CONVINCE	1+	0=	1
DECLARE	1+	0=	1
DEMAND	1+	0=	1
DEMONSTRATE	1+	0=	1
EXPECT	1+	0=	1
EXPLAIN	1+	0=	1
GRANT	0+	1=	1
GUESS	0+	1=	1
HOLD	1+	0=	1
INDICATE	1+	0=	1
LEARN	1+	0=	1
MAKE CLEAR	1+	0=	1
RECALL	1+	0=	1
REMEMBER	1+	0=	1
REMIND	1+	0=	1
SIGH	1+	0=	1
STRESS	1+	0=	1
SUPPOSE	0+	1=	1
THANK	1+	0=	1
TRUST	1+	0=	1
Sum	123+	21=	144

TC J			
SHOW	11+	0=	11
ASSUME	7+	0=	7
EXPECT	6+	0=	6
INDICATE	6+	0=	6
NOTE	6+	0=	6
BELIEVE	3+	0=	3
OBSERVE	3+	0=	3
POINT OUT	3+	0=	3
DEDUCE	3+	0=	3
DEMONSTRATE	3+	0=	3
RECOGNIZE	2+	0=	2
REQUIRE	2+	0=	2
SEE	2+	0=	2
THINK	2+	0=	2
ADD	1+	0=	1
AGREE	1+	0=	1
ARGUE	1+	0=	1
ASSURE	1+	0=	1
CLAIM	1+	0=	1
CONCLUDE	1+	0=	1
EMPHASIZE	1+	0=	1
ENSURE	1+	0=	1
FEEL	1+	0=	1
FIND	0+	1=	1
IMPLY	1+	0=	1
INFER	1+	0=	1
LEARN	1+	0=	1
MAKE CERTAIN	1+	0=	1
MEAN	1+	0=	1
NOTICE	1+	0=	1
REMEMBER	1+	0=	1
REVEAL	1+	0=	1
STATE	1+	0=	1
SUGGEST	1+	0=	1
Sum	78+	1=	79

Table 2a (cont.)

TC N								
THINK	6+29= 35	FIGGER	0+ 3= 3	BET	0+ 1= 1
KNOW	7+ 9= 16	FIND	3+ 0= 3	DECIDE	0+ 1= 1
SEE	11+ 1= 12	MAKE SURE	0+ 3= 3	EXPLAIN	1+ 0= 1
SAY	2+ 8= 10	MEAN	0+ 3= 3	FORESEE	1+ 0= 1
TELL	2+ 8= 10	ASCERTAIN	2+ 0= 2	FORGET	1+ 0= 1
BELIEVE	1+ 3= 4	DOUBT	2+ 0= 2	IMAGINE	0+ 1= 1
FEEL	3+ 1= 4	ESTIMATE	1+ 1= 2	NOTICE	1+ 0= 1
GUESS	0+ 4= 4	GRASP	2+ 0= 2	ORDER	1+ 0= 1
HOPE	3+ 1= 4	HEAR	2+ 0= 2	PRETEND	0+ 1= 1
INDICATE	3+ 1= 4	INSIST	1+ 1= 2	PROVE	1+ 0= 1
REALIZE	3+ 1= 4	LEARN	1+ 1= 2	SHOW	1+ 0= 1
RECKON	0+ 4= 4	SUPPOSE	0+ 2= 2	SUGGEST	1+ 0= 1
WISH	1+ 3= 4	SWEAR	1+ 1= 2	WHISPER	1+ 0= 1
		ADMIT	1+ 0= 1	Sum	67+93=160
		ASSUME	0+ 1= 1		

Table 2b. Distribution THAT/ZERO according to frequency of active matrix verb taking non-extraposed, non-complementary object NOCL. THAT+ZERO=NOCL, with percentages of ZERO in brackets.

Number of occurrences of matrix verb in respective TCs	TC A	TC G	TC N
5 or more	71+122=193 (63.2%)	52+ 12= 64 (18.8%)	28+ 55= 83 (66.3%)
3 - 4	27+ 17= 44 (38.6%)	27+ 3= 30 (10.0%)	17+ 27= 44 (61.4%)
1 - 2	40+ 11= 51 (21.6%)	44+ 6= 50 (12.0%)	22+ 11= 33 (33.3%)
S U M	138+150=288 (52.1%)	123+ 21=144 (14.6%)	67+ 93=160 (58.1%)
Chi-square	X^2 =31.80; d.f.=2; p<0.01	-	X^2 =10.78; d.f.=2; p<0.01

and ZEROs. The statements found in some grammars about certain verbs normally taking THATs, others normally taking ZEROs, thus appear to have been too categorical.

It is nevertheless a striking feature of these lists that a majority of the most frequent matrix verbs in TC A, TC G and TC N seem to have been recorded with comparatively high proportions of ZEROs. In Table 2b the matrix verbs occurring in each of these TCs are divided into

three groups according to frequency of occurrence, and the THAT/ZERO distribution is given in respect of each group.

It will be seen that in both TC A and TC N the proportion of ZEROs is considerably higher in the group of the most frequent matrix verbs, those occurring five times or more, than in the group of the least frequent matrix verbs, those occurring just once or twice: the ratios are roughly 3:1 in TC A and 2:1 in TC N. In both TCs the middle group, consisting of the matrix verbs occurring three or four times, occupies an intermediate position. The Chi-square test shows these variations to be statistically significant at 1% level in both TC A and TC N. As regards the types of English which these two TCs represent, where THATs and ZEROs are about equally frequent, my findings thus confirm the assumption that more common verbs tend to favour ZERO objects, less common verbs THAT objects. As frequency of occurrence is sometimes associated with style, more frequent items being claimed generally to be characteristic of informal or neutral style, less frequent items of more formal style, the differences observed in TC A and TC N might be related to stylistic differences within each TC.

In TC G, too, the proportion of ZEROs is largest among the most frequent matrix verbs. The most conspicuous feature of TC G, however, is that the differences are small (clearly not statistically significant), because of the low proportions of ZEROs throughout.

It may finally be noted that no significant differences in the THAT/ZERO distribution have been detected depending on the semantic classes of matrix verbs distinguished in Elsness (1981), or any other semantic classification.

5.2 Intervening adverbials and types of subject in TC A and TC N

The rest of this article will concentrate on the object NOCLs recorded in the two TCs where THATs and ZEROs have been found to occur with roughly the same frequencies, TC A and TC N, as in these TCs the choice between THAT and ZERO can be assumed to be affected less by the general style factor and more by more specific conditioning factors than in the two other TCs.

It is moreover a fact that certain NOCL constructions are of a type which makes comparison with other such constructions less interesting with respect to the THAT/ZERO distribution. For example, NOCLs with

subjunctive verbs are in my material invariably THATs, irrespective of TC. (Among the recorded object NOCLs 12 have subjunctive verbs in TC A, 2 in TC N.) Conversely, NOCLs in which the subject is raised into the matrix clause ('... the man (who) the President said was responsible.') are without exception ZEROs. NOCLs which are preceded by co-ordinate clauses (NOCL or other) or by indirect objects are also atypical from the point of view of the THAT/ZERO distribution, as are cases in which the NOCL is an existential *there* construction, since in such constructions the matrix verb is not immediately followed by the NOCL subject. In our further analysis of the distribution between THATs and ZEROs among object NOCLs all constructions of these types will be omitted.

One factor potentially affecting the THAT/ZERO distribution in the remaining constructions with object NOCLs is the occurrence or non-occurrence of adverbials between the matrix verb and the NOCL subject. As with zero connective it will not always be clear whether the adverbial belongs in the matrix clause or in the NOCL ('John said this morning the girl was gone.'), constructions with such intervening adverbials have been reported to be more likely to take *that* connective than other constructions (see above, 2 (C)).

Another factor which might be thought to affect the THAT/ZERO distribution is the type of subject occurring in the NOCL, for instance whether or not that subject is realised by a personal pronoun (see Ulvestad 1956:42).

The distribution between THATs and ZEROs was therefore examined both according to the occurrence or non-occurrence of adverbials between the matrix verb and the NOCL subject and according to whether the NOCL subject is realised by a personal pronoun or by some other kind of noun phrase. Both comparisons revealed great differences, as can be seen from Table 3.

Concerning constructions with intervening adverbials first, the proportion of ZEROs will be seen to be much lower among these than among the other constructions in both TCs. It is nevertheless noteworthy that as many as 11 ZEROs of this type were recorded in TC A, although none in TC N. One of these ZEROs in TC A occurs in the very first sentence of the Syntax Data Corpus (which is also the first sentence of the Brown Corpus):

Table 3. Distribution THAT/ZERO among object NOCLs according to type of NOCL subject and occurrence of adverbials between matrix verb and NOCL subject. THAT+ZERO=NOCL, with percentages of ZERO in brackets. All NOCLs with subjunctive verbs, existential there, raised subjects, preceding co-ordinate clauses or preceding indirect objects excluded.

NOCL subject	No interv. adverbial	Intervening adverbial	S U M	Chi-square
TC A				
Personal pronoun	10+ 56= 66 (84.8%)	6+ 6= 12 (50.0%)	16+ 62= 78 (79.5%)	X^2 = 5.58; 0.01<p<0.05
Other	63+ 60=123 (48.8%)	22+ 5= 27 (18.5%)	85+ 65=150 (43.3%)	X^2 = 7.07; p<0.01
S U M	73+116=189 (61.4%)	28+ 11= 39 (28.2%)	101+127=228 (55.7%)	
Chi-square	X^2 =22.07; p<0.01	-		
TC N				
Personal pronoun	18+ 60= 78 (76.9%)	6+ 0= 6 (0.0%)	24+ 60= 84 (71.4%)	X^2 =12.60; p<0.01
Other	31+ 15= 46 (32.6%)	2+ 0= 2 (0.0%)	33+ 15= 48 (31.3%)	-
S U M	49+ 75=124 (60.5%)	8+ 0= 8 (0.0%)	57+ 75=132 (56.8%)	
Chi-square	X^2 =21.96; p<0.01	-		

The Fulton County Grand Jury said *Friday* an investigation of Atlanta's recent primary election produced "no evidence" that any irregularities took place. (A01:1)

Here the intervening adverbial *Friday* is clearly part of the preceding matrix clause. The following example, on the other hand, has a finite adverbial clause introducing the NOCL:

"But I believe *if people were better informed on this question*, most of them would oppose it also. (A02:61)

In fact only one other ZERO construction has an intervening adverbial which is part of the NOCL. This adverbial, too, is a finite clause:

He said *no matter what stand he takes* it would be misconstrued that he was sympathetic to one or the other of the Republicans. (A06:175)

In the case of the 9 ZERO constructions with intervening adverbials which belong in the matrix clause, the adverbials are without exception realised by single words or two-worded phrases denoting time, of the type *today, yesterday, last night, Monday, Monday night,* besides *Friday* in the example quoted above. That the adverbial is part of the matrix clause in as many as 9 of the 11 ZEROs with intervening adverbials is interesting, as the 28 THATs with intervening adverbials are evenly divided between constructions with matrix adverbials and constructions with NOCL adverbials.

Although the total number of constructions is too small to allow any definite conclusions, it may thus seem that zero connective is more readily available if the intervening adverbial belongs in the matrix clause, possibly because such constructions are in better agreement with the most common weight-distributional pattern in English clauses ('weight' in syntactic and/or semantic - and in speech also phonetic - terms), characterised by progressively heavier elements from beginning to end; adverbials being comparatively heavy constituents, the need for an overt boundary marker may be felt less strongly if the adverbial is clause-final than if it is clause-initial, since in the former case the weight-distributional pattern helps to signal the clause boundary. However, zero connective may seem to be more common with initial NOCL adverbials when these are realised by finite clauses, perhaps because a finite clause is so heavy that its function is felt to be clear enough without the overt connective.

While in the case of intervening adverbials the reason for the lower proportion of ZEROs is evidently a desire on the part of the writer to avert ambiguity (such desires may or may not be conscious), the reason for the higher proportion of ZEROs in constructions with NOCL subjects realised by personal pronouns is less obvious. One possible explanation might be that since some personal pronouns have distinct nominative forms, the risk of syntactic confusion is smaller with such subjects: with a matrix verb like BELIEVE, for instance, the decoder may be in temporary doubt about the syntactic function of the following noun phrase in a sentence like 'I believe Mike did it.', since *Mike* might itself be the object of *believe*, as in 'I believe Mike.'; if the NOCL subject is realised by a personal pronoun like

he, no such temporary confusion is possible.

This type of confusion can arise with only some of the verbs taking object NOCLs, since many, perhaps most, of the verbs occurring in such constructions do not normally take ordinary noun- or pronoun-headed noun phrases as objects: *there are no* *'I said Mike.', *'I thought Mike.', etc. - unless *Mike* is a mention form or the object is elliptical. It therefore seemed doubtful that this could account for the very marked difference in the proportion of ZEROs depending on type of NOCL subject that was recorded in both TC A and TC N.

Furthermore, if this was the reason for the variation in the proportion of ZEROs, one would expect only those personal pronouns which have distinct nominative forms (*I*, *he*, *she*, *we*, *they*) to occur with higher proportions of ZEROs, and the other personal pronouns (*you*, *it*) to take roughly the same proportions of ZEROs as subjects not realised by personal pronouns. However, no such difference could be traced in my material.

The recorded difference in the THAT/ZERO distribution might instead be explained by reference to the same weight-distributional principle that was referred to above: it could be that because personal pronouns are particularly light elements, they are more readily taken to be clause-initial than other noun phrases, with the result that the need for an overt marker of the clause boundary is felt less strongly.

It might also be that the greater structural complexity characterising many of the noun phrases not realised by personal pronouns in itself makes *that* connective more likely to be selected, as a contribution to greater syntactic clarity.

Another possible explanation of the fact that NOCLs with personal pronoun subjects show higher proportions of ZEROs could be that such a NOCL is felt to be more closely attached to the preceding matrix clause, because of the lighter subject, and that an overt syntactic marker between the clauses would therefore tend to be avoided. That would mean that zero connective is used to mark a closer clause juncture than *that* connective. In that case the THAT/ZERO distribution might also be expected to be affected by the person of the NOCL subject, and possibly also by that of the matrix subject, and by the

Table 4a. Distribution THAT/ZERO according to person and coreference of matrix and NOCL subjects. THAT+ZERO=NOCL, with percentages of ZERO in brackets. All NOCLs with subjunctive verbs, existential there, raised subjects, preceding co-ordinate clauses or indirect objects, or intervening adverbials excluded.

