Clause patterns in Modern British English: A corpus-based (quantitative) study

Nelleke Oostdijk and Pieter de Haan
University of Nijmegen

Abstract: In the Department of Language and Speech (Corpus Linguistics Section) of Nijmegen University a research project is being carried out which aims to provide a survey of the frequency of occurrence and the distribution of a range of syntactic structures in Modern British English. The project makes use of the Nijmegen Corpus. This computerized corpus, comprising approximately 130,000 words, has undergone a detailed syntactic analysis and is available for exploratory studies. This article reports on findings with regard to the clause patterns encountered in the material.

1. Introduction

Large-scale quantitative studies of syntactic structures and phenomena are long overdue. While word frequency counts and concordances have been a common good to the linguistic community for quite some time now, corpora that have undergone a detailed syntactic analysis are few, and so are the quantitative studies that are based on these. In the Department of Language and Speech (Corpus Linguistics Section) of Nijmegen University a research project is being carried out which aims to provide a survey of the frequency of occurrence and the distribution of a range of syntactic structures in Modern British English.[1]

The usefulness of structure frequency counts is obvious: the results can bring new insights to the world of descriptive linguistics,[2] while those concerned with natural language processing will value the information about frequency and distribution when it comes to deriving probabilities for their parsers. Quantitative data that derive from (large-scale) corpus-based studies can serve to strike a balance between intuitive notions of what is common or typical of certain language varieties, and observed actual language use.

Grammars that have emerged from traditional descriptive linguistics, such as the comprehensive grammatical handbooks of Kruisinga (1909-32), Poutsma (1904-26), Jespersen (1909-49), and the more recent handbooks by Quirk *et al*. (1972, 1985), provide a wealth of information as far as the structural description of constructions and their usage is concerned. Information about the frequency of occurrence and the distribution of various linguistic structures is largely non-existent. Where such information is ventured, it is often based on subjective judgments on the part of the author(s) and not on any systematic examination of any amount of textual evidence.[3] While concerned with identifying the 'common core' of the language, aiming to describe what is 'usual' in language use, grammarians tend to pass judgment on the 'desirability' of constructions, the extent to which they are 'acceptable' to speakers, or typical of a certain style, etc.

The corpus-based (quantitative) studies that have been carried out to date have generally focused on specific constructions or phenomena. The main studies are listed in Table 1.1.[4] It must also be observed that these studies involved a great deal of manual labour at the cost of consistency and scale.[5] Collecting data from corpora has so far been done mainly by hand, even if subsequent analyses of these data were carried out with the help of the computer (cf. Van Ek, 1966; Huddleston, 1971; Ellegård, 1978; De Haan, 1989a)

Table 1.1: Some quantitative corpus-based studies (cf. de Haan, 1989a: 50)

	object of study:
Lebrun (1965)	*can* and *may*
Van Ek (1966)	predication structures
Svartvik (1966)	voice
Hough III (1970)	modification
Yotsukura (1970)	articles
Aarts (1971)	NP structures
Huddleston (1971)	various/general description
Scheffer (1975)	progressives
Erdmann (1976)	*there* sentences
Wekker (1976)	future time

Vestergaard (1977)	prepositional phrases & prep. verbs
Ellegård (1978)	syntactic structures
Hermerén (1978)	modals
Olofsson (1981)	relative junctions
Biber (1988)	stylistic variation
De Haan (1989a)	postmodifying clauses in the English NP
Mair (1990)	infinitival complement clauses
Meyer (1992)	apposition

Unlike the studies above, the current project can make use of the data contained in a computerized corpus, the Nijmegen Corpus, which has been automatically analyzed in great detail by means of a formal grammar. The analyses, in the form of tree diagrams, have been stored in a syntactic database system where they are available for further examination. While the project eventually aims to provide a survey of the frequency of occurrence and the distribution of a range of syntactic structures, in the present article the focus is on clause structure.

2. Data processing and data analysis

The Nijmegen Corpus was compiled and computerized at the University of Nijmegen during the early 1970s. In the course of the Dutch Computer Corpus Pilot Project (Keulen, 1986) it was analyzed (semi)automatically and the results were stored in the *Linguistic DataBase* (LDB, cf. van Halteren and van den Heuvel, 1990).

The corpus is relatively small: it comprises approximately 130,000 words of running text.[6] It is rather unique, however, in that it was compiled with the intention of studying language variation at a syntactic level and therefore contains rather largish samples of some 20,000 words each. The samples have been taken from a number of different text categories and authors. In all, the corpus contains approximately 120,000 words that originate from printed sources (the remainder is spoken sports commentary) and is therefore biased towards the written language. The bibliographic references to the source texts of the samples in the corpus are given in Appendix A.

The Nijmegen Corpus was manually tagged by postgraduate students of the English Departments of most Dutch Universities, under the supervision of the English Department at Nijmegen. For the purpose of analyzing the corpus a formal grammar was written that could be transformed automatically into a parser (an analysis program). The grammar was largely based on the descriptive grammar by Quirk *et al.* (1972), although the formalization forced the grammar writers into being more rigid with respect to their descriptive system. In this system a structure is assumed which is based on immediate constituency and which represents the rank scale. Constituents are labelled for their function and category. Thus the labelling holds information both about the syntactic characteristics of a single descriptive unit, and about its role in a larger linguistic structure.

All the analyzed utterances have been stored in the Linguistic DataBase in the form of analysis trees, containing function and category information at every node. The database can be queried by defining so-called search patterns to be matched in the analysis trees. For instance, a search can be conducted on instances of prepositional phrases that function in subject noun phrases. An account of an LDB study of the Nijmegen Corpus can be found in van Halteren and Oostdijk (1988).