NOCL subj. \ Matrix subj.		1st/2nd person	3rd person, pers. pron.	3rd person, other	None	S U M
TC A						
1st/ 2nd pers.	+Coref.	0+ 4= 4 (100.0%)	-	-	0+ 1= 1 (100.0%)	0+ 5= 5 (100.0%)
	-Coref.	0+ 1= 1 (100.0%)	1+ 0= 1 (0.0%)	1+ 0= 1 (0.0%)	-	2+ 1= 3 (33.3%)
3rd p., pers. pron.	+Coref.	-	0+ 10= 10 (100.0%)	4+ 33= 37 (89.2%)	-	4+ 43= 47 (91.5%)
	-Coref.	0+ 1= 1 (100.0%)	1+ 1= 2 (50.0%)	3+ 5= 8 (62.5%)	-	4+ 7= 11 (63.6%)
3rd p., other	-Coref.	2+ 4= 6 (66.7%)	15+ 21= 36 (58.3%)	44+ 35= 79 (44.3%)	2+ 0= 2 (0.0%)	63+ 60=123 (48.8%)
S U M	+Coref.	0+ 4= 4 (100.0%)	0+ 10= 10 (100.0%)	4+ 33= 37 (89.2%)	0+ 1= 1 (100.0%)	4+ 48= 52 (92.3%)
	-Coref.	2+ 6= 8 (75.0%)	17+ 22= 39 (56.4%)	48+ 40= 88 (45.5%)	2+ 0= 2 (0.0%)	69+ 68=137 (49.6%)
TC N						
1st/ 2nd pers.	+Coref.	0+ 11= 11 (100.0%)	-	-	0+ 2= 2 (100.0%)	0+ 13= 13 (100.0%)
	-Coref.	0+ 8= 8 (100.0%)	0+ 3= 3 (100.0%)	0+ 1= 1 (100.0%)	0+ 4= 4 (100.0%)	0+ 16= 16 (100.0%)
3rd p., pers. pron.	+Coref.	-	3+ 8= 11 (72.7%)	2+ 4= 6 (66.7%)	-	5+ 12= 17 (70.6%)
	-Coref.	4+ 8= 12 (66.7%)	3+ 3= 6 (50.0%)	5+ 7= 12 (58.3%)	1+ 1= 2 (50.0%)	13+ 19= 32 (59.4%)
3rd p., other	-Coref.	6+ 9= 15 (60.0%)	14+ 4= 18 (22.2%)	11+ 2= 13 (15.4%)	-	31+ 15= 46 (32.6%)
S U M	+Coref.	0+ 11= 11 (100.0%)	3+ 8= 11 (72.7%)	2+ 4= 6 (66.7%)	0+ 2= 2 (100.0%)	5+ 25= 30 (83.3%)
	-Coref.	10+ 25= 35 (71.4%)	17+ 10= 27 (37.0%)	16+ 10= 26 (38.5%)	1+ 5= 6 (83.3%)	44+ 50= 94 (53.2%)

relationship between the two subjects: presumably, the association between the two clauses will be felt to be particularly close in cases where the NOCL subject is coreferential with the matrix subject (this will often be the only possibility for coreference with elements in the matrix clause in constructions without indirect objects).

To find out, I examined the constructions without intervening adverbials more closely. Table 4a distinguishes the persons of the matrix and NOCL subjects, and in the case of 3rd person subjects also whether these are realised by personal pronouns or by other noun phrases; NOCL subjects realised by personal pronouns are further distinguished according to whether or not they are coreferential with the matrix subject. In the few cases without expressed matrix subjects the intended subject could easily be inferred from the context, and so the question of coreference determined even in respect of those constructions. As no clear differences between constructions with 1st and with 2nd person subjects in either matrix clause or NOCL were detectable (apart from those related to coreference), these persons are merged in the table.

Table 4a shows that in both TC A and TC N the THAT/ZERO distribution varies markedly according to all the four conditioning factors person of matrix subject, person of NOCL subject, opposition pronoun/other noun phrase realising NOCL subject, and coreference between subjects. Although in some cases the number of occurrences is very small, it is interesting to note the clear tendency for the proportion of ZEROs to decrease as one moves from 1st/2nd person to 3rd person matrix subjects, and the similar tendency in the case of NOCL subjects, and also the tendency for the proportion of ZEROs to be higher with coreferential than with non-coreferential NOCL subjects. There is, moreover, a tendency among constructions with NOCL subjects in the 3rd person for the proportion of ZEROs to be higher if the subject is realised by a personal pronoun than if it is realised by some other kind of noun phrase, but no clear tendency in this direction in respect of the matrix subject in either TC.

In order to test for statistical significance by means of the Chi-square test one ought to vary just one of these conditioning factors at a time, but in that case the number of occurrences would frequently be too small for the test to yield significant results. Most of

the tendencies nevertheless seem convincing enough.

The differences depending on coreference are particularly interesting. The aggregate figures for coreferential and for non-coreferential NOCL subjects among constructions where these subjects are realised by personal pronouns are brought together in Table 4b, which thus ignores the distinctions depending on the person of the NOCL subject.

Table 4b. Distribution THAT/ZERO among object NOCLs with subjects realised by personal pronouns according to coreference between NOCL and matrix subjects. THAT+ZERO=NOCL, with percentages of ZERO in brackets. All NOCLs with subjunctive verbs, existential there, raised subjects, preceding co-ordinate clauses or indirect objects, or intervening adverbials excluded.

Matrix/NOCL subject	TC A	TC N
+Coreference	4+ 48= 52 (92.3%)	5+ 25= 30 (83.3%)
-Coreference	6+ 8= 14 (57.1%)	13+ 35= 48 (72.9%)
S U M	10+ 56= 66 (84.8%)	18+ 60= 78 (76.9%)
Chi-square	$X^2 = 8.05$; $p<0.01$	-

The Chi-square test shows the difference in the THAT/ZERO distribution depending on whether or not the two subjects are coreferential to be statistically significant in TC A but not in TC N. (The difference is significant if both TCs are considered together: $X^2 = 7.51$.)

The very marked difference recorded in TC A might be due to differences among the various matrix verbs as regards the relative frequencies with which they co-occur with coreferential and with non-coreferential NOCL subjects. The matrix verb SAY, which accounts for more than half the constructions of this type occurring in TC A, is of special interest in this connection. When checked, however, the proportion of NOCLs which have coreferential subjects turned out to be exactly the same with this verb as with all the matrix verbs combined: 78.9% v. 78.8%!

The results set out in Table 4a and Table 4b seem to lend strong support to the hypothesis that zero connective marks a closer clause

juncture than *that* connective. This may explain why ZEROs are particularly common among NOCLs with coreferential subjects. And the further differences depending on the person of both the matrix and the NOCL subjects appear to point in the same direction: it is not surprising if the link between the two clauses is felt to be especially close in cases where the writer and/or the addressee are directly involved in the message being conveyed. In particular, it is noteworthy that not a single THAT was recorded in either TC among constructions where both the matrix and the NOCL subjects are in either the 1st or the 2nd person, irrespective of coreference.

The fact that among constructions with non-coreferential NOCL subjects in the 3rd person ZEROs are more common if this subject is realised by a personal pronoun than if it is realised by some other, possibly quite complex, noun phrase is evidence that the vertical differences in Table 3 were not only due to variation of the person of the NOCL subject: the differences depending on the distinction between personal pronouns and other noun phrases obtain also when the factor person of NOCL subject is held constant, which corroborates our assumption that the weight and/or complexity of the NOCL subject is important for the choice of connective.

It may be wondered whether structural complexity in the NOCL subject generally makes for higher proportions of THATs. One straightforward measure of the complexity of a noun phrase is the number of words it is made up of. Table 5 gives the THAT/ZERO distributions among object NOCLs in TC A and TC N according to the length of the NOCL subject, expressed in number of words, for the constructions where this subject is not realised by a personal pronoun.

The table shows that in both TCs the proportion of ZEROs is higher among constructions with one-worded NOCL subjects than among other constructions, even when those with NOCL subjects realised by personal pronouns are disregarded. However, the difference is rather small in TC A, where there is further a slight tendency in the opposite direction when one compares the figures for 2 words and those for 3 or more words. The tendency for the proportion of THATs to increase with the length of the NOCL subject is much more conspicuous in TC N, but there the number of occurrences is too small for the Chi-square test to yield reliable results. However, when one takes into account

Table 5. Distribution THAT/ZERO among object NOCLs with subjects not realised by personal pronouns according to length of NOCL subject, in number of words. THAT+ZERO=NOCL, with percentages of ZERO in brackets. All NOCLs with subjunctive verbs, existential *there*, raised subjects, preceding co-ordinate clauses or indirect objects, or intervening adverbials excluded.

Length of NOCL subject	TC A	TC N
1 word	7+ 11= 18 (61.1%)	6+ 7= 13 (53.8%)
2 words	26+ 20= 46 (43.5%)	14+ 7= 21 (33.3%)
3 or more words	30+ 29= 59 (49.2%)	11+ 1= 12 (8.3%)
S U M	63+ 60=123 (48.8%)	31+ 15= 46 (32.6%)

that ZEROs have been found to be particularly common in the case of NOCL subjects realised by personal pronouns, there can be seen to be a clear overall tendency in both TC A and TC N for the proportion of THATs to increase with the length of the NOCL subject. This suggests that a motive for selecting *that* connective can be a desire to contribute to greater syntactic clarity in cases of structural complexity just after the matrix/NOCL boundary.

Structural complexity of this kind does not, of course, mean ambiguity, either temporary or permanent: constructions with long NOCL subjects are no more ambiguous than constructions with short NOCL subjects. This is thus a different type of conditioning factor from the one we noted in our comparison of constructions with and without intervening adverbials.

The differences depending on coreference are still unaccounted for. In an attempt to find out how it could be that, especially in TC A, constructions with NOCL subjects realised by personal pronouns exhibit a higher proportion of ZEROs if the pronoun is coreferential with the matrix subject, I examined the structures of the other NOCL subjects more closely. It was found that a rough but useful classifi-

Table 6a. Distribution THAT/ZERO among object NOCLs according to type of NOCL subject. THAT+ZERO=NOCL, with percentages of ZERO in brackets. All NOCLs with subjunctive verbs, existential there, raised subjects, preceding co-ordinate clauses or indirect objects, or intervening adverbials excluded. 'X' signifies any word or word group, or zero.

Type of NOCL subject	TC A	TC N	Type acc. to THAT/ZERO predom.
Personal pronoun	10+ 57= 67[8] (85.1%)	18+ 61= 79[9] (77.2%)	I
the + X	18+ 33= 51 (64.7%)	10+ 7= 17 (41.2%)	I
a(n) + X	5+ 3= 8 (37.5%)	1+ 0= 1 (0.0%)	II
Possessive + X	1+ 3= 4 (75.0%)	4+ 0= 4 (0.0%)	-
Demonstrative + X	1+ 5= 6 (83.3%)	0+ 2= 2 (100.0%)	I
(not +) Indefinite det./pron. + X[10]	11+ 3= 14 (21.4%)	6+ 0= 6 (0.0%)	II
Numeral + X	2+ 0= 2 (0.0%)	3+ 0= 3 (0.0%)	II
(N/Adj +) N_{common} + X	16+ 4= 20 (20.0%)	1+ 0= 1 (0.0%)	II
(N +) N_{proper} + X[11]	9+ 7= 16 (43.8%)	4+ 5= 9 (55.6%)	I
N_{proper}'s + X + N	0+ 1= 1 (100.0%)	2+ 0= 2 (0.0%)	-
S U M	73+116=189 (61.4%)	49+ 75=124 (60.5%)	

cation of these noun phrases could be achieved on the basis of (i) type of premodifying structure, if any, and (ii) type of head word.

The classification is set out in Table 6a, along with the respective distributions between THATs and ZEROs in TC A and TC N. For the sake of completeness, the constructions which have NOCL subjects realised by personal pronouns are also included in the table. With the two exceptions pointed out in the notes, these latter noun phrases consist of single words, whilst many of the noun phrases of the other types consist of two or more words. However, no distinction is made in this table according to length, since most of the classes are so small that such a distinction would have been uninteresting.

It can be seen that the THAT/ZERO distribution varies greatly according to the types of NOCL subject distinguished in the table. Although in most cases the number of constructions is too small to warrant any definite conclusions, it is noteworthy that the variation follows strikingly similar patterns in the two TCs: in both TC A and TC N ZEROs are common among constructions in which the NOCL subject is realised by a noun phrase which begins with the definite article or a demonstrative or is headed by a proper noun, besides those realised by personal pronouns; in TC N these are the only types of NOCL among which ZEROs were recorded at all; in TC A they all show more than forty per cent ZERO. The two TCs are in direct conflict only in the case of NOCL subjects introduced by possessives or genitive nouns, but then the number of such constructions is very small.

The constructions in which ZEROs are common, accounting for more than forty per cent of the recorded NOCLs in both TC A and TC N, will be referred to as Type I; those in which THATs predominate, accounting for more than sixty per cent of the recorded NOCLs in both TC A and TC N, will be referred to as Type II. Only the NOCLs with initial possessives or genitive nouns are incapable of being classified according to this distinction.

One should not attach too much importance to the individual results of the classification, because of the small number of constructions recorded in most of the classes of noun phrases we have distinguished. The question arises whether it is possible to find any features which

are shared by all or most of the constructions of either Type I or Type II but not by (most of) the constructions of the other type.

We notice that in TC A zero connective is common before the definite article and demonstratives, which in speech all begin with the voiced-lenis dental fricative, /ð̆/, a sound that may be somewhat awkward to pronounce just after the connective /ðət/, ending in the voiceless-fortis alveolar plosive. The fact that the connective begins with the same dental fricative may add to this feeling of awkwardness. On the assumption that factors which are directly relevant only to speech can spill over into written language, this might (help to) explain the recorded variation in the THAT/ZERO distribution. As, however, noun phrases beginning with a dental fricative are extremely rare in the other classes, this hypothesis cannot be further tested against my material. It may be noted, however, that no difference was detectable depending on whether the NOCL subject begins with a vowel or with a consonant.

It will further be seen that the distinction between the constructions of Type I and those of Type II is related to but does not coincide with the distinction between noun phrases introduced by grammatical words and noun phrases introduced by lexical words. This distinction is not, of course, entirely clear-cut, and it is furthermore doubtful whether it can be applied to proper names, which are used to refer but which, it can be argued, do not have sense, as these terms are used by many contemporary linguists (see Lyons 1977: 174ff.). With these qualifications it can be seen that on the whole constructions with NOCL subjects beginning with grammatical words tend to be ZERO, whilst constructions with NOCL subjects beginning with lexical words tend to be THAT.

As regards possible explanations for this tendency, it should be borne in mind that one of the main functions of the grammatical words under consideration is to signal (the beginning of) a noun phrase, i.e. they serve as syntactic transition markers, and it may therefore be that with such a word occurring initially in the NOCL the need for an additional transition marker, in the form of *that* connective, is felt less strongly. This argument is weakened, however, by the fact that with many matrix verbs the noun phrase signalled by such a grammatical word may have the function of direct

verbal object just as well as that of a subordinate clause (with the exception of pronouns with distinct nominative forms, which have been discussed before), so that there may still be a need for the connective to contribute to greater syntactic clarity.

Another possible explanation may be that lexical words are on the whole heavier than grammatical words: they carry a more distinct semantic load, they are more apt to be polysyllabic, and in speech they will usually be heavily stressed. Again the variation in the THAT/ZERO distribution may therefore be accounted for by weight-distributional relations: the need for an overt clause-boundary marker may be felt more strongly in constructions where, atypically, a heavy word occupies clause-initial position.

Some grammatical words are also heavy in terms of stress and/or number of syllables. That applies to many indefinite determiners/pronouns and to numerals and demonstratives. The fact that constructions with the two former types of NOCL subject are predominantly THAT thus corroborates the weight-distributional hypothesis.

Table 6b. Distribution THAT/ZERO among object NOCLs according to likely weight of first word in NOCL (subject). THAT+ZERO=NOCL, with percentages of ZERO in brackets. All NOCLs with subjunctive verbs, existential *there*, raised subjects, preceding co-ordinate clauses or indirect objects, or intervening adverbials excluded.

Likely weight of initial word in NOCL (subject)	TC A		TC N	
	With pers. pronouns	Without p. pronouns	With pers. pronouns	Without p. pronouns
Light	34+ 96=130 (73.8%)	24+ 39= 63 (61.9%)	33+ 68=101 (67.3%)	15+ 7= 22 (31.8%)
Heavy	39+ 20= 59 (33.9%)	39+ 20= 59 (33.9%)	16+ 7= 23 (30.4%)	16+ 7= 23 (30.4%)
S U M	73+116=189 (61.4%)	63+ 59=122 (48.4%)	49+ 75=124 (60.5%)	31+ 14= 45 (31.1%)
Chi-square	X^2 =25.66; p<0.01	X^2 = 8.48; p<0.01	X^2 = 9.18; p<0.01	-

The overall distributions between THATs and ZEROs according to the likely weight of the first word in the NOCL (subject) are set out in Table 6b, where all indefinite determiners/pronouns are counted

as heavy, along with numerals, demonstratives, and the various lexical words. Since the total figures are bound to be heavily influenced by the figures for personal pronouns and there may be several reasons, some of them irrelevant to our present concern, why these constructions are predominantly ZERO, it may be of some interest to compare the distributions also when the constructions with personal pronouns are disregarded. This and the following table therefore give overall distributions both with and without personal pronouns. The table shows that in TC A the difference in the THAT/ZERO distribution is statistically significant at 1% level even if constructions with NOCL subjects realised by personal pronouns are disregarded, whilst in that case the distributions are virtually the same in TC N.

Table 6a reveals that certain discrepancies still remain between the explanations we have suggested and the distinction we made between Type I and Type II constructions, on the basis of the observed distributions between THATs and ZEROs. There will be no discrepancies if instead one refers to the kinds of reference the various noun phrases are likely to express, distinguishing between definite or even unique reference on the one hand and indefinite reference on the other hand: the noun phrases of Type I are such as will normally be used for definite reference (personal pronouns and noun phrases beginning with the definite article or a demonstrative) or unique reference (proper nouns), whereas the noun phrases of Type II will typically be used to express indefinite reference (indefinite pronouns, common nouns not preceded by determiners, and noun phrases beginning with the indefinite article, indefinite determiners, numerals, or adjectives). The constructions which could not be classed as either Type I or Type II because they show different tendencies in the two TCs, those with NOCL subjects beginning with possessives or genitive nouns, will normally express definite reference.

It is true, of course, that the reference of a noun phrase cannot be predicted with certainty from its form, least of all when the form is as broadly described as in our classification, but these seem to be the most common kinds of reference expressed by each type of noun phrase. The respective distributions between THATs and ZEROs can be studied in Table 6c.

Table 6c. Distribution THAT/ZERO among object NOCLs according to likely reference of NOCL subject. THAT+ZERO=NOCL, with percentages of ZERO in brackets. All NOCLs with subjunctive verbs, existential there, raised subjects, preceding co-ordinate clauses or indirect objects, or intervening adverbials excluded.

Likely reference of NOCL subject	TC A		TC N	
	With pers. pronouns	Without p. pronouns	With pers. pronouns	Without p. pronouns
Definite/unique	39+106=145 (73.1%)	29+ 49= 78 (62.8%)	38+ 75=113 (66.4%)	20+ 14= 34 (41.2%)
Indefinite	34+ 10= 44 (22.7%)	34+ 10= 44 (22.7%)	11+ 0= 11 (0.0%)	11+ 0= 11 (0.0%)
S U M	73+116=189 (61.4%)	63+ 59=122 (48.4%)	49+ 75=124 (60.5%)	31+ 14= 45 (31.1%)
Chi-square	X^2 =34.04; p<0.01	X^2 =16.54; p<0.01	X^2 =15.80; p<0.01	X^2 = 4.79; 0.01<p<0.05

It will be seen that the Chi-square scores are in each case higher than in Table 6b. This time there is a significant difference even in TC N when constructions with personal pronouns are disregarded, although only at 5% level.

As is well known, definite reference can be of several different types. Anaphoric reference is probably the most common type in most kinds of written language, where extra-linguistic reference is usually less relevant (except in direct speech, quite common in TC N). This type of reference is perhaps more common with noun phrases containing the definite article than with certain other types of definite noun phrase, for instance noun phrases introduced by possessives or genitive nouns, which will often have a reference similar to that of cataphoric uses of the definite article: '*his/John's* house' = '*the* house that he/John owns', etc., but then we recall that constructions with such NOCL subjects varied a great deal in their choice between THAT and ZERO in the few cases where they were recorded. Besides, definite noun phrases may express generic reference.

It seems reasonable to assume, however, that in written language most types of definite noun phrase will usually have anaphoric

reference. The differences depending on the kind of reference the NOCL subject is likely to express are therefore reminiscent of the differences we observed during our examination of NOCLs with subjects realised by personal pronouns, when we suggested that the reason for the particularly high proportion of ZEROs among NOCLs whose subjects were coreferential with the respective matrix subjects might be that zero connective marks a closer clause juncture. It may more generally be the case that in constructions where the subject of an object NOCL has definite, anaphoric reference the NOCL will be felt to be more closely attached to the matrix clause than in other constructions, and that *that* connective will then be seen as too heavy a boundary marker, zero connective being selected instead.

With proper nouns it would be misleading, perhaps, to speak of the reference being anaphoric, but it is nevertheless a fact, ascertained by checking the respective contexts, that a majority of the NOCL subjects recorded of that type do refer to entities which have also been referred to in the preceding context; besides, it will be recalled that constructions with proper nouns exhibit lower proportions of ZEROs than some of the other classes likely to have definite/unique reference (see Table 6a).

The differences in the THAT/ZERO distribution that have been observed among object NOCLs depending on whether the NOCL subject is likely to have definite/unique reference on the one hand or indefinite reference on the other hand can thus be taken as support for our hypothesis that zero connective marks a closer clause juncture than *that* connective. It should be borne in mind, however, that the numerical differences are small in our material between this and the other distinctions (e.g. depending on weight) that have been suggested to account for the recorded THAT/ZERO distributions. Clearly, the fact that it was the distinctions based on the likely reference of the NOCL subject that was found to have the greatest conditioning force may be due to chance.

A further reason why this result should be treated with caution is that the distinction we set up according to type of reference was based on a rather broad classification of the forms of the recorded NOCL subjects and not on any examination of the reference they actually express in their respective contexts.

Finally, our discussion of the effect of different referential relations on the choice of connective has focused exclusively on the reference of the NOCL subject. In a full analysis other elements would also have to be taken into account.

Even so, the fact remains that the distinction based on the kind of reference the NOCL subject is likely to express corresponds exactly with the distinction between Type I and Type II, which was based on the THAT/ZERO distributions recorded in our material, in all cases where the latter distinction could be made. The hypothesis that zero connective marks a closer clause juncture than *that* connective is particularly difficult to dismiss because of the independent evidence provided by the difference among constructions with NOCL subjects realised by personal pronouns according to whether these pronouns are coreferential with the respective matrix subjects.

Perhaps the most reasonable reading of our findings is that the various factors we have discussed complement and often reinforce one another: that the choice of connective in object NOCLs may be conditioned both by whether the NOCL itself begins with another light word and by how closely the NOCL is felt to be attached to the preceding context; in addition there is some evidence to suggest that at least in TC A the resulting sound sequence can also affect the choice of connective, *that* being infrequent in that TC just in front of another word beginning with /ð/, viz. the definite article or one of the demonstratives.

The choice of connective in NOCLs is similar to the choice of relative pronoun in restrictive relative clauses where the pronoun does not function as subject and is not preceded by a preposition, in that both variables may assume zero as one of their possible values. The frequently noted tendency for zero relative to be preferred (besides *that*) after superlative antecedents ('the best book I have ever read') can thus be seen as a parallel of the tendency we have observed for zero connective to be selected in constructions with definite NOCL subjects: just as the anaphoric reference of many of these subjects can be assumed to strengthen the link between matrix clause and NOCL, so the cataphoric reference of superlative antecedents may strengthen the link between matrix clause and relative clause. In both cases zero seems to mark a particularly close clause juncture.

On the other hand, our conclusion that anaphoric NOCL subjects tend to favour zero connective may seem to be at variance with Bolinger's theory about the anaphoric force of *that* connective (see above, 2 (E)). What Bolinger claims, however, is not that *that* connective is preferred in cases where the NOCL itself, considered in isolation from its connective, is anaphoric, but rather that the connective is used to *give* the NOCL anaphoric force in cases where that is required by the context and the NOCL does not contain any other anaphoric element (see esp. Bolinger 1972:56ff.). In other words: *that* connective converts a non-anaphoric NOCL into an anaphoric one. Seen in that light, it is not surprising that ZEROs predominate among constructions where the NOCL subject is likely to have anaphoric reference, since in such constructions the connective would be redundant according to Bolinger's theory.

In order to test Bolinger's theory one needs to make a distinction between contexts which require and contexts which do not require anaphoric NOCLs, and then to study paradigmatic variations systematically, something which is far from easy on the basis of a corpus investigation. The question of the validity of Bolinger's theory will not, therefore, be further pursued.

5.3 The matrix verb phrase

It was established above that structural complexity at the beginning of the NOCL, as indicated by the length of the NOCL subject in number of words, is conducive to higher proportions of THATs (see Table 5). One may ask whether structural complexity just *before* the matrix/NOCL boundary has a similar effect on the choice of connective. In Table 7 the THAT/ZERO distribution is given according to the form of the main verb in the matrix verb phrase: whether this appears in the present or past tense, or as an infinitive, an *-ing* form or a past participle. The present and past tense forms will normally occur on their own, whilst the non-finite forms will frequently be part of more complex verbal constructions (often finite ones).

Table 7 shows that in TC A but not in TC N the proportion of ZEROs is consistently higher among the constructions in which the main matrix verb is finite than in the constructions in which this verb is non-finite. According to the Chi-square test, the difference is statistically significant at 5% level in TC A. (If applied to the

Table 7. Distribution THAT/ZERO among object NOCLs according to form of main matrix verb. THAT+ZERO=NOCL, with percentages of ZERO in brackets. All NOCLs with subjunctive verbs, existential *there*, raised subjects, preceding co-ordinate clauses or indirect objects, or intervening adverbials excluded.

Main matrix verb	TC A	TC N
Present	11+ 19= 30 (63.3%)	4+ 25= 29 (86.2%)
Past	42+ 83=125 (66.4%)	26+ 26= 52 (50.0%)
Sum finite	53+102=155 (65.8%)	30+ 51= 81 (63.0%)
Infinitive	6+ 7= 13 (53.8%)	15+ 15= 30 (50.0%)
-ing	9+ 3= 12 (25.0%)	3+ 7= 10 (70.0%)
-ed	5+ 4= 9 (44.4%)	1+ 2= 3 (66.7%)
Sum non-finite	20+ 14= 34 (41.2%)	19+ 24= 43 (55.8%)
Sum total	73+116=189 (61.4%)	49+ 75=124 (60.5%)

aggregate figures for finite and non-finite main matrix verbs, the test yields $X^2 = 6.13$.) The most obvious explanation is that this is a parallel to the difference we observed in our discussion of the NOCL subject: *that* connective tends to be preferred in cases of a high degree of structural complexity near the matrix/NOCL boundary, so as to contribute to greater syntactic clarity.

In TC N a more conspicuous difference is the one within the finite group, between constructions in which the main matrix verb is in the present tense and constructions in which it is in the past tense: ZEROs are considerably more common if this verb is in the present tense, a difference which can be shown to be statistically significant at 1% level ($X^2 = 8.97$). This difference is all the more noteworthy because it agrees with a difference observed by Aijmer (1967) in a corpus which was made up of British novels, and thus was fairly similar to our TC N in terms of genre.

The reason for the recorded difference in TC N is not immediately clear. It may have to do with differences among the various lexical verbs taking object NOCLs as regards the relative frequency with

which they occur in the present v. the past tense. Concerning TC A, it is a fact that the verb SAY, which is extremely common in that TC and which favours ZEROs, most often occurs in the past tense, and thus contributes to the high proportion in TC A of ZEROs among constructions in which the main matrix verb is in the past tense. However, a more detailed analysis of the relationship between the tenses of the various matrix verbs and their preferences for THATs or ZEROs would be required before one could say anything more definite about the reasons for the observed differences between present and past tense verbs in TC N, not parallelled by any similar difference in TC A. It should be recalled, however, that no differences were found between THAT and ZERO constructions with respect to the semantic class of the matrix verb in any of the TCs.

At least in TC A we have found further evidence that structural complexity near the matrix/NOCL boundary contributes to higher proportions of THATs among object NOCLs. This leads us to ask whether structural complexity generally makes for higher proportions of THATs. Several measures of structural complexity can be conceived of. We shall look at two: sentence length and sentence depth.

5.4 Sentence length

It might be thought that the THAT/ZERO distribution would depend on the length of the NOCL, for example. However, one's intuitive impression of that length proved exceedingly difficult to measure objectively, as in many cases it depends crucially on whether or not adverbials, non-finite subclauses, etc., are considered to be part of the NOCL, and objective criteria to decide these questions are hard to come by. It was concluded that the only constructional unit whose length could be determined reliably was the typographical sentence containing the NOCL. Table 8 gives the THAT/ZERO distributions among the recorded constructions, distinguished according to the three main types of NOCL subject we have operated with, depending on whether the typographical sentence contains less than 20 words, or 20 words or more.