It can also be specified what action must be taken when a match is found. In the most elementary applications control will be handed back to the user, enabling him or her to examine the sentence containing the match on the screen. Another possibility is to have the sentence containing the match, or just the match, stored in a new file, which can later be inspected, or printed. Alternatively, the LDB could be instructed to keep count of the number of matches found. In the hypothetical example above we could specify that tables should be generated, telling us how many occurrences of each individual preposition were found in the construction specified, in each of the subcorpora.

There are various statistical methods for determining significant relationships between variable features. For the analysis of a 2 × 2 cross-tabulation, or contingency table, i.e. a table with two dichotomous variables, e.g. *sex* (male female) × *education* (academic non-academic) a simple chi-square test will usually indicate whether the distribution found is statistically significant. For the analysis of an $n \times n$ table, i.e. one with at least one non-dichotomous variable, e.g. *sentence structure* and *sentence status*, a look at the standardized residual scores usually gives a fairly reliable indication of significant relationships. However, a table which involves more than two variable features (so-called *n*-way tables, in which $n > 2$) cannot be analyzed adequately with either of these techniques. For the analysis of *n*-way tables in this study we therefore used a loglinear analysis (cf. de Haan and van Hout, 1986; 1988).

It may be useful to mention the main advantages of this technique over chi-square tests performed on 2 × 2 tables, without going into too many technical details:

- more than two variable features can be accommodated. In this investigation we performed a loglinear analysis on a table with the variables sentence structure, sentence status (matrix or embedded) and sort of sentence (finite, non-finite or elliptical), as well as one on a table with the variables sentence structure, sentence status and number of adverbials;
- more than two values per variable can be handled. In our analyses we distinguished two types of sentence status (matrix and embedded), six sentence structures (see below, Section 3), three sentence sorts (see above) and three classes of numbers of adverbials (no adverbials, one adverbial, more than one adverbial);
- not only the significance of single variables (i.e. effects) can be determined, but also that of the combination of two or more effects;
- ultimately, the loglinear analysis enables the investigator to draw up various models, containing different numbers of effects, in order to arrive at the model which provides the most adequate explanation for the distribution found.

3. Object of study: Frequency of occurrence and distribution of clause patterns

English is commonly described as a ‘fixed word-order’ language. The reason for this is, as Quirk *et al*. (1985: 51) observe, that ‘in English the positions of subject, verb and object are relatively fixed. In declarative clauses, they occur readily in the order S V O, unless there are particular conditions ... which lead to a disturbance of this order.’

The unmarked word-order is the ‘normal’ order in a simple declarative sentence (the canonical form of the sentence) in which constituents such as subject, verb and object occur. Sentence (or: clause) patterns showing unmarked word-order represent the normal flow of information in a sentence or clause, i.e. the subject is the topic, which therefore occurs sentence/clause-initially, and the predicate contains the comment, and therefore follows the subject (the principle of **end-focus**, cf. Leech, 1983). Moreover, in cases of multiple complementation (i.e. OI-OD or OD-CO) the first complement will be shorter than the second (the principle of **end-weight**, cf. Gleason, 1965; Leech, 1983).

Marked word-order typically occurs in cases where the principles of end-focus and end-weight clash. For example, when a subject is very long, or in cases where the topic is not the subject, or where for some reason prominence is given to a constituent other than the subject, by placing it in initial position. The latter can be due to the fact that a contrast is intended (cf. Chafe, 1976).

Li and Thompson (1976) point out that Indo-European languages are **subject-prominent** (as opposed to e.g. Chinese, which is a **topic-prominent** language). This means that if the end-weight principle dictates a 'heavy' subject to be moved to a final position, there occurs a 'dummy' subject in the original subject (i.e. initial) position. It is this same fact which accounts for the occurrence of dummy subjects in sentences like *it is raining*.

Huddleston (1971), following Halliday (1969), refers to the leftmost element, or group of elements, of the sentence (or clause) as the 'theme'. He uses this term to distinguish between sentences with marked theme and unmarked theme. A marked theme is an element, other than a *wh*-item or a conjunction, preceding the subject. He mentions cases of marked object theme (as in: '*This sum* we might call the torque'), marked attribute theme (as in: '*more effective, and certainly more interesting*, however, is a structure...'). Sentence-initial adverbials are called 'marked adjunct themes'.

Discussions of sentence structure in grammatical handbooks (e.g. Quirk *et al*. 1972, 1985; Aarts and Aarts, 1982) do not give any systematic account of the composition of sentences and the alternative orders in which their immediate constituents may occur, nor do they provide any information about the frequency of occurrence and the distribution of clause patterns. The present study was therefore undertaken with the objective of establishing what clause patterns actually occur. More in particular, we wanted to find answers to the following questions:

1. What patterns are most commonly used, and what is their distribution? What is the ratio of sentences and clauses in which we find permutations of the basic clause patterns? What permutations are actually found in 'real' data?
2. Are there any differences between matrix sentences and embedded sentences (= clauses), and if so, what are these?
3. Are there any differences between finite and non-finite sentences and clauses with respect to the clause structures that are found in them?

4. Are embedded sentences that are immediate constituents (ICs) of other clauses any different from those that are ICs of phrases?
5. To what extent are clause patterns extended by means of optional adverbials?