As can be seen from Table 8, there is a slight overall tendency in both TCs for the proportion of THATs to increase with the length of the sentence, although this tendency is not consistent through all the sub-classes set up on the basis of type of NOCL subject. It is

obvious that the differences are not statistically significant in respect of any of the sub-classes, or in respect of the overall distribution in TC A. The table shows that in TC N, however, the difference in the THAT/ZERO distribution is significant at 5% level when all three sub-classes are put together, but then such a comparison of overall occurrences can be misleading, because it ignores the interdependence between sentence length and type of NOCL subject.

Table 8. Distribution THAT/ZERO among object NOCLs according to length of typographical sentence. THAT+ZERO=NOCL, with percentages of ZERO in brackets. All NOCLs with subjunctive verbs, existential *there*, raised subjects, preceding co-ordinate clauses or indirect objects, or intervening adverbials excluded.

Number of words in sentence / NOCL subject	Coref. prs. pronoun	Non-coref. p. pronoun	Other	S U M
TC A				
Less than 20	1+ 17= 18 (94.4%)	3+ 3= 6 (50.0%)	15+ 17= 32 (53.1%)	19+ 37= 56 (66.1%)
20 or more	3+ 31= 34 (91.2%)	3+ 5= 8 (62.5%)	48+ 43= 91 (47.3%)	54+ 79=133 (59.4%)
S U M	4+ 48= 52 (92.3%)	6+ 8= 14 (57.1%)	63+ 60=123 (48.8%)	73+116=189 (61.4%)
Chi-square	-	-	-	-
TC N				
Less than 20	4+ 20= 24 (83.3%)	8+ 26= 34 (76.5%)	13+ 9= 22 (40.9%)	25+ 55= 80 (68.8%)
20 or more	1+ 5= 6 (83.3%)	5+ 9= 14 (64.3%)	18+ 6= 24 (25.0%)	24+ 20= 44 (45.5%)
S U M	5+ 25= 30 (83.3%)	13+ 35= 48 (72.9%)	31+ 15= 46 (32.6%)	49+ 75=124 (60.5%)
Chi-square	-	-	-	$X^2 = 5.51$; $0.01 < p < 0.05$

That there is nevertheless a distinct difference in average sentence length between the THATs and the ZEROs recorded in our material comes out more clearly in Table 9, which simply gives average number of words per typographical sentence in respect of each sub-class in the two TCs. In order to make it easier to assess the importance of the

Table 9. Average length of typographical sentence, in number of words (connective that not counted), of THATs and ZEROs according to type of NOCL subject, with number of recorded constructions in brackets. All NOCLs with subjunctive verbs, existential there, raised subjects, preceding co-ordinate clauses or indirect objects, or intervening adverbials excluded.

NOCL subject		Coref. prs. pronoun	Non-coref. p. pronoun	Other	All types
TC A	THAT	21.8 (4)	25.5 (6)	27.1 (63)	26.7 (73)
	ZERO	22.9 (48)	21.9 (8)	24.5 (60)	23.6 (116)
TC N	THAT	18.8 (5)	17.6 (13)	21.3 (31)	20.1 (49)
	ZERO	17.2 (25)	15.3 (35)	16.9 (15)	16.3 (75)

various figures, the number of constructions recorded in each subclass is given in brackets.

It will be seen that with the exception of the constructions in TC A where the NOCL subject is realised by a coreferential personal pronoun - where the small number of recorded THATs makes the comparison highly unreliable - the average length of THAT sentences is in each case higher than that of ZERO sentences. Even if these findings do reflect real underlying differences, however, these do not seem to be much greater than what is accounted for by the differences depending on the complexity of the NOCL subject and the matrix verb phrase that were noted above. We can conclude that as far as the complexity of linear structure is concerned it is mainly elements located near the matrix/NOCL boundary that affect the distribution between THATs and ZEROs.

5.5 Sentence depth

Another, more direct, measure of structural complexity is the depth of embedding reached in each sentence. By means of the computer I examined how deeply each NOCL was embedded, and also how many levels of clauses were embedded within each NOCL. ('Clause' is here used in the sense in which the term is employed in the Syntax Data Corpus, i.e. non-finite as well as finite clauses were counted - *want to go*,

for example, would be analysed as two clauses. For further details, see Ellegård 1978.)

The number of clauses embedded below each NOCL turned out not to have any definite effect on the THAT/ZERO distribution. The distribution between THATs and ZEROs did vary, however, depending on how deeply the NOCL itself was embedded. The figures are set out in Table 10,[12] which shows that in both TC A and TC N the proportion of ZEROs increases markedly with the depth of the embedding, although it should be noted that the percentages for three and four levels of embedding are not very reliable, because of the small number of constructions recorded; the Chi-square test does not, therefore, show statistical significance in either TC. However, if both TCs are considered together - something which is not wholly unjustified, since the THAT/ZERO distribution is closely similar in the two TCs - the difference between NOCLs at one or two levels of embedding on the one hand and those at three or four levels on the other hand is statistically significant at 5% level ($X^2 = 5.12$).

Table 10. Distribution THAT/ZERO among object NOCLs according to level of embedding. THAT+ZERO=NOCL, with percentages of ZERO in brackets.

Clause levels above NOCL	TC A	TC N
One	100+100=200 (50.0%)	47+ 52= 99 (52.5%)
Two	33+ 38= 71 (53.5%)	17+ 29= 46 (63.0%)
Three	4+ 9= 13 (69.2%)	2+ 8= 10 (80.0%)
Four	1+ 3= 4 (75.0%)	1+ 4= 5 (80.0%)
S U M	138+150=288 (52.1%)	67+ 93=160 (58.1%)

On the assumption that TC A and TC N are representative of their respective genres on this point, it may be wondered why the proportion of ZEROs increases with the depth at which the NOCL is embedded, and furthermore is independent of the number of levels below the NOCL. If structural complexity generally makes for a smaller proportion of

ZEROs, one would rather have expected the proportion of ZEROs to decrease as the number of clause levels, either above or below the NOCL, increased.

One thing which may help to explain the recorded difference in the effect on the THAT/ZERO distribution is the linear difference that there will normally be between embeddings above and below object NOCLs: clauses appearing above the NOCL in the hierarchical sentence structure are likely to precede it in the surface realisation of the sentence, whilst clauses appearing below are likely to follow it. One reason why the number of levels above but not the number of levels below the NOCL influences the choice of connective may therefore be that at the time when that choice is made, the writer is not always conscious of elements which follow the connective, often at some distance removed from it.

Another thing which probably contributes to the difference in their effect on the distribution between THATs and ZEROs is that, when checked, the vast majority of the clauses appearing above the NOCLs were found to be finite ones, whilst a majority of the clauses embedded below are non-finite, and thus on the whole shorter, lighter constructions.

As regards the question why the proportion of ZEROs increases with the depth of embedding of the NOCL itself, it should be noted that Ellegård recorded a tendency for the clauses in the Syntax Data Corpus to get shorter as the depth of embedding increases (as 'clause' was defined in the Syntax Data project; see Ellegård 1978:23ff.).

Furthermore - and this is probably the more important factor - it does not seem unreasonable to assume that the more clause boundaries there are in a sentence, the less significant each boundary will be felt to be. The larger proportions of ZEROs among the more deeply embedded NOCLs can thus be seen as a manifestation of the general tendency for the proportion of ZEROs to increase with the closeness of the association between NOCL and matrix clause.

6 SUMMARY AND CONCLUSION

Our investigation of the distribution between THATs and ZEROs among the NOCLs occurring in the Syntax Data Corpus has revealed evidence of the following conditioning factors:

(a) *Style*. The assumption that THATs predominate in formal language and that the proportion of ZEROs increases as the style becomes less formal has been amply confirmed through differences among the various TCs: in TC G and especially TC J THATs have been found to be the rule and ZEROs the exception, whilst in TC A and TC N THATs and ZEROs are about equally frequent. The assertion that ZEROs are particularly rare in scientific writings (see above, 2 (A)) has thus been borne out. Further evidence of the influence of the stylistic factor can be seen in the fact that among object NOCLs in TC A and TC N, and to some extent also TC G, there is a tendency for the proportion of ZEROs to increase with the frequency of the matrix verb.

(b) *Syntactic function*. In most of the syntactic functions we have distinguished the number of recorded NOCLs has been too small to warrant any definite conclusions about the THAT/ZERO distribution. However, in all four TCs the proportion of ZEROs is somewhat higher in the most numerous syntactic function, that of non-extraposed, non-complementary object in an active matrix clause, than among all syntactic functions combined. (This difference is not very interesting in respect of TC J, with just two recorded ZEROs.) In the one TC where both passive matrix constructions and ZEROs are numerous, TC A, combinations of the two have been found to be quite common. More surprisingly, and contrary to what several grammarians have reported before (see above, 2 (B)), a certain number of ZEROs have also been recorded among the NOCLs complementing nouns, although the proportion of ZEROs in this function is lower than among non-complementary NOCLs and NOCLs complementing adjectives.

(c) *Potential ambiguity*. Examination of the constructions recorded with object NOCLs in TC A and TC N has shown that the proportion of THATs increases markedly if an adverbial occurs between the matrix verb and the NOCL subject. This must be due to a (conscious or unconscious) desire on the part of the writer to avert ambiguity.

(d) *Structural complexity near clause boundary*. Among object NOCLs in TC A and TC N the proportion of THATs increases with the complexity of the NOCL subject, and in TC A also with the assumed complexity of the matrix verb phrase: in both TCs THATs are less frequent if the NOCL subject is realised by a personal pronoun than if it is realised by some other kind of noun phrase, and in the latter case

the proportion of THATs shows a further tendency to increase with the length of the noun phrase; in TC A, moreover, THATs are more frequent if the main matrix verb is non-finite than if it is finite. Although there is no risk of ambiguity in such cases, one may see the selection of *that* connective as a contribution to greater syntactic clarity.

(e) *Weight-distributional relations*. Some evidence has been unearthed to suggest that *that* connective is more likely to be chosen if either the matrix clause or the NOCL deviates from the most common weight-distributional pattern in English clauses, characterised by light elements in initial position and heavier elements towards the end. Such a principle may explain why the non-clausal intervening adverbials recorded in ZERO constructions (in TC A) without exception belong in the matrix clause rather than in the NOCL, and it can also be seen as a supplementary (or even alternative) explanation of the observed tendency for the proportion of THATs to increase with the complexity of the NOCL subject. Like (d) above, this conditioning factor may thus be regarded as evidence of a tendency for *that* connective to be selected in constructions where there is felt to be a need for greater syntactic clarity, either because of a comparatively high degree of structural complexity near the clause boundary - conditioning factor (d) - or because of conflicting linguistic signals - conditioning factor (e).

(f) *The closeness of the clause juncture*. We have advanced the theory that, at least as far as object NOCLs are concerned, zero connective marks a closer link between the matrix clause (preceding context) and the NOCL than *that* connective. Evidence for this has been found among constructions with object NOCLs in TC A and TC N. The most important evidence is:

(i) the fact that among constructions with NOCL subjects realised by personal pronouns zero connective has been found to be (even) more frequent if the NOCL subject is coreferential with the matrix subject, especially in TC A; and

(ii) the fact that zero connective has been found to be common if the NOCL subject is likely to have definite, presumably often anaphoric, reference, whilst *that* connective predominates in other cases, although the demonstrative force of this observation is weakened somewhat by the close interdependence between the distinction based on

likely reference and certain other distinctions that could be regarded as potential conditioning factors, such as the one based on the weight of (the initial word in) the NOCL subject.
The theory is further corroborated by:
(iii) the observed tendency for the proportion of ZEROs to increase with the depth at which the NOCL is embedded (even though the number of constructions recorded with more than two clause levels above the NOCL is small, so that the statistical significance of this tendency is somewhat uncertain); we have suggested that a major reason for this tendency may be that the importance attached to each clause boundary diminishes as the number of such boundaries within the sentence increases; and
(iv) the fact that only zero connective has been recorded in constructions where both the matrix and the NOCL subjects are in the 1st/2nd person; it is possible that in such cases the link between the two clauses is felt to be particularly close.

Our study has largely been confined to TC A and TC N - factors (c)-(f) above are based on evidence from these two TCs only - where the stylistic factor seems to have less direct influence on the choice between THAT and ZERO than in TC G and TC J, and where a search for specific contextual conditioning factors could therefore be expected to be more rewarding. We have not examined whether, or to what extent, factors (c)-(f) are also operative in TC G and TC J. Since, however, the style of a text may itself function as a conditioning factor of various other conditioning factors, it is perfectly possible that part of the reason for the low proportion of ZEROs in TC G and TC J is, for example, that NOCL subjects realised by personal pronouns are particularly rare in those TCs, and that other types of NOCL subjects make for higher proportions of THATs there too.

Certain differences have been observed between TC A and TC N as regards the relative weight the various conditioning factors carry in these two TCs: the reference of the NOCL subject has a more marked effect on the choice of connective in TC A than in TC N, as shown both by the distinction among personal pronouns depending on coreference, and by the distinction among other noun phrases depending on definiteness; the tendency for zero connective to be selected in front of NOCL subjects beginning with the voiced-lenis dental fricative is distinct only in TC A; the finite/non-finite

opposition among main matrix verbs has been found to have greater conditioning force in TC A than in TC N; and ZERO constructions with intervening adverbials have been recorded only in the former TC; on the other hand, the effect of the length of the NOCL subject is more conspicuous in TC N, and only in that TC does the opposition between present and past tense matrix verbs appear to be important for the choice of connective.

Although these differences do not seem to conform to any clear pattern, there can be seen to be some evidence to suggest that factors which are most immediately relevant to spoken language have greater conditioning force in TC A than in TC N. Could it be that the kind of English used in newspaper reports, often written in great haste, is more directly influenced by the conditioning process characteristic of spoken English than the kind of English used in fiction, presumably composed with greater care and deliberation? The present investigation does not provide any basis for a definite answer to this question.

Conditioning factors (a), (b) and (c) in the above list are identical with those referred to under items (A), (B) and (C), respectively, in our introductory list of commonly recognised conditioning factors (see section 2 above). As regards the two other types of conditioning factors referred to in that list, (D): the matrix verb and (E): semantic contrast, few of the matrix verbs we have recorded in any considerable numbers occur exclusively with either THAT or ZERO objects, although we have found that the most frequent matrix verbs are more apt to take ZERO objects than the less frequent ones; as for Bolinger's semantic theory, this has not been subjected to any systematic testing, but item (f) in our new list of conditioning factors is related to Bolinger's theory in that it involves the reference of the NOCL.

Apart from that, items (d), (e) and (f) in the above list represent conditioning factors which do not appear to have been recognised before. The evidence for (d) and (e) is somewhat less conclusive than for the other conditioning factors we have posited: the various aspects of structural complexity just before and after the matrix/NOCL boundary will have to be gone into in greater detail before one can say with certainty what effect such complexity has on the choice

of connective, and a similar comment can be made about weight-distributional relations. Our investigation leaves little doubt, however, that *that* connective is used as a contribution to greater syntactic clarity not only in constructions where there would otherwise be a risk of ambiguity, but also in other cases where a need for added clarity may be felt.

As regards our theory that zero connective marks a closer clause juncture than *that* connective, the evidence is substantial, as we have just seen. Since, furthermore, this is a conditioning factor which seems to have been overlooked by previous investigators, it stands out as perhaps the most important result of our study.

Further research would be required to find out whether the conclusions we have reached on the basis of an examination of a limited selection of written American English are valid also for other varieties of English, such as spoken American English, spoken and written British English, etc. What we have found suggests that at least in contexts where the direct influence of the stylistic factor on this variable seems slight, the conditioning process determining the choice of connective in English nominal clauses is more complex than has usually been acknowledged.

NOTES

1 I am indebted to Prof. H. Spang-Hanssen for generous advice on the statistical methods employed and on other aspects of this study, and to cand.philol. Ivar Fonnes, Senior Lecturer at the University of Oslo, who did the necessary computer programming.

2 For a lucid account of this and the competing dualist view of the relationship between meaning and style, see Leech and Short (1981, esp. pp. 14-40).

3 In this and subsequent references to the corpus the initial letter indicates TC, the following double digit figure the number of the text within the TC, and the figure after the colon the line number within the text, according to the computer tape version of the Syntax Data Corpus received from the University of Gothenburg. (Line numbers may differ somewhat from other editions of the Brown Corpus.)

4 The number is slightly smaller than in Elsness (1981), where two clauses had been included by mistake, owing to tagging errors. Moreover, a couple of enumerative constructions in which each of several co-ordinate NOCLs is preceded by a numeral or alphabetically ordered letter were held to be uninteresting from the point

of view of the THAT/ZERO distribution, and hence omitted from the investigation reported in the present article.

5 On the compilation of the Brown Corpus, see Francis (1979).

6 These and subsequent italics are mine.

7 In this and subsequent applications of the Chi-square test to 2x2 tables I used a version of the test corrected for continuity recommended by Siegel (1956:107-10) for all cases of $N>40$, and for $20 \leq N \leq 40$ if all expected frequencies are 5 or more.

8 One construction with the NOCL subject *they (the detectives)* has not been included in previous references to NOCL subjects realised by personal pronouns.

9 One construction with the NOCL subject *us Baptists* has not been included in previous references to NOCL subjects realised by personal pronouns.

10 This class comprises forms like *all, many, several, both, some, any, more, most, whatever, other*, which can function as both determiners and pronouns, and also *none* and forms of the type *everybody, anything*, which invariably occur in pronoun function.

11 The nouns which in a few cases premodify the recorded proper names are titles of the type *President* and *Mr.*, and further *apprentice* and *halfback*, which can be said to have a clearer lexical content.

12 In this table *all* object NOCLs are included, irrespective of the occurrence of subjunctives, intervening adverbials, etc., so that the figures given here are not directly comparable with the ones in the preceding tables. The reasons for this difference are technical.

REFERENCES

Aijmer, K. 1967. An Investigation of the Factors Conditioning the Choice of Connective in Object Clauses in Recent English Fiction. Unpublished thesis for the degree of *fil.lic.*, University of Stockholm.

Bolinger, D. 1972. *That's That*. Janua Linguarum, Series Minor, 155. Mouton, The Hague and Paris.

Bolinger, D. 1977. *Meaning and Form*. English Language Series No. 11. Longman, London and New York.

Ellegård, A. 1978. *The Syntactic Structure of English Texts. A Computer-Based Study of Four Kinds of Text in the Brown University Corpus*. Gothenburg Studies in English 43. University of Gothenburg.

Ellinger, J. 1933. 'Substantivsätze mit oder ohne *that* in der neueren englischen Literatur'. *Anglia* 57. 78-109.

Elsness, J. 1981. 'On the Syntactic and Semantic Functions of *That*-Clauses'. In S. Johansson and B. Tysdahl, eds., *Papers from the First Nordic Conference for English Studies, Oslo, 17-19 September 1980*. University of Oslo. 281-303.

Fowler, H.W. 1965. *A Dictionary of Modern English Usage* (2nd ed.), revised by E. Gowers. Clarendon, Oxford.

Francis, W.N. 1979. 'Problems of Assembling and Computerizing Large Corpora'. In H. Bergenholtz and B. Schaeder, eds., *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora*. Scriptor Verlag: Königstein. 110-23. Also in S. Johansson, ed. 1982. *Computer Corpora in English Language Research*. Norwegian Computing Centre for the Humanities, Bergen. 7-24.

Hornby, A.S. 1954. *A Guide to Patterns and Usage in English*. Oxford University Press, London.

Jespersen, O. 1928. *A Modern English Grammar on Historical Principles III*. George Allen & Unwin Ltd, London, Ejnar Munksgaard, Copenhagen.

Kruisinga, E. 1932. *A Handbook of Present-Day English. Part II: English Accidence and Syntax*. P. Noordhoff, Groningen.

Leech, G.N. and M.H. Short. 1981. *Style in Fiction. A Linguistic Introduction to English Fictional Prose*. English Language Series No. 13. Longman, London and New York.

Lyons, J. 1977. *Semantics*. Vol. 1. Cambridge University Press, Cambridge.

McDavid, V. 1964. 'The Alternation of "That" and Zero in Noun Clauses'. *American Speech* 39. 102-13.

Oxford English Dictionary. 1933. J.A.H. Murray *et al.*, eds. Vols. 1-12 & supplement. Oxford.

Poutsma, H. 1929. *A Grammar of Late Modern English. Part I: The Sentence, Second Half: The Composite Sentence*. P. Noordhoff, Groningen.

Quirk, R., Greenbaum, S., Leech, G.N. and J. Svartvik. 1972. *A Grammar of Contemporary English*. Longman, London.

Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Kogakusha, Ltd, Tokyo.

Storms, G. 1966. '*That*-Clauses in Modern English'. *English Studies* 47. 249-70.

Ulvestad, B. 1956. 'An Approach to Describing Usage of Language Variants'. *International Journal of American Linguistics, Memoir 12*. 37-59.

Zandvoort, R.W. 1957. *A Handbook of English Grammar*. Longmans, Green and Co, London, New York, Toronto.

TEXTUAL ASPECTS OF TOPICALIZATION IN A CORPUS OF ENGLISH

Marita Gustafsson
University of Turku, Finland

1 INTRODUCTION

Topic, topicalization, theme, and *focus* are all problematic terms in present-day linguistics. Dahl (1974) tries to sort out the confusion on the basis of a logically oriented grammar. He starts with the old idea that the topic of a sentence is 'what the sentence is about' and that the comment is 'what is said about the topic' (p. 4). He then shows that this definition of the topic 'explains at least one restriction on topics, namely that they must be definite in some sense of this term' (p. 7). Quirk seems to think similarly when he says that 'It is arguable that even attempts to begin discourse with complete strangers are continuations, in the sense that we take as our "starting point" *topics* (a key word in this connection) known to be conventionally established as common ground in our society' (Quirk 1978:30). The normal place for the topic is the beginning of the sentence, where it serves as background for what is to be said. Quirk even uses the term *topicalized*, when he refers to the subject *Mary* in the sentence

(1) Mary reviewed a book of his quite recently. (p. 35)

In generative grammar, however, the term *topicalization* is used to denote rules which move constituents to sentence-initial position, e.g. object topicalization in

(2) *Him* I don't like, and *her* I've always hated.

If we compare *Mary* in (1) with *him* and *her* in (2), we will see that the 'topicalized' items have quite different motivations for their initial positions. While (1) is the unmarked order of a statement in English, (2) shows a highly marked word order which requires special motivation. Quirk is naturally aware of the various marked forms, and he gives an example of another type where stress and intonation are used to give the subject *Mary* rhetorical focus

(3) MARY reviewed a book of his. (Quirk 1978:35)

A third definition of the term *topicalization* is given by Enkvist, who makes a distinction between topicalizations and thematizations.

The former term is reserved for 'those topicalizing and commenting operations that merely move items about in the sentence, without affecting their syntactic functions'. Thematizations are operations like passivization where also the surface-syntactic functions of moved items change (Enkvist and von Wright 1978:52). Add to these definitions Chafe's statement that 'the so-called *topic* is simply a focus of contrast that has for some reason been placed in an unusual position at the beginning of a sentence' (Chafe 1975:49f.), and there could hardly be more confusion about the basic terminology.

It seems to me that at least part of the confusion arises from the difference between sentence-oriented and text-oriented approaches. In the present study, however, I am going to make use of both approaches by, first, collecting the material on the basis of generative topicalizations, and, then, analysing the material using textual parameters. This study is a modest attempt at empirical text linguistics called for by Enkvist, when he pointed out 'that text theory has been comparatively well developed and that further empirical work with the discoursal analysis of corpora of texts will now be in order to test the theories' (Enkvist 1978:188).

2 MATERIAL AND METHODS

The development of large computer corpora of English has made it possible to scan long enough passages in various genres for the detection of even infrequent phenomena. Unfortunately, so far only a few of the corpora include grammatical tagging, although work for that purpose is proceeding on several corpora (see *ICAME News*, nos. 2 and 3). Grammatical tagging is essential for all kinds of syntactic analysis. The corpus used in the present study is a tagged version of the original Brown Corpus, prepared by Alvar Ellegård and his students from the English Department at Gothenburg University. Four genres of the Brown Corpus were selected for detailed grammatical tagging, viz. journalism, science, popular fiction, and literary essays. The Gothenburg corpus consists of 64 texts (16 from each genre) of about 2,000 words, or approximately 128,000 words in all. The texts and their grammatical codes were prepared for automatic data processing, and the whole material is now available on magnetic tape (for more information about the corpora, see Kučera and Francis 1967; Ellegård 1978).

Topicalizations in the generative sense result in a word order where the sentence begins with an element other than the subject-NP, and we get surface orders like OSV, AdvSV, CSV, etc. I was interested in finding out (1) how frequent these marked word order patterns are in various genres of English, and (2) what the possible motivations are for these patterns, as far as one can determine the motivation from the context. For the detection of these cases a computer programme was compiled at the University of Turku.[1] The clause is taken as the basic unit for the search, and the constituent order of each clause is checked for patterns specified at the beginning of the search. If a clause with the specified order is found, the whole clause is printed with all the information available in the Gothenburg version. The programme can cope with all types of embedded and discontinuous clauses, but the search was deliberately restricted to finite clauses, whether continuous or discontinuous in surface structure. The word order patterns specified for the search were

(1) OSV and OVS (O = direct object)
(2) O_iSV and O_iVS (O_i = indirect object)
(3) O_pSV and O_pVS (O_p = prepositional object)
(4) AdvSV and AdvVS (Adv = all kinds of adverbials, Ellegård's classes A, B, C, M, D, and E)
(5) C_sSV and C_sVS (C_s = subject complement)
(6) C_oSV and C_oVS (C_o = object complement)

In other words, the programme prepared for me a printout with all finite clauses which had initial objects, adverbials, or complements. Pure conjunctions were excluded from the search, as they are restricted to initial position, but more than one item of the specified type as well as combinations of specified types were allowed for, e.g. AdvAdvSV or OAdvSV.

As the grammatical tagging of the Gothenburg version contains no textual parameters, the cases yielded by the search were recoded using textual parameters. The new coding scheme first contains information necessary for the identification of cases from the complete Brown Corpus, then some grammatical information (type and mood of the clause) from the tagged corpus, and finally fifteen new variables for obtaining information on some textual aspects of the topicalized cases. The adverbial class is first subclassified into obligatory adjuncts (= valency adverbs; see Enkvist 1976), free

adjuncts, disjuncts, and conjuncts. Fronted adverbials are further classified according to their semantic class, because the classification of adverbials in Ellegård's code is too vague to warrant any decision-making on disjuncts and conjuncts, for example. The system used follows Quirk and Greenbaum (1973) in a slightly simplified form. The textual parameters concentrate on the moved constituent as well as the final nominal constituent within the same clause. These constituents are assessed as to their function, number, structure, and length. The givenness of the fronted and final constituents is also evaluated, and in the case of given items the linguistic device used in referring is considered. As the search programme picks out all clauses with initial objects, adverbials, and complements, there will be many cases where the word order is not the result of topicalization proper but of relativization or question transformations. In these cases the existing word order is the only grammatical one. Therefore the coding scheme contains a parameter for the assessment of the probable motivation for the movement. Another parameter classifies further the linguistic devices used in the obligatory grammatical movements. The complete coding scheme is presented in the Appendix. In designing the code I benefited greatly from a similar coding system used in a Finnish project (Hakulinen, Karlsson, and Vilkuna 1980; see also Kohonen 1978). After recoding the material was punched onto cards and subjected to a statistical analysis using SPSS (= Statistical Package for Social Sciences).

3 SOME STATISTICAL RESULTS OF THE ANALYSIS

3.1 Frequency of topicalizations

The total number of clauses in the Gothenburg corpus is 17,862; out of these approx. 13,050 (72.9%) are finite clauses, which constitute the material of the present study. The total number of topicalized cases among finite clauses is 3,511. This means that in approx. 27% of all finite clauses the word order is other than S- or V-initial (VS is also regarded as an unmarked word order, e.g. in questions). The actual number of clauses is slightly smaller, however, as 530 fronted elements have been moved together with another element. This result, by the way, contrasts with some ideas presented about the applicability of topicalization to only one element in a clause (Kohonen 1978:160). The distribution of clause types in the Gothen-

burg corpus (henceforth = GC) is as follows: main clauses 53.6%, subordinate, non-relative clauses 32.1%, and relative clauses 15.3%. The corresponding figures for the topicalized material (= TM) are 74%, 8%, and 18%. The proportion of relative clauses is almost the same, but the low figure for subordinate clauses in TM shows that topicalization is more common in main clauses. This result agrees with some earlier studies (see Kohonen 1978:254ff.). When the clause type is crosstabulated by the probable motivation for the topicalization (variable 20), we will see that almost half of the cases in subordinate clauses (43.6%) are grammatically obligatory, most of them indirect questions. The number of 'real' topicalizations is thus cut down to a little over 100 in subordinate clauses, which is only 3% of all cases.