In approaching this subject matter we started from the definition of the basic clause types as found in the grammatical description that was used in the analysis of the corpus material and which largely coincides with the clause type definitions given in Quirk *et al.* (1972).[7] We came to distinguish five clause types on the basis of merely the obligatory functional constituents that each of these contain.[8] They are:

- *intransitive*: SU-V(intransitive): the sentence consists of a subject and an (intransitive) verb; e.g. *Jane laughed.*
- *intensive*: SU-V(intensive)-CS: apart from the subject the sentence consists of an intensive verb and a subject complement; e.g. *He is a buddhist.*
- *monotransitive*: SU-V(monotransitive)-OD: the sentence consists of a subject, a monotransitive verb and a direct object; e.g. *I've found my glasses.*
- *ditransitive*: SU-V(ditransitive)-OI-OD: the sentence consists of a subject, a ditransitive verb, an indirect and a direct object; e.g. *She gave me the keys.*
- *complex transitive*: SU-V-OD-CO: apart from the subject, the sentence consists of a complex transitive verb, a direct object and an object complement; e.g. *The meeting elected Harry chairman.*

Henceforth we shall use the labels 'intransitive', 'intensive', etc., whenever we refer to any clausal structure that answers to the above descriptions, not taking into account any extensions by means of optional adverbials and similar elements, nor paying any attention to whether the constituents occur in unmarked order or whether for some reason their order is marked.[9]

4. Distribution of patterns

The corpus comprises 15,125 sentences. 7434 (49.15%) of these are matrix sentences, while 7691 (50.85%) are embedded sentences.[10] The majority of the matrix sentences are finite: 7271 (97.8%) are finite, 109 (1.5%) are non-finite, and the remaining 54 (0.7%) are elliptical. With embedded sentences the distribution over each of these categories is rather different: here 4454 (57.9%) of the sentences are finite, while 2792 (36.3%) are non-finite and 445 (5.8%) are elliptical.

In Table 4.1 overall figures are given for each of the clause patterns and their absolute frequencies of occurrence. The distribution of the various clause patterns is basically similar for both matrix sentences and embedded sentences. In both types of sentence the intransitive pattern is the most frequent, followed by the monotransitive and the intensive pattern. The three patterns together account for 79.6% and 87.6% of the total number of sentences respectively.

Table 4.1: Clause patterns: Overall figures (n = 15,125)

status	intrans	intens	mono	ditrans	complex	other	total
matrix	2254	1850	1814	90	82	1344	7434
embedded	3029	1320	2389	61	131	761	7691
total	5283	3170	4203	151	213	2105	15,125

However, while the intensive and monotransitive patterns are evenly distributed within matrix sentences, monotransitive patterns are nearly twice as frequent as intensive patterns in embedded sentences. This is shown in Figure 4.1, which displays the relative distribution of the various patterns in the two types of clause. The chi-square test and the standardized residuals of the scores of Table 4.1 show these differences to be statistically significant.

A further difference between the group of matrix sentences and that of embedded sentences is that with the two minor clause patterns, i.e. the ditransitive pattern and the complex transitive pattern, in matrix sentences the ditransitive pattern occurs only slightly more frequently than the complex transitive pattern, whereas in embedded sentences the complex transitive pattern occurs twice as frequently as the ditransitive pattern.

The pattern ‘other’, finally, needs some comment here. Strictly speaking, it comprises all the patterns other than the five mentioned. In practice, however, the ‘other’ patterns in matrix sentences are virtually all cases of coordinated sentences, the conjoins having been counted as embedded

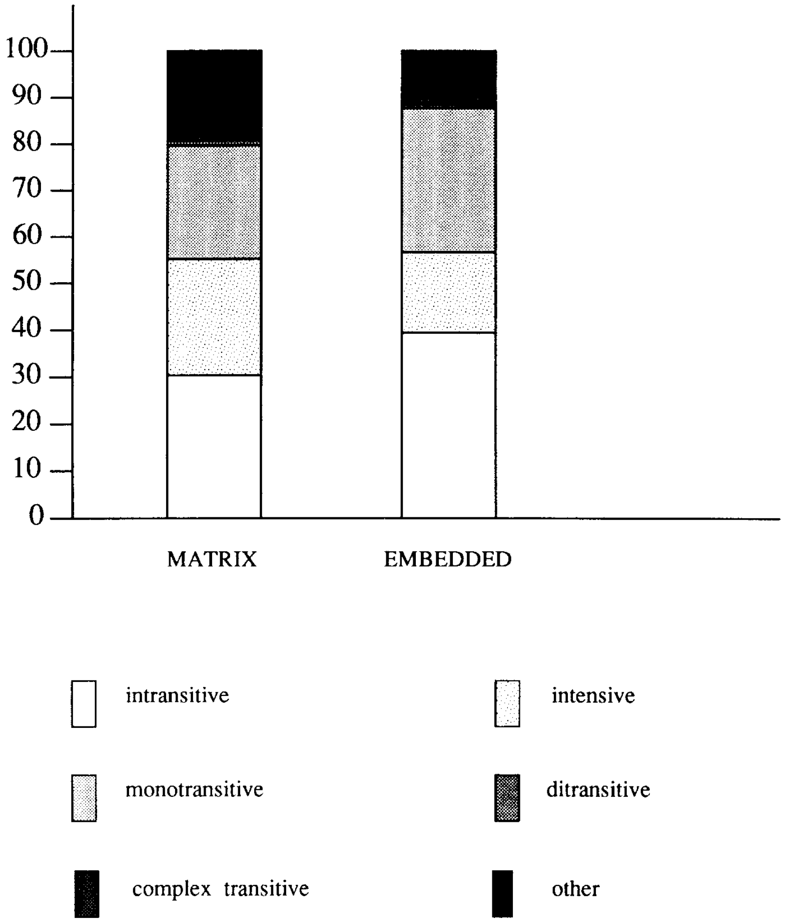

Fig. 4.1 Proportion of clause patterns in matrix and embedded sentences

sentences. In embedded sentences the pattern ‘other’ usually does not signify a coordination. We shall come back to this below.

The information contained in Table 4.1 becomes rather more interesting when the figures are broken down and a distinction is made between finite, non-finite and elliptical sentences (cf. Table 4.2). What emerges then is that in matrix sentences intensive patterns are highly infrequent when the sentence is non-finite, while they are quite common in finite sentences. This tendency also occurs with intensive patterns in embedded sentences, although here the opposition is less outspoken: intensive patterns occur relatively more frequently in finite sentences than in non-finite sentences (21.9% vs. 11.8%).