As far as the mood of the clause is concerned, TM seems to follow the distribution in GC. The overwhelming majority of finite clauses are declarative (93.4%), compared with 95.4% of *all* clauses in GC. Imperatives and exclamations are almost non-existent both in GC and TM, but interrogatives are slightly more represented in TM than in GC (5.7% vs. 2.6%).

As mentioned above, the total number of topicalizations is 3,511. Table 1 gives the distribution of these cases according to topicalized sentence element and genre. The labels used for the genres are those of the Brown Corpus: A = journalism, J = science, N = popular fiction, and G = literary essays.

Table 1 reveals some interesting differences between the frequencies of topicalized elements in different genres. Topicalization of free adjuncts is by far the most common type in all genres; more than half of all cases belong to this class. In journalism (A) and popular fiction (N) objects take the second place with approximately one third of cases in each genre in that category. Science represents the other extreme with less than 4% of object topicalizations. A closer look at the latter cases even showed that they are all grammatically motivated, i.e. relative pronouns or question words. I have elsewhere (Gustafsson, forthcoming) discussed object topicalizations in more detail, but it could be pointed out here that the high number of initial objects in genres (A) and (N) is mostly due to the frequency of quotations, whether direct or indirect. Newspapers cite their sources by mentioning the speaker in a comment

Table 1 Distribution of cases according to element and genre.

Fronted sentence element / Genre	A		J		N		G		total	
	N	%	N	%	N	%	N	%	N	%
1 Direct object	206	33.0	27	3.8	357	31.1	155	15.2	745	21.2
2 Indirect object	-		-		1	0.1	-		1	0.0
3 Prepositional object	6	1.0	2	0.3	9	0.8	26	2.5	43	1.2
4 Obligatory adjunct	8	1.3	-		2	0.2	-		10	0.3
5 Free adjunct	301	48.2	420	58.7	614	53.4	585	57.2	1920	54.7
6 Disjunct	14	2.2	53	7.4	36	3.1	58	5.7	161	4.6
7 Conjunct	79	12.7	197	27.5	101	8.8	174	17.0	551	15.7
8 Subject complement	9	1.4	17	2.4	27	2.3	24	2.3	77	2.2
9 Object complement	1	0.2	-		2	0.2	-		3	0.1
	624	100.0	716	100.0	1149	100.0	1022	100.0	3511	100.0

clause that follows the citation, and popular fiction contains a lot of dialogue. This is evident on the basis of the type of verb in the clauses where the topicalization takes place. The proportion of verbs of saying and thinking is quite high in the quoting genres: 40.9% in (A) and 24.7% in (N) as opposed to 3.4% in (J) and 9.4% in (G). As a matter of fact, the object-initial order is so common and non-emphatic in (A) and (N) that it has been classified as unmarked as far as the motivation is concerned (variable 20).

Another difference seen in Table 1 is the proportion of disjuncts and conjuncts in various genres. It is hardly surprising to notice that the use of these cohesive devices is highest in genres (J) and (G). Scientific texts, in particular, are argumentative and explicative in function, and their writers have to be clear and logical. They also want to show explicitly how the text is organized, and what truth value can be given to the opinions and results presented in the text. Therefore, disjuncts are well represented in science, as they are also in literary essays, where they probably emphasize the writers' personal opinions and likings rather than detached reasoning.

A final note on the frequency of topicalizations concerns obligatory adjuncts. These are adverbs which have a close bond with the verb,

e.g. put *on the table*. Enkvist calls them valency adverbials, and points out that they are harder to topicalize as they are deeper in the clausal structure (Enkvist 1976:56). The present results agree with his findings.

3.2 Structure and length of the moved and final constituents

The organization of elements in a sentence is guided by several forces working simultaneously. First and foremost, constituent order in English is determined by grammar, but as text linguists, beginning with Vilém Mathesius, have shown, various grammatical devices can be used to convey information in a textually appropriate perspective. The principle of *end-weight* works alongside the principles of *end-focus* and *actuality*. The first two terms are well-known (see e.g. Quirk and Greenbaum 1973:406ff.). The third comes from Jespersen's *Modern English Grammar*. Jespersen explains the principle of actuality by saying that the speaker tends to express first what is uppermost in his mind. In ordinary circumstances the front position is given to the subject, but under the influence of stress and emotion other elements may take the front position, thus counteracting the principles of end-weight and end-focus (Jespersen 1961,VII:54).

From the point of view of constituent order the initial and final positions in a clause seem to be more important than the rest. Apparently, there is also some kind of interplay between these positions, and elements can be moved from one to the other. That is why, in the present analysis, both positions have been studied and their interaction determined by crosstabulation. First, the structures of topicalized elements are presented separately for the fronted elements and the final constituents.

There is a clear difference between the structures of fronted and final constituents. The most common fronted element consists of one word, and the other structures follow in order of complexity. One exception to the tendency appears in complex clauses which, especially in journalism (A), are more numerous than simple non-finite clauses. This is probably due to the unmarked cases of quotation mentioned earlier (p.50).

Table 2 Structure of fronted element according to genre.

Str. of fronted element	Genre A N	%	J N	%	N N	%	G N	%	Total N	%
1 One word	197	31.6	278	38.8	493	42.9	438	42.9	1406	40.0
2 Phrase	200	32.1	284	39.7	279	24.3	366	35.8	1129	32.2
3 Simple fin. clause	111	17.8	75	10.5	225	19.6	111	10.9	522	14.9
4 Simple non-fin. clause	22	3.5	32	4.5	63	5.5	43	4.2	160	4.6
5 Complex cl.	94	15.1	47	6.6	89	7.7	64	6.3	294	8.4
	624	100.0	716	100.0	1149	100.0	1022	100.0	3511	100.0

Table 3 Structure of final constituent according to genre.

Str. of final const. Genre A N	%	J N	%	N N	%	G N	%	Total N	%
1 One word	17	2.7	32	4.5	151	13.1	74	7.2	274	7.8
2 Phrase	200	32.1	295	41.2	380	33.1	386	37.8	1261	35.9
3 Simple fin. clause	40	6.4	25	3.5	49	4.3	29	2.8	143	4.1
4 Simple non-fin. clause	20	3.2	34	4.7	41	3.6	55	5.4	150	4.3
5 Complex cl.	216	34.6	259	36.2	265	23.1	312	30.5	1052	30.0
6 No final const.	131	21.0	71	9.9	263	22.9	166	16.2	631	18.0
	624	100.0	716	100.0	1149	100.0	1022	100.0	3511	100.0

Table 3 displays a situation where two classes, phrase and complex clause, are far more common than any of the others. Simple finite and non-finite clauses are very sparsely represented in final position, and so are one-word constituents. The fairly high number of clauses with no final constituent after the verb is also interesting. Naturally, if we work with the idea of topicalization, then moving an element to initial position may leave behind it an empty slot at the end of the clause. One could argue, however, that the movement may have been triggered by a need to rhematize the verb. This kind

of motivation was not taken into consideration in the present coding scheme.

The structure of the fronted and final elements and their length in number of words are naturally correlated: complex clauses are not only more complicated in structure but also longer. A phrase in this classification is a noun phrase with more than one word, but there is no upper limit to its non-clausal pre- and post-modification. Clausal modification comes under complex clauses. Crosstabulation of both the structure of fronted and final elements and the length of fronted and final elements shows that there is highly significant dependence between the two positions in terms of the variables of structure and length. The test used was the chi-square test, which is applicable to samples with nominal scaling (see e.g. Siegel 1956: 104-10). From several possible contingency tables I shall here present only one: the dependence between the structures in the fronted and final elements.

If Table 4 is interpreted in linguistic terms, we can say that there is a tendency for complex final items to appear in clauses in which the fronted element is simple in structure, one word or a phrase. This does not exclude the possibility of a very complicated sentence with complex items at both ends; there are 90 such cases in the corpus. It was pointed out above (p. 53) that final constituents are often phrasal in structure. Phrases are also common in initial position, which gives the cell (2,2 = phrase at both ends) the second highest loading in the whole material. What is surprising, though, is the small number of non-finite clauses in final position. It is apparently so that post-verbal non-finite clauses typically function as postmodifiers of noun phrases and have been classified according to their head nouns.

The results of the statistical analysis show that the majority of topicalized items are short and simple in structure: over 60% of them are one- or two-word constructions in length, and over 70% have a structure which is either a word or a phrase. The principle of end-weight seems to have been followed in spite of the movement of an element to front position, but one should remember, however, that 18% of topicalized clauses are without a post-verbal final element.

Table 4 Dependence between structures in fronted and final elements.

Variable 9 (str. of fronted el.)	18 (str. of final element) 1	2	3	4	5	6	Row total
1 One word	131	531	46	67	351	280	1406
	9.3	37.8	3.3	4.8	25.0	19.9	
	47.8	42.1	32.2	44.7	33.4	44.4	
2 Phrase	60	457	33	52	396	131	1129
	5.3	40.5	2.9	4.6	35.1	11.6	
	21.9	36.2	23.1	34.7	37.6	20.8	
3 Simple fin. clause	53	114	34	16	154	151	522
	10.2	21.8	6.5	3.1	29.5	28.9	
	19.3	9.0	23.8	10.7	14.6	23.9	
4 Simple non-fin. clause	7	76	5	6	61	5	160
	4.4	47.5	3.1	3.8	38.1	3.1	
	2.6	6.0	3.5	4.0	5.8	0.8	
5 Complex cl.	23	83	25	9	90	64	294
	7.8	28.2	8.5	3.1	30.6	21.8	
	8.4	6.6	17.5	6.0	8.6	10.1	
Column total	274	1261	143	150	1052	631	3511

The coding keys in variable 18 are the same as in variable 9 (6 = no final constituent). In each cell the figure in the first line gives the absolute frequencies, the second line gives the row percentages, and the third the column percentages.

Chi square = 209.2, df = 20, significance $p < 0.001$.

3.3 Givenness of the fronted and final elements

The front position in the sentence reflects the kind of information conveyed by it. According to de Beaugrande, 'for efficient communication, it is sensible to present material already established before making additions and modifications. It follows that the early portion of a sentence would be preferentially used for mapping what is previously known. ... By the same token, new or focused knowledge would be strategically well positioned in the predicate. For special focus, marked sentence structures can be employed' (1980:122). This principle was systematized in the Functional Sentence Perspective by Prague School linguists, and it is now generally accepted. Recent discussion has, however, pointed out that it might be better to use

a dichotomy *textually bound* or *unbound* instead of or in addition to the *given - new* distinction, as it is possible for the initial element to refer to new referents but still be textually bound by anaphoric pronouns, connective adverbs, etc. (see e.g. Karlsson 1978: 295-6). Karlsson shows also that, for example, in Finnish the object may take the initial position without being specially focused. Therefore, he makes a distinction between *focalization* (= fronting for special reasons, such as emphasis or contrast) and non-emphatic, thematically determined preverbal word order (1978:299).

Most text linguistic definitions of the problematic term *topic* discussed in the Introduction somehow connected the concept with the type of information conveyed by this element. *Topic* is 'common knowledge, old or given information, what the sentence is about'. However, the element moved by a topicalization operation need not be the topic of the sentence in the above sense. On the contrary, if we consider the triggering effect behind the fronting, it may well be the case that the moved element refers to new information and has been fronted to give it more emphasis.

When discussing the informativity of texts, de Beaugrande and Dressler distinguish between three orders of informativity. First-order occurrences are trivial both syntactically and semantically, such as function words, but also content words in their ordinary meanings, although the latter are generally more informative. When the writer or speaker chooses an unusual syntactic order or semantically less probable items, the predictability decreases, and we get second-order informativity. According to de Beaugrande and Dressler, 'the presence of at least some second-order occurrences would be the normal standard for textual communication, since texts purely on the first order would be difficult to construct and extremely uninteresting' (1981:141-4). Third-order informativity arises from discontinuities and discrepancies, which are infrequent and demand much attention and processing.

As far as the information value of topicalizations is concerned, some of the cases are clearly instances of second-order informativity, sometimes even third-order, but there are also many first-order cases, such as relative words and many initial adverbials. It is a pity I did not yet know of de Beaugrande's and Dressler's ideas at

the time the coding scheme was planned, as it might have made it easier to assess the information value of the fronted elements. The present coding system contains variables (13), (14), and (17), relevant for the assessment of information. Variables (13) and (17) evaluate the givenness of the fronted and final elements in terms of seven categories. Category (1) 'mentioned' denotes coreferential items, while (2) 'same meaning' refers to co-sense, but not co-reference, between the item in question and another item earlier in the text. Categories (1) - (5) all imply old or common information in the sense that the items are either specifically mentioned in the context or otherwise generally known. In the case of long topicalizations it was not always easy to assess the information value of the whole item. A clausal object, for example, may include anaphoric reference to earlier parts of the text, but at the same time mention entirely new things about them. In such cases the whole clause has been classified as new information.

I shall first present separately the givenness of the fronted and final elements according to genre, and then discuss the possible contingencies between the two positions in this respect.

Table 5 Givenness of the fronted element according to genre.

Givenness of fronted el. / Genre	A		J		N		G		Total	
	N	%	N	%	N	%	N	%	N	%
1 Mentioned	151	24.2	217	30.3	260	22.6	313	30.6	941	26.8
2 Same meaning	48	7.7	3	0.4	3	0.3	4	0.4	58	1.7
3 Generic	1	0.2	-		1	0.1	-		2	0.1
4 Implied	54	8.7	22	3.1	104	9.1	47	4.6	227	6.5
5 Pragmatically known	87	13.9	166	23.2	317	27.6	227	22.2	797	22.7
6 New	236	37.8	125	17.5	340	29.6	272	26.6	973	27.7
7 Irrelevant	47	7.5	183	25.6	124	10.8	159	15.6	513	14.6
	624	100.0	716	100.0	1149	100.0	1022	100.0	3511	100.0

Table 6 Givenness of the final element according to genre.

Givenness of final el.	Genre A		J		N		G		Total	
	N	%	N	%	N	%	N	%	N	%
1 Mentioned	17	2.7	54	7.5	166	14.4	93	9.1	330	9.4
2 Same meaning	3	0.5	-		-		-		3	0.1
3 Generic	-		-		1	0.1	-		1	0.0
4 Implied	30	4.8	17	2.4	80	7.0	28	2.7	155	4.4
5 Pragmatically known	24	3.8	41	5.7	90	7.8	33	3.2	188	5.4
6 New	410	65.7	528	73.7	512	44.6	688	67.3	2138	60.9
7 Irrelevant	9	1.4	5	0.7	37	3.2	14	1.4	65	1.9
8 No final constituent	131	21.0	71	9.9	263	22.9	166	16.2	631	18.0
	624	100.0	716	100.0	1149	100.0	1022	100.0	3511	100.0

It is easier to start the discussion with the final element, because the distribution of givenness there seems to follow the basic thematic structure in that the overwhelming majority of final elements conveys new information (82% of cases with a final constituent).

In the fronted element the situation is different. If the fronted element were the topic in the text linguistic sense, i.e. a definite element that conveys old information, one would expect categories (1) - (5) to cover almost all occurrences of the topic. If, on the other hand, the fronted element has been topicalized in the generative sense, i.e. as 'a focus of contrast' in Chafe's words, one might easily expect new information in that slot. It seems to me that the diffuse distribution of cases between old, new, and irrelevant in the givenness variable supports the view that only some of the O-, Adv-, and C-initial clauses exhibit 'real' topicalizations in the generative sense. First of all, one has to exclude cases with relative and question words which reflect the requirements of basic grammar. Also, for most disjuncts and conjuncts the initial position is the unmarked one. Besides, their information content is in most cases irrelevant from the point of view of the actual message, and they have been classified accordingly into category (7).

The heterogeneous nature of information in the initial element is reflected also in the crosstabulation of the fronted and final elements in this respect. New information in the final constituent co-occurs with both old and new information in the fronted element: when there is no final constituent in the clause, the fronted element refers to new information as often as it does to old. The chi-square computed for the contingency table is highly significant, but it is not totally reliable, as the number of cells with low expected frequencies is too large (see Siegel 1956:109).

When the fronted item conveyed old information by the relation of coreference or co-sense (categories (1) and (2) in variable (13)), I was also interested in the type of link that had been used. The most common were various anaphoric pronouns and adverbs (over 70% of cases), while repetition and semantic (= lexical) links each covered approx. 13% of cases. One should notice again that relative words increase the amount of anaphora. Also, one should bear in mind that long topicalizations may be textually bound by devices like anaphora, but have been classified as new if the central point of the message is new in the context.

4 PROBLEM OF INITIAL ADVERBIALS

The frequencies of topicalized items show that adverbial is the sentence element which is most frequently moved to the initial position (Table 1 on p.51). This may naturally be due to the high number of adverbials in English sentences as a whole. According to Ellegård, there are approx. 70 non-clausal adverbials per 100 clauses in English, while the corresponding figures for objects and complements are 44 and 20, respectively (1978:42f.). But there is also general agreement among grammarians that the adverbial is a fairly mobile element in English sentence structure, at least some types of adverbials are. Some adverbials, notably those called disjuncts and conjuncts by Quirk *et al.* (1972), are not even integrated in the sentence structure, which can be seen in their behaviour under certain syntactic operations. It is quite probable that for some adverbials the initial position is the unmarked one, or at least a non-emphatic alternative, and we cannot really postulate that they have been moved from a postverbal position.

Greenbaum (1969) offers some information on adverbs which function as disjuncts or conjuncts in English. He comes to the conclusion that some conjuncts are altogether immobile: their only acceptable place is the initial position. The favoured position for all conjuncts is in front of the clause (almost 75% appear there). If only mobile conjuncts are taken into consideration, preference for the initial position is still 2:1. As far as disjuncts are concerned, the same tendency seems to operate, but the distinction is not as clear as it is in conjuncts. Considering the fairly loose connection that conjuncts and disjuncts have with the rest of the sentence, Greenbaum suggests that they should be generated from a deep structure where the adverbial is directly dominated by the S node. Style disjuncts, in particular, can be explained by postulating a performative clause in which the disjunct functions as an adjunct (1969:82f., 232). Similar solutions have been proposed, for example, by Schreiber (1972) and Jacobson (1978).

In the present study topicalized adverbials were first classified according to their function into four categories: obligatory adjunct, free adjunct, disjunct, and conjunct. Obligatory adjuncts are what Enkvist calls valency adverbials, i.e. adverbials that can be treated as part of the semantic specification of the verb. Their bond to the verb is a tight one, and they are harder to topicalize. Free adjuncts correspond to Enkvist's adverbials of setting, denoting background information not essential to the action itself. Free adjuncts are more mobile in English sentence structure, and can also be topicalized more freely (Enkvist 1976:54-6).

The frequencies of adverbial topicalizations presented in Table 1 (p.51) support Enkvist's views on the tightness of the bond in valency adverbials. There are only ten initial obligatory adjuncts in the whole material, as opposed to 1,920 topicalized free adjuncts. Intertextual differences in free adjuncts are rather small. Differences between genres are more noticeable in disjuncts and conjuncts; these adverbials are clearly style markers in expository and argumentative prose (see Greenbaum 1969:80, 194).

As mentioned above (p.49), Ellegård's classification of adverbials was unsatisfactory for my purposes, and that is why the topicalized adverbials were recoded semantically. The new categories follow ...

basically Quirk and Greenbaum (1973), and the results are presented in Table 7.

Table 7 Semantic classes of adverbials according to genre.

Semantic class	Genre	A		J		N		G		Total	
		N	%	N	%	N	%	N	%	N	%
1 Viewpoint		5	0.8	8	1.1	1	0.1	6	0.6	20	0.6
2 Focusing		12	1.9	45	6.3	28	2.4	32	3.1	117	3.3
3 Intensifying		2	0.3	2	0.3	10	0.9	4	0.4	18	0.5
4 Process		40	6.4	70	9.8	134	11.7	114	11.2	358	10.2
5 Subject		24	3.8	6	0.8	22	1.9	25	2.4	77	2.2
6 Place		69	11.1	180	25.1	110	9.6	201	19.7	560	15.9
7 Time		156	25.0	108	15.1	311	27.1	203	19.9	778	22.2
8 Style		-		31	4.3	-		20	2.0	51	1.5
9 Attitudinal		14	2.2	22	3.1	36	3.1	38	3.7	110	3.1
A Additive		12	1.9	38	5.3	10	0.9	13	1.3	73	2.1
B Transition		5	0.8	22	3.1	18	1.6	23	2.3	68	1.9
C Result + reason		13	2.1	47	6.6	5	0.4	32	3.1	97	2.8
D Contrast		9	1.4	11	1.5	10	0.9	20	2.0	50	1.4
E Concessive		12	1.9	42	5.9	15	1.3	48	4.7	117	3.3
F Condition		28	4.5	38	5.3	43	3.7	38	3.7	147	4.2
G No adverbial in the clause		223	35.7	46	6.4	396	34.5	205	20.1	870	24.8
		624	100.0	716	100.0	1149	100.0	1022	100.0	3511	100.0

Functionally, categories (1) - (7) are adjuncts, categories (8) and (9) disjuncts, and categories (A) - (F) conjuncts. Category (G) simply gives the number of non-adverbial topicalizations in the corpus.[2]

The typical position of adverbial classes varies to some extent. I have already pointed out (p. 59) that disjuncts and conjuncts favour the initial position, and so do some of the adjuncts. According to Quirk *et al.* (1972), viewpoint adjuncts are usually in initial position, and also subject adjuncts often occur initially. The first

group can be paraphrased by 'if we consider what we are saying from X point of view', while the second group relates the process of the verb to the subject, e.g. *Bitterly, he buried his children* (Quirk and Greenbaum 1973:210ff.). When these adverbials occur in initial position, one can hardly argue that they have been moved there from a postverbal position.

If we want to find 'real' topicalizations, we must look for them in the more mobile classes of adverbials, such as process, place, and time adverbials. Process adjuncts, by the way, include the traditional adverbials of manner, means, and instrument. These classes are clearly the largest in the corpus, and together they cover almost 65% of topicalized adverbials. In order to be able to determine whether adverbial topicalizations are rare or not, we ought to know something about the overall frequencies of adverbials in English. Unfortunately, there is very little material for comparison; Greenbaum (1969) discusses only adverbs functioning as disjuncts or conjuncts, and Ellegård's class E (= other adverbials) is too heterogeneous for comparison, as it includes adjuncts in addition to disjuncts and conjuncts. Jacobson's extensive work on adverbials does not relate their occurrences to other sentence elements or clause structure (Jacobson 1964, 1975, and 1978).

One could start with the total number of adverbials in Ellegård, which is 83 adverbials per 100 clauses: 14.1 of these are clausal (13.9 major adverbials and 0.2 manner adverbials), while the majority, 68.9, are phrasal. Of the latter 56 belong to major adverbials, 6.5 are manner adverbials, and 6.4 are sentence adverbials. The proportion 83:100 means that the present material (approx. 13,000 finite clauses) contains almost 10,800 adverbials. The number of topicalized adverbials was 2,641, which would mean that almost every fourth adverbial has been fronted. This, incidentally, is much more than in other sentence elements, thus confirming the well-known mobility of adverbials in English. The question still remains whether the initial position represents a marked word order or not.

I have argued earlier (p. 58) that if the topicalized elements had been fronted because of emphasis or contrast, one would expect them to refer to new information in discourse. The figures presented in Table 5 (p.57) on the givenness of the fronted element show that

in approx. 58% of cases the fronted element conveys old information, in 28% new information. There are, however, differences between sentence elements in this respect. The proportion of *old* information is the same in direct objects and adverbials, but topicalized objects contain *new* information more often than adverbials (37.4% vs. 24.5%). The main difference lies in the irrelevant cases. While there are hardly any topicalized objects whose givenness has been classified as irrelevant, that category covers 18.5% of fronted adverbials. This is mainly due to the role disjuncts and conjuncts play in texts. They are connective elements, which add to the cohesion of the text. Quite often, and this is especially true of disjuncts, they function on a metatextual level, explaining the organization of the textual structure. When one studies adverbial fronting, one must assume that initial position is the unmarked one for disjuncts and conjuncts, i.e. they have not been topicalized.

This leaves us with the adjuncts. The proportion of old information is now slightly higher; almost 69% of fronted adjuncts convey given information. The result suggests, at least to me, that the majority of adjuncts have been fronted for other than genuine topicalizing reasons, i.e. other than emphasis, contrast, etc. In the case of relative adverbs, like *when*, *where*, *how*, and *why*, the initial position is naturally the only grammatical one, but that explains only 10% of the fronted adjuncts conveying old information. The rest, almost 1,200 cases, may naturally be 'real' topicalizations despite their old information content, but there are at least two other possibilities: (1) these adjuncts have been moved for thematic, i.e. textual, reasons, or, (2) the initial position is simply one of the unmarked positions for some adjuncts, and we cannot postulate a basic SV(O)A order for all adjuncts.

We get some support for the first of the 'quasi-topicalization' alternatives by crosstabulating the givenness variable by the semantic class of the adverbial. The highest loadings appear in cells (1,6; mentioned place), (5,6; implied place), and (5,7; implied time). This result indicates that it is quite common to start the sentence with an adverbial of setting which ties the following information with what has gone before. Reference to the same time or place is a convenient way of performing the link. If we assume that the unmarked position of these adjuncts is after the verb, then we must say that

they have been fronted for textual reasons.

There are, however, fronted adjuncts which do not denote given information, and thus their movement cannot be attributed to thematic forces. They refer to new information, but their context does not give support to any emphatic or contrastive interpretations. If they are read aloud, they do not obtain a contrastive focus, which would be a sign of a marked pattern. One can only conclude that the initial position is a neutral one at least for some adjuncts. The contingency table mentioned above (givenness by semantic class) reveals a high loading in cells (6,4; new process) and (6,7; new time), which points to the frequency of sentences beginning with time adjuncts like *today, yesterday, last week, next year*, etc. Indeed, it seems to me that for time adjuncts in particular the initial position is an unmarked one, and we must adopt also the second alternative presented above. The high number of process adjuncts, on the other hand, can partly be explained by the presence of interrogative adverbs like *how* and *why* in indirect questions.

5 MOTIVATION FOR TOPICALIZATION

As pointed out above, it is generally assumed that the basic word order in English is SVO(A), although there are several suggestions for the deep structure position of adverbials. When the constituents appear in different positions in surface structure, they have undergone movement transformations. The starting point in the present study was the surface structure, and the topicalized material includes all clauses with an O-, A-, or C-initial word order. In order to find out the proportion of what I have called 'real' topicalizations, I tried to estimate the probable motivation behind the fronting. Sometimes the task is easy: if the resulting word order is the only grammatical one, no special trigger is needed. In many cases, though, one has to look for clues in a wide context, and sometimes the positioning could be attributed to several forces working simultaneously. Table 8 (next page) presents the distribution in variable (20).

Some of the categories require explanation. Category (1, compensatory) was adopted from Hakulinen, Karlsson, Vilkuna (1980), who use it to refer to cases where an item is moved to the front to fill an otherwise empty slot before the verb. This device is often used in

Finnish existential sentences, and in English it seems to work in similar cases, but often causes a VS inversion in the clause. I have also used the feature to denote cases where the motivating force may be a need to alleviate an unduly heavy end for a better balance in the sentence.

Table 8 Probable motivation for topicalization according to genre.

Triggering effect / Genre	A		J		N		G		Total	
	N	%	N	%	N	%	N	%	N	%
1 Compensatory	64	10.3	134	18.7	86	7.5	135	13.2	419	11.9
2 Grammatical	146	23.4	119	16.6	275	23.9	267	26.1	807	23.0
3 Emphasis	145	23.2	160	22.3	297	25.8	263	25.7	865	24.6
4 Contrast	48	7.7	18	2.5	15	1.3	19	1.9	100	2.8
5 Unmarked	183	29.3	257	35.9	444	38.6	307	30.0	1191	33.9
6 Combined effect of forces	25	4.0	17	2.4	20	1.7	13	1.3	75	2.1
7 Parallelism	13	2.1	10	1.4	12	1.0	18	1.8	53	1.5
8 Disambiguation	-		1	0.1	-		-		1	0.0
	624	100.0	716	100.0	1149	100.0	1022	100.0	3511	100.0

Categories (3, emphasis) and (4, contrast) are the factors which are generally mentioned in connection with topicalization. Category (5, unmarked) covers all cases where no special reasons could be detected for the fronting. By parallelism (category 7) I mean similarity in syntactic structure. This effect often produces several topicalizations, one after another, only the first of which is triggered by emphasis or contrast. Category (8, disambiguation) was included in the coding scheme for the purpose of finding out if the possibility of misunderstanding affects the placement of elements in a sentence. Adverbials would be the likely candidates in this category as their scope can be varied according to their position, but as the results show, only one case was found. This need not mean that disambiguation plays no role in fronting, but rather that it is difficult to distinguish from the other forces.

When we look at the results presented in Table 8, we will see that emphasis and contrast cover less than 30% of all cases. Actually, grammatically obligatory and unmarked cases together form a majority in the corpus (57%), in spite of the fact that I tried to use bias in favour of 'real' topicalizations. In other words, emphasis and contrast were classified as triggering effects whenever the context gave *any* support for such interpretation. If the decision-making had been based on other factors, e.g. textual boundness, the number of emphatic and contrastive cases would have been even smaller.