Table 4.2: Clause pattern distribution in matrix and embedded sentences: finite, non-finite and elliptical sentences compared (N = 15,125)

status	sort	intrans	intens	mono	ditrans	complex	other	total
matrix	finite	2188	1831	1764	87	81	1320	7271
	non-finite	47	4	35	1	1	21	109
	elliptical	19	15	15	2	-	3	54
	total	2254	1850	1814	90	82	1344	7434
em- bedded	finite	1836	997	1362	33	73	253	4554
	non-finite	1168	318	1025	28	58	95	2692
	elliptical	25	5	2	-	-	413	445
	total	3029	1320	2389	61	131	761	7691

The interpretation of Table 4.2 is not straightforward. We used a loglinear model in order to understand the complexities of the interactions that are present in this table. Basically, there is a three-way effect, meaning that clause patterns are potentially related to both their form (finite, non-finite or elliptical) and their status (matrix or embedded). Moreover, there are three two-way effects, indicating that there may be relationships between pattern and status, pattern and form, and between form and status. Finally, there are three one-way effects (pattern, status and form). Each of these seven effects can (and will) influence the

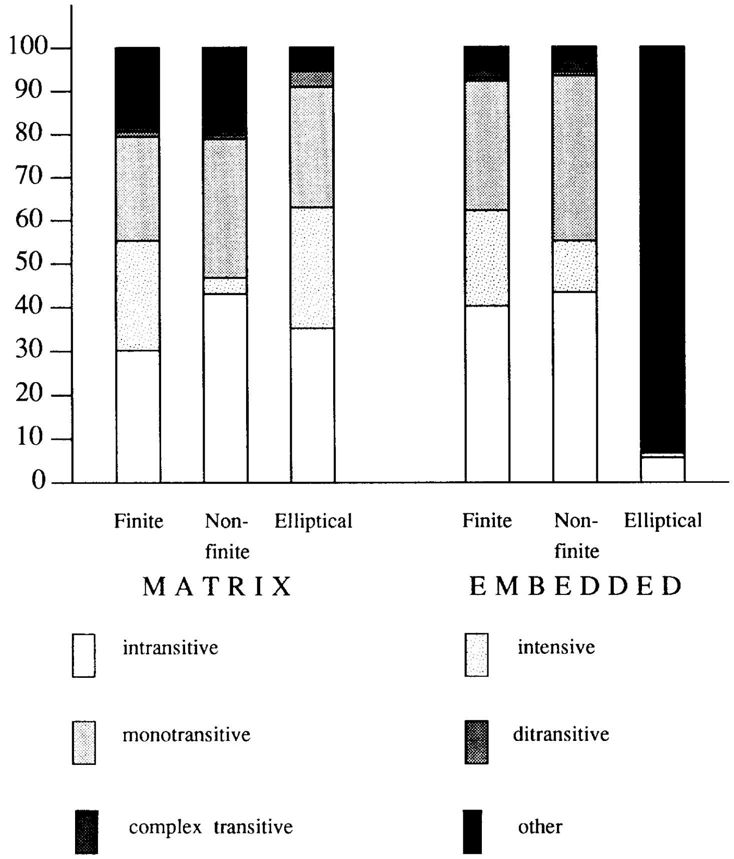

Fig. 4.2 Distribution of clause patterns by sentence form in matrix and embedded sentences

distribution found in Table 4.2, some of them more significantly than others. In order to facilitate the discussion of the interpretation of Table 4.2 we have included Figure 4.2, which puts the numbers found in Table 4.2 in their proper relative perspective.

The various clause patterns are not directly related to the status of the sentences. Any preferences found for matrix or embedded sentences are indirect. For example, the fact that monotransitive patterns occur very frequently in embedded sentences can only be accounted for in the following way. Monotransitive patterns occur relatively often in non-finite sentences. Non-finite sentences are found mainly as embedded sentences. This leads to the observation that monotransitive patterns are found relatively often in embedded sentences. The loglinear model shows no direct relationship between monotransitive patterns and embedded sentences, but in the three-way effect it shows that monotransitive patterns occur often both in finite and in non-finite embedded sentences. At the same time, the model shows that, on the whole, monotransitive patterns do not occur particularly often in finite sentences. This effect is clearly reinforced in the three-way effect.

The opposite also occurs. The analysis of the two-way effects shows that intensive patterns occur more often in finite than in non-finite sentences. The matrix sentences are almost all finite, whereas most of the non-finite sentences are embedded (see Table 4.2). The combination of these two facts does **not** lead to the conclusion that therefore intensive patterns do not occur often in embedded sentences: the three-way effect shows that the opposite is true: intensive patterns do occur relatively often in embedded non-finite sentences. The reason for this is that the intensive patterns that are found in non-finite sentences at all, are almost all found in the embedded sentences. This means that their score in embedded sentences contrasts sharply with that of the non-finite matrix sentences.

Figure 4.2 clearly shows that the three ‘simple’ patterns, viz. intransitive, intensive and monotransitive, occur the most often in all sentence forms, both in matrix and embedded sentences. This is confirmed in the loglinear analysis by the single effect clause pattern. There is one exception, viz. elliptical embedded sentences. The reason for the large number of ‘other’ patterns here is precisely the fact that in many of these cases a ‘proper’ clause pattern cannot be determined.

The only explicit reference to the relationships discussed above that we have come across so far is by Ellegård (1978: 56), who claims that ‘... the most frequent of all sequences is VO, i.e. a clause consisting of just a verb and its object' and '... there is no great word order difference between finite and nonfinite, and between main and subordinate clauses in English.' Our findings do not agree with this. Depending on how we are to interpret his notion of 'word order difference', which we take to mean 'clause structure', according to our findings it is not true (1) that the monotransitive pattern is the most frequent pattern, (2) that there is no difference between finite and non-finite sentences, or (3) that there is no difference between main and subordinate sentences.

Table 4.3: Distribution of embedded sentences over various categories (embedded sentence considered as the IC of a next higher categorial constituent; embedded sentence is IC of X, where X ∊ {SF, SB, SN, PC, ELL, 'PHRASE', other})

	finite sentence	subordin. sentence	non-fin. sentence	parenth. sentence	ellipt.	phrase	other	total
N	2542	2269	192	43	104	2505	36	7691
%	33.0	29.5	2.5	.6	1.4	32.5	.5	100.0

Table 4.3 shows the distribution of embedded sentences over their superordinate structures. It is shown that embedded sentences are found in almost equal numbers as ICs of finite sentences, 'subordinate' sentences (i.e. sentences explicitly introduced by subordinators, the majority of them being also finite),[11] and phrases (roughly 30% each), while a minority of them are found as ICs of non-finite sentences, and elliptical and other structures. In Table 4.4 an overview is given of the various clause patterns over the sentences distinguished in Table 4.3. In Table 4.4 elliptical structures have been merged with 'other' structures.

Table 4.4: Clause patterns in embedded sentences (N = 7691)

status	intrans	intens	mono	ditrans	complex	other	total
phrase-IC	1147	357	835	15	61	90	2505
clause-IC	1836	943	1490	41	68	668	5046
other	46	20	64	5	2	3	140
total	3029	1320	2389	61	131	761	7691

The standardized residual scores of Table 4.4 point to significant scores for the three 'simple' patterns, viz. intransitive, intensive and monotransitive. All this is caused by the relative absence of intensive patterns in sentences which are ICs in phrases. This can be explained by the fact that intensive patterns almost exclusively represent predicative structures. Most of the ICs in phrases are **modifying** ICs in noun phrases (see also Table 4.5), which means that for sentences with intensive patterns there are alternatives, in phrases, in the form of attributive adjective phrases,[12] which would seem to account for their relative absence in ICs in phrases. The relative distribution of the various patterns in Table 4.4 is shown in Figure 4.4.

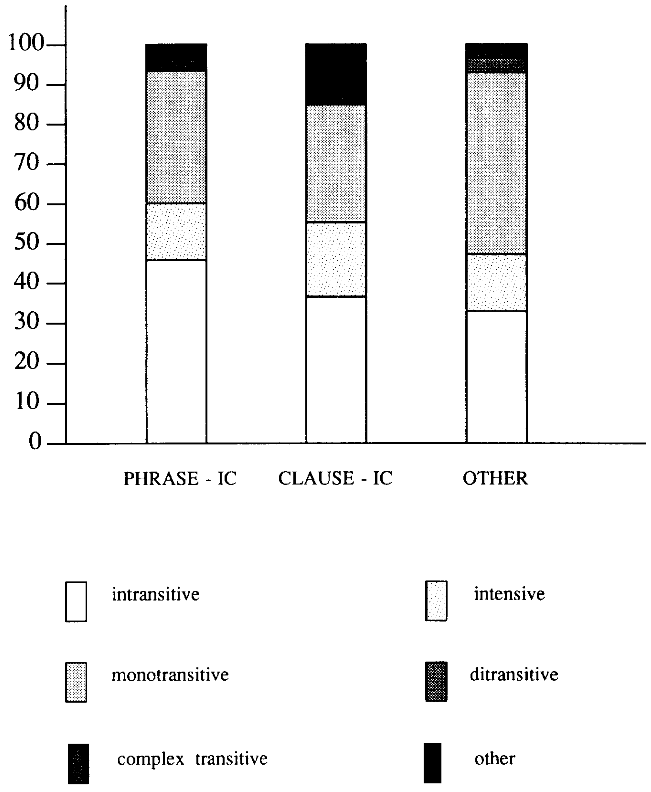

Fig. 4.4 Distribution of clause patterns in embedded sentences by superordinate constituents

Table 4.5: Distribution of embedded sentences as ICs of phrases (POM = postmodifier, NP = noun phrase, AJP = adjective phrase, AVP = adverb phrase, CP = prepositional complement, PP = prepositional phrase)

	POM in NP	POM in AJP	POM in AVP	CP in PP	total
N	1797	210	6	492	2505
%	71.7	8.4	.3	19.6	100.0

The assumption we made about the nature of sentences as ICs in phrases appears to be confirmed by what we see in Table 4.5. Over 80% of ICs in phrases are indeed phrase postmodifiers (the majority of them noun phrase postmodifiers), with less than 20% prepositional complements. Breaking down these figures, we see, in Table 4.6, how the various clause patterns are distributed in the different ICs in phrases.

Table 4.6: Clause pattern distribution in embedded sentences that are ICs of phrases (N = 2505)

status	intrans	intens	mono	ditrans	complex	other	total
POM in NP	943	263	478	11	41	61	1797
POM in AJP	90	33	80	-	2	5	210
POM in AVP	-	-	2	-	-	-	6
CP in PP	110	61	275	4	18	24	492
total	1147	357	835	15	61	90	2505

Although we assumed the low proportion of intensive patterns in ICs in phrases to be due especially to their absence in noun phrase postmodifiers, there is no significant score for noun phrase postmodifiers: intensive patterns score roughly equally low in almost all the ICs in phrases. On the other hand, we do see a kind of complementary distribution for intransitive and monotransitive patterns, with a massive representation of intransitive patterns in noun phrase postmodifiers, and an equally large representation of monotransitive patterns in prepositional complements. The relative distribution of the figures in Table 4.6 is shown in Figure 4.6.