It seems to me that the motivating factors in variable (20) can be divided into structural and semantic groups. Structural motivation can work on both clausal and textual levels. Clausal level motivation simply covers those cases where basic syntax does not allow any other word order (= category (2)). This includes relative clauses, interrogative clauses, exclamations, and some clauses with initial negative or restrictive expressions. But structural motivation can work on longer passages than a clause or a sentence. Text linguists have shown that the distribution of given and new information, textual boundness and unboundness, constituent length, etc., are factors which affect the ordering of constituents in sentences. In the coding scheme categories (1, compensatory) and (5, unmarked) reflect structural motivation on textual level. A *syntactically marked* word order pattern may be *textually unmarked* and the best solution in a particular context. Enkvist (1976:65) has expressed similar views concerning adverbial placement in English.

Emphasis and contrast differ from the structural factors in that they are semantically motivated. What we usually mean by emphasis and contrast is that some new aspects enter the discourse. It may be a new referent not mentioned before, or it may be new information about a known referent, something that changes or upsets our previous presuppositions. Emphasis and contrast reveal the speaker's or writer's subjective attitudes towards the content of the message. This explains also why it is sometimes difficult for people to agree on emphatic or contrastive items, as the interpretation depends on our view of the world, previous knowledge, etc. However, I think that we *can* agree on the semantic nature of emphasis, contrast, and other types of strong motivation. In these cases the marked surface order is a reflection of the semantic markedness.

The structural and semantic levels were separated in the above treatment, but in actual communication they work together. Beaugrande points out that an element may be syntactically probable and semantically improbable, or vice versa. He says, 'Probable content in a probable format would be uniformly easy to process and not informative. Improbable content in an improbable format would be uniformly difficult to process and intensely problematic. But improbable content in a probable format, or probable content in an improbable format would be challenging and yet not unreasonably problematic' (Beaugrande 1980:104-5). The topicalizations in the present corpus come from all these contingencies, and therefore their contribution to the information content and the style of the texts varies enormously.

6 CONCLUDING REMARKS

The study reported in this paper was an attempt to use a computer corpus for quantitative syntactic purposes. The corpus used was the Gothenburg version of the Brown Corpus, consisting of 128,000 running words from four different genres. The topicalized sub-corpus was extracted from the big corpus by a search programme which picked out all object-, adverbial-, and complement-initial clauses. The total yield was 3,511 cases with a topicalized word order. The study concentrates on the question whether all these cases are 'real' topicalizations, which are usually supposed to be triggered by emphasis, contrast, and other 'special' reasons. The results show that the majority of so-called topicalizations can be attributed to other factors. A large number of cases represent the only possible grammatical word order, and their surface order is generated by Relative Clause and Question Transformations. But in addition to the grammatical cases, which could actually have been excluded from the corpus, there remain over 1,600 cases (46% of the corpus) in which one cannot find any special reasons for the fronting. I have argued above that most of these cases are textually unmarked, though grammatically marked. The syntactic pattern has conceded to the thematic requirements of a proper text. This tendency is perceivable in the statistical information concerning the given-new dichotomy, the length and structure of the fronted and final constituents, and the type of reference used in the fronted element. The thematic

forces work together with some other well-known forces, such as the principles of end-focus and end-weight, and it is the total effect of these forces in a context that has to be taken into account when various surface order patterns are assessed.

The ideas presented above are by no means new. Similar views have been expressed by several text linguists, but usually without substantial evidence from concrete corpora. If a corpus has been used, it may have revealed stylistic idiosyncracies rather than a generalizable norm (see e.g. Enkvist and von Wright 1978:54-70).

Recently work on larger corpora has become easier, as several of them either have been coded or are being coded grammatically. After a quarter of a century of introspective linguistics, there is room for corpus-based studies, particularly now that syntactic variation has become a favoured topic. As Ulvestad points out, sometimes linguists are interested in establishing grammaticality, sometimes they want to describe 'within-class variation in the domain of grammatical constructions'. For the first type of study, the concepts of frequency and probability are of no concern, while the second type cannot be pursued without knowledge of frequencies in large enough corpora (Ulvestad 1979:89-90; see also Johansson 1979:292). The present study is an attempt of the second type. However, one must beware of the danger of simplification in interpreting any statistical results. Discrete figures in nicely-laid tables easily hide the linguistic complexity of a text, where everything depends on everything else.

NOTES

1 I am grateful to Mr Seppo Rantala, M. Sc., from the Computer Centre of the University of Turku, for the compilation of the computer programme.

2 There is a difference of one case in the number of conjuncts between Tables 1 and 7. This must be due to a punching error which has not been detected, in spite of several check runs.

REFERENCES

Andersson, E., ed. 1978. *Working Papers on Computer Processing of Syntactic Data.* Publications of the Research Institute of the Åbo Akademi Foundation, 37. Åbo.

Beaugrande, R. de. 1980. *Text, Discourse, and Process: Toward a Multidisciplinary Science of Texts.* London.

Beaugrande, R. de and W. Dressler. 1981. *Introduction to Text Linguistics.* London.

Bergenholtz, H. and B. Schaeder, eds. 1979. *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora.* Königstein.

Chafe, W.L. 1975. 'Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View'. In C.N. Li, ed. 25-55.

Dahl, Ö., ed. 1974. *Topic and Comment, Contextual Boundness, and Focus.* Hamburg.

Dressler, W., ed. 1978. *Current Trends in Text Linguistics.* Berlin, New York.

Ellegård, A. 1978. *The Syntactic Structure of English Texts: A Computer-Based Study of Four Kinds of Text in the Brown University Corpus.* Gothenburg Studies in English, 43. Gothenburg.

Enkvist, N.E. 1976. 'Notes on Valency, Semantic Scope, and Thematic Perspective as Parameters of Adverbial Placement in English'. In N.E. Enkvist and V. Kohonen, eds. 51-74.

Enkvist, N.E. 1978. 'Stylistics and Text Linguistics'. In W. Dressler, ed. 174-190.

Enkvist, N.E. and V. Kohonen, eds. 1976. *Reports on Text Linguistics: Approaches to Word Order.* Publications of the Research Institute of the Åbo Akademi Foundation, 8. Åbo.

Enkvist, N.E. and M. von Wright. 1978. 'Problems in the Study of Textual Factors in Topicalization'. In E. Andersson, ed. 45-71.

Greenbaum, S. 1969. *Studies in Adverbial Usage.* London.

Gustafsson, M. (forthcoming). 'Are There Any Topicalized Objects?'. To be published in *Studies Presented to Y.M. Biese on the Occasion of His Eightieth Birthday.* Helsinki.

Hakulinen, A., Karlsson, F. and M. Vilkuna. 1980. *Suomen tekstilauseiden piirteitä: Kvantitatiivinen tutkimus.* Publications of the Department of General Linguistics, University of Helsinki, 6. Helsinki.

ICAME News, Newsletter of the International Computer Archive of Modern English, University of Bergen, Bergen. Nos. 2 and 3.

Jacobson, S. 1964. *Adverbial Positions in English.* Uppsala.

Jacobson, S. 1975. *Factors Influencing the Placement of English Adverbs in Relation to Auxiliaries.* Stockholm Studies in English, 33. Stockholm.

Jacobson, S. 1978. *On the Use, Meaning, and Syntax of English Preverbal Adverbs.* Stockholm Studies in English, 44. Stockholm.

Jespersen, O. 1961. *A Modern English Grammar on Historical Principles,* Part VII. Repr. London and Copenhagen. (First published in 1949)

Johansson, S. 1979. 'The Use of a Corpus in Register Analysis: The Case of Learned and Scientific English'. In H. Bergenholtz and B. Schaeder, eds. 281-93.

Karlsson, F. 1978. 'Nominaalilausekkeiden tematiikkaa ja eksistentiaalilauseiden ongelma'. In *Rakenteita*. 293-305.

Kohonen, V. 1978. *On the Development of English Word Order in Religious Prose Around 1000 and 1200 A.D.: A Quantitative Study of Word Order in Context*. Publications of the Research Institute of the Åbo Akademi Foundation, 38. Åbo.

Kučera, H. and W.N. Francis. 1967. *Computational Analysis of Present-Day American English*. Providence, Rhode Island.

Li, C.N., ed. 1975. *Subject and Topic*. New York.

Quirk, R. 1978. 'Focus, Scope, and Lyrical Beginnings'. *Language and Style*, 11. 30-39.

Quirk, R., Greenbaum, S., Leech, G. and J. Svartvik. 1972. *A Grammar of Contemporary English*. London.

Quirk, R. and S. Greenbaum. 1973. *A University Grammar of English*. London.

Rakenteita. Juhlakirja Osmo Ikolan 60-vuotispäiväksi 8.2.1978. Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja, 6. Turku.

Schreiber, P.A. 1972. 'Style Disjuncts and the Performative Analysis.' *Linguistic Inquiry*, 3. 321-47.

Siegel, S. 1956. *Nonparametric Statistics for Behavioural Sciences*. Tokyo.

Ulvestad, B. 1979. 'Corpus vs. Intuition in Syntactical Research.' In H. Bergenholtz and B. Schaeder, eds. 89-108.

APPENDIX: Coding scheme

Column	Variable no.	Parameter description and classification
1-4	1	Item number
5-8	2	Line number in the corpus print
9	3	Genre A = journalism G = literary essays J = science N = popular fiction
10-11	4	Text number in the corpus
12	5	Clause type Z = main clause P = subordinate clause (other than relative) R = relative clause
13	6	Mood of the clause D = declarative clause Q = interrogative clause J = imperative clause E = exclamatory clause
14	7	Function of the moved constituent 1 = direct object 2 = indirect object 3 = prepositional object 4 = obligatory adjunct 5 = free adjunct 6 = disjunct 7 = conjunct 8 = subject complement 9 = object complement
15	8	Position and scope of the movement 1 = moved alone to the beginning of a sentence 2 = moved together with another constituent to the beginning of a sentence 3 = moved alone to the beginning of a clause 4 = moved together with another constituent to the beginning of a clause
16	9	Structure of the moved constituent 1 = one word 2 = phrase 3 = simple finite clause 4 = simple non-finite clause 5 = complex clause (contains embedding)
17	10	Order of main constituents 1 = SV 2 = VS 3 = ellipted S

Column	Variable no.	Parameter description and classification
18	11	Length of the moved constituent 1 = one word 2 = 2-3 words 3 = 4-6 words 4 = 7-10 words 5 = 11- words
19	12	Length of the final constituent in the same clause as the fronting 1 = one word 2 = 2-3 words 3 = 4-6 words 4 = 7-10 words 5 = 11- words
20	13	Givenness of the moved constituent 1 = mentioned (coreferential items) 2 = same meaning (not coreferential) 3 = generic 4 = implied 5 = pragmatically known 6 = new 7 = irrelevant
21	14	Cohesive device in moved constituents mentioned earlier (1 + 2 in variable (13)) 1 = ellipsis 2 = anaphoric pronoun 3 = cataphoric pronoun 4 = repetition 5 = semantic cohesion 6 = grammatical link 7 = exophoric reference 8 = other device 9 = no link
22	15	Semantic class of a moved adverbial 1 = viewpoint 2 = focusing 3 = intensifying — 1 - 7 = adjuncts 4 = process 5 = subject 6 = place 7 = time 8 = style — 8 - 9 = disjuncts 9 = attitudinal A = additive B = transition C = result + reason — A - E = conjuncts D = contrast E = concessive F = condition

Column	Variable no.	Parameter description and classification
23	16	Type of verb 1 = stative verb 2 = dynamic verb 3 = verb of saying and thinking
24	17	Givenness of the final constituent 1 = mentioned 2 = same meaning 3 = generic 4 = implied 5 = pragmatically known 6 = new 7 = irrelevant
25	18	Structure of the final constituent 1 = one word 2 = phrase 3 = simple finite clause 4 = simple non-finite clause 5 = complex clause
26	19	Function of the final constituent 1 = object 2 = obligatory adjunct 3 = free adjunct 4 = disjunct or conjunct 5 = complement 6 = other (e.g. subject)
27	20	Probable motivation for fronting 1 = compensatory 2 = grammatical 3 = emphasis 4 = contrast 5 = unmarked 6 = combined effect of several forces 7 = parallelism 8 = disambiguation
28	21	Function of grammatically moved constituents (2 in variable (20)) 1 = relative pronoun 2 = question word or phrase 3 = exclamation 4 = negative or restrictive constituent 5 = missing relative pronoun 6 = lifted constituent (see Ellegård) 7 = conjunction

CURRENT WORK ON ENGLISH COMPUTER CORPORA

The two articles in the present issue of *ICAME News* are based on the Syntax Data Corpus prepared by Alvar Ellegård (see *ICAME News* 2). They illustrate the possibilities offered by using a grammatically tagged corpus. Current work on English computer corpora, in particular grammatical tagging, is reported in Stig Johansson, ed., *Computer Corpora in English Language Research* (Bergen: Norwegian Computing Centre for the Humanities, 1982). The book, which also includes some more general articles and bibliographies for the Brown Corpus, the LOB Corpus, and the London-Lund Corpus, should give a good picture of the 'state of the art' (see further the enclosed order form). Other recent or forthcoming publications are:

Francis, W. Nelson and Henry Kučera. *Frequency Analysis of English Usage: Vocabulary and Grammar*. Houghton Mifflin Co. (based on the tagged version of the Brown Corpus prepared at Brown University)

Hofland, Knut and Stig Johansson. 1982. *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities. (mainly based on the LOB Corpus; includes a comparison of word frequencies in the LOB Corpus and the Brown Corpus; see the enclosed order form)

Svartvik, Jan, Eeg-Olofsson, Mats, Forsheden, Oscar, Oreström, Bengt and Cecilia Thavenius. 1982. *Survey of Spoken English: Report on Research 1975-81*. Lund Studies in English 63. Lund: C.W.K. Gleerup. (report on the computerization of the London-Lund Corpus and on research based on this material)

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.

... 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*, 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.

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

CONDITIONS ON THE USE OF ICAME CORPUS MATERIAL

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

(a) collecting and distributing information on (i) English language material available for computer processing; and (ii) linguistic research completed or in progress on this material;

(b) compiling an archive of corpora to be located at the University of Bergen, from where copies of the material can be obtained at cost.

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

1 No copies of corpora, or parts of corpora, are to be distributed under any circumstances without the written permission of ICAME.

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

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

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

EDITORIAL NOTE

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

ICAME NEWS is published by the Norwegian Computing Centre
for the Humanities (NAVF.s EDB-senter for humanistisk forskning)
Address: Harald Hårfagresgate 31, P. O. 53, 5014 Bergen - University, Norway