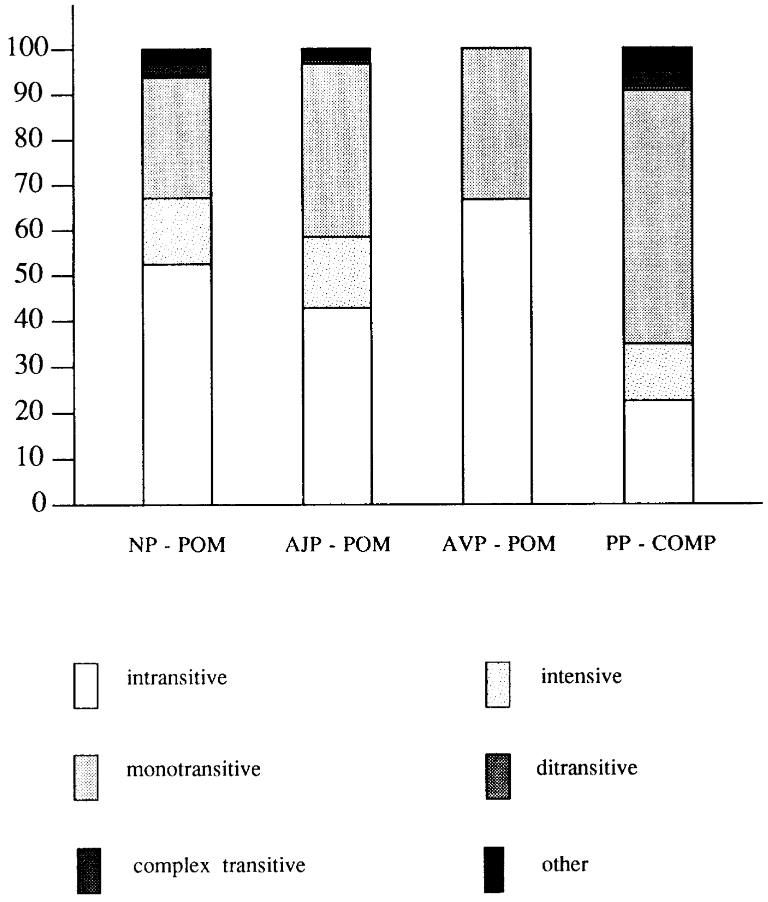

Fig. 4.6 Distribution of clause patterns in phrases

Unlike ICs in phrases, which are almost all postmodifiers, the ICs in clauses show a wide variety of functions. They are shown in Table 4.7, which reads as follows: 421 embedded sentences are adverbials in finite sentences; 41 embedded sentences are adverbials in non-finite sentences; 47 embedded sentences are 'cleft tails' in finite sentences, etc.

Table 4.7: Distribution of embedded sentences as ICs of clause

function	finite sentence	subordin. sentence	non-fin. sentence	parenth. sentence	total
adverbial	421	-	41	-	462
subordinator compl.	-	2269	-	29	2298
cleft sentence tail	47	-	-	-	47
object complement	6	-	-	-	6
subject complement	123	-	2	-	125
verb complement	557	-	82	7	646
elliptical sentence	412	-	-	2	414
notional object	11	-	-	-	11
notional subject	111	-	-	1	112
direct object	567	-	62	4	633
subject	73	-	-	-	73
reported utterance	214	-	5	-	219
total	2542	2269	192	43	5046

The patterns of the sentences in the most common functions in finite sentences are shown in Table 4.8, and those in the most common functions in non-finite sentences are shown in Table 4.9

Table 4.8: Clause pattern distribution in embedded sentences that occur as immediate constituents in a finite sentence restricted to A, SU, NOSU, CS, OD, NOOD and CO (N = 1311)

function in SF	intrans	intens	mono	ditrans	complex	other	total
adverbial	167	45	172	9	8	19	421
subject	29	7	32	1	2	2	73
notion. subj.	33	3	64	1	4	6	111
subj. compl.	35	15	56	1	9	7	123
dir. object	192	153	186	5	5	26	567
notion	5	-	6	-	-	-	11
obj. compl.	1	5	-	-	-	-	6
total	462	228	516	17	28	60	1311

Table 4.9: Clause pattern distribution in embedded sentences that occur as immediate constituents in a non-finite sentence restricted to A, CS, and OD (N = 105)

function in SF	intrans	intens	mono	ditrans	complex	other	total
adverbial	14	2	23	-	1	1	41
subj. compl.	-	-	-	-	2	-	2
dir. object	27	16	15	-	2	2	62
total	41	18	38	-	5	3	105

The only significant scores in Table 4.8 are those for the intensive patterns (remember that intensive patterns occurred especially frequently in ICs in clauses). They are found particularly often in direct object clauses and are noticeably absent in adverbial clauses. The scores in Table 4.9 do not point to any significant distribution of patterns. A comparison of embedded sentences as adverbials and direct objects in finite sentences vs. those occurring in non-finite sentences (Tables 4.8 and 4.9) shows that with finite sentences the difference between the intransitive pattern and the monotransitive pattern is marginal. However, adverbial sentences in non-finite clauses clearly favour the monotransitive pattern, while embedded sentences that occur in the function of direct object in non-finite sentences show a relatively high proportion of intransitives. Given the overall distribution of intransitive and monotransitive patterns in the sentences in this group, however, the latter difference is not significant.

5. The distribution of adverbials

So far we have been looking at the frequency of occurrence and distribution of clause patterns without taking into account any extensions by means of optional adverbials and similar elements. In this section we investigate the relative distribution of adverbial adjuncts over the various clause patterns we have distinguished.

Each of the clause patterns we have distinguished can be extended by means of the addition of one or more adverbial adjuncts. Figure 5.1 shows the distribution of sentences with and without adverbials in matrix and embedded sentences. The various shades of grey of the columns indicate the number of adverbials: increasing darkness indicates a larger number of adverbials.

Over 60% of the matrix sentences contain at least one adverbial, while for embedded sentences this figure is slightly lower: just under 55%. The breakdown of these figures for clause patterns is shown in Tables 5.2 (for matrix sentences) and 5.3 (for embedded sentences).

Table 5.2: Extensions of clause patterns in matrix sentences (N = 6090)

number of adverbials	intrans	intens	mono	ditrans	complex	total
0	549	915	756	60	35	2315
1	846	581	624	22	28	2101
2	530	268	305	7	11	1121
3	236	71	96	1	8	412
4	72	12	28	-	-	112
5	20	3	5	-	-	28
6	-	-	-	-	-	-
7	1	1	-	-	-	2
total	2254	1850	1814	90	82	6090

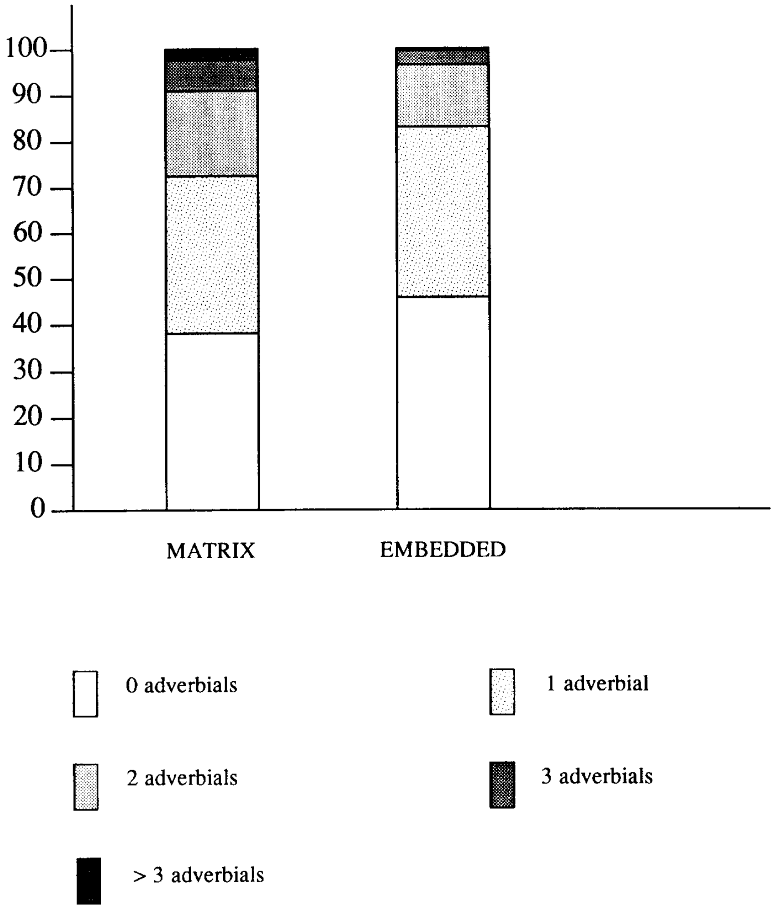

Fig. 5.1 Proportion of number of adverbials per sentence in matrix and embedded sentences

The three 'simple' patterns are all frequently extended by means of adverbials: more than half of them have at least one adverbial. There are, however, differences among them. Intransitive patterns occur the most often with at least one adverbial: over 75% of them have at least one. For monotransitive patterns this figure is slightly lower (slightly under 60%), while for intensive patterns it is even lower (just over 50%). Also, these three patterns are more heavily extended, i.e. many of them take more adverbials. Ditransitive patterns are seen not to be extended by adverbials to the degree that the others are: two out of three have no adverbial at all. Figure 5.2 shows the relative distribution of the number of adverbials in the various patterns.

Table 5.3: Extensions of clause patterns in embedded sentences (N = 6930)

number of adverbials	intrans	intens	mono	ditrans	complex	total
0	893	814	1325	44	98	3174
1	1344	405	785	16	29	2579
2	608	89	233	1	3	934
3	158	12	40	-	1	211
4	20	-	5	-	-	25
5	6	-	1	-	-	7
total	3029	1320	2389	61	131	6930

On the whole, fewer embedded sentences are extended. The number of sentences without adverbials is larger than for matrix sentences. This goes for all the patterns. This is probably related to the fact that embedded sentences already add to the complexity of the matrix sentence, which makes information processing more difficult. This would only be further complicated by the inclusion of additional elements. Also the maximum number of adverbials in embedded sentences is lower than for matrix sentences. Still, we find seven cases of embedded sentences with as many as 5 adverbials (in this and the following examples we have put embedded sentences in italics if the relevant structure occurs in the embedded sentence):

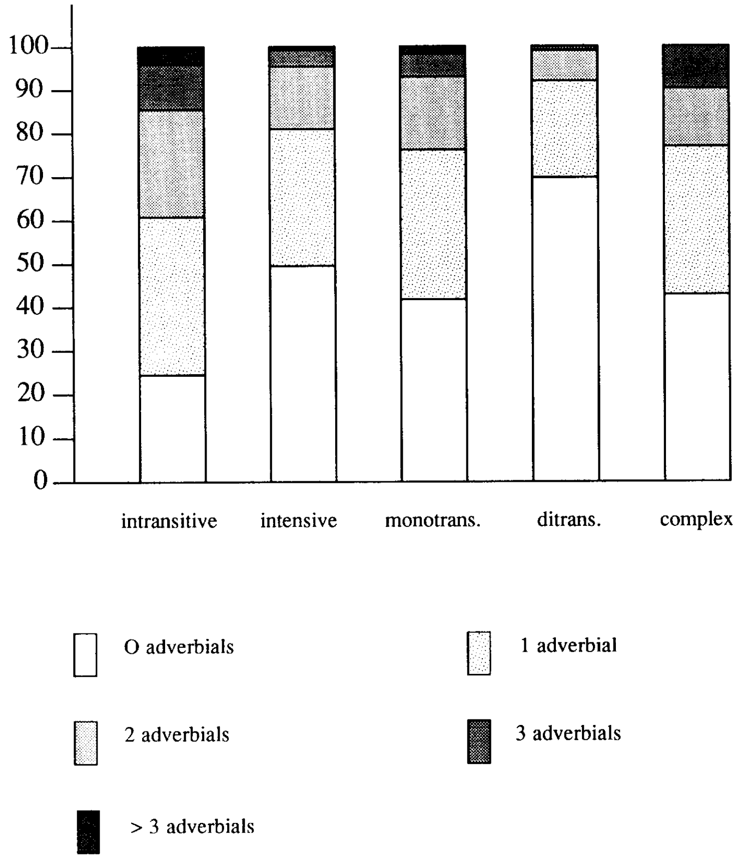

Fig. 5.2 Distribution of number of adverbials per sentence by pattern in MATRIX sentences

> We may safely conclude *that isolated or even repeated experiences in later life do not necessarily give rise to permanent new attitudes when these are in conflict with older and more deep-seated ones unless they are constantly reinforced*, and that when permanent attitudes do arise (as in the case of the sexually-assaulted woman) it is because in fact they fit in very well with previously-existing traits.

On the whole, however, as is shown in Figure 5.3, we see a similar distribution of adverbials in the embedded sentences to that in the matrix sentences, with most adverbials occurring in the sentences with the 'simple' patterns, and the more complex patterns having virtually no adverbials at all.

In order to make a better assessment of the differences in distribution of numbers of adverbials in the various clause patterns in matrix and embedded sentences, we carried out a loglinear analysis of the three way table *clause pattern* × *sentence status* × *number of adverbials*, in which we merged all the cases of more than one adverbial into one category. This was done in order not to create too many cells in the table, too many of which would not contain any observations at all.

The loglinear analysis provides confirmation of what we have observed so far. A significantly low score for 0 adverbials and a ditto high score for 1 adverbial in the matrix sentences confirms that matrix sentences on the whole contain more adverbials, irrespective of the clause patterns. On the level of the clause patterns, without looking at sentence status, we see that sentences with intransitive and monotransitive patterns, in particular, often have more than one adverbial. The sentences with ditransitive and complex transitive patterns, on the other hand, are most often found without adverbials.

The analysis, moreover, provides confirmation of our observations in connection with Figure 4.1, viz. that intransitive and monotransitive patterns are found rather more often in embedded sentences, whereas the other patterns are found more often in matrix sentences.

The three-way effect, in which the interactions among the variables in the table are shown, adds two more significant observations. The first one is that there are relatively many intransitive matrix sentences with 0 adverbials, and few with more than 1 adverbial, while for the embedded intransitive sentences the opposite is true. This is another case of two lower-order effects cancelling each other out in the higher-order effect. For although relatively many matrix sentences contain more

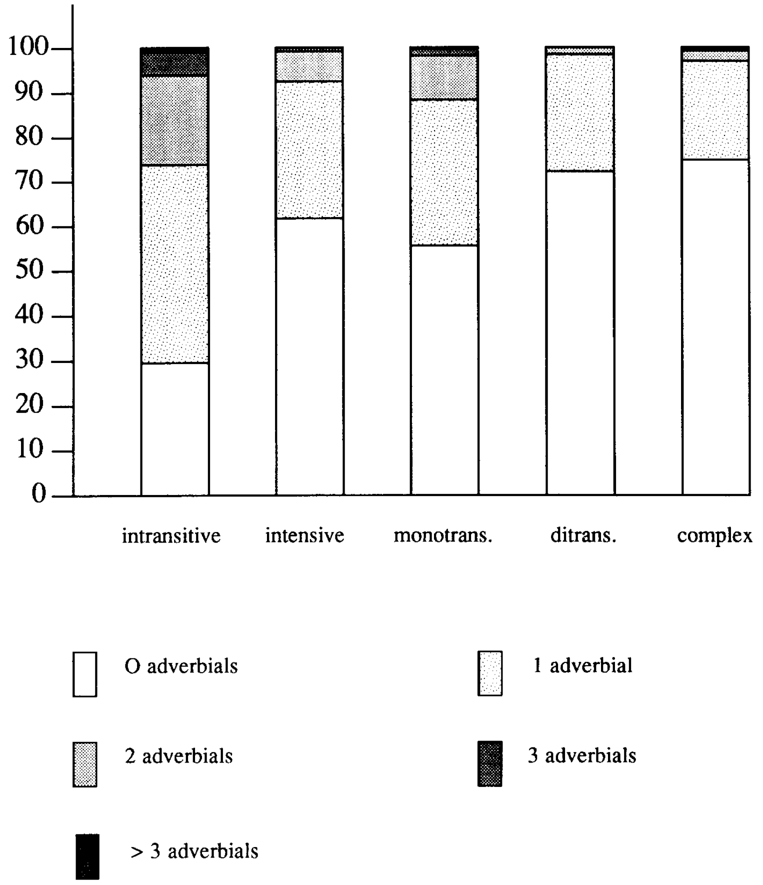

Fig. 5.3 Distribution of number of adverbials per sentence by pattern in EMBEDDED sentences

than one adverbial, and relatively many intransitive sentences contain more than one adverbial, this does **not** lead to the conclusion that many intransitive matrix sentences contain more than one adverbial. Rather the opposite proves to be true.

The reason for this is that the figures for the matrix sentences are contrasted with those for the embedded sentences. If we take another look at Figure 5.3 we can see that the relative score for embedded intransitive sentences with 0 adverbials is low, compared to the other patterns. This keeps the score for embedded intransitive sentences with 0 adverbials low and, as a consequence, pushes up the corresponding score for matrix intransitive sentences with 0 adverbials. This, again, has an effect on the score for matrix intransitive sentences with more than one adverbial. The correct interpretation of this observation, therefore, is that relatively few of the embedded sentences without adverbials have an intransitive pattern.

The second observation in the three-way effect is that there are relatively few matrix monotransitive sentences with more than one adverbial. The interpretation of this observation runs entirely parallel to that of the intransitive patterns.

6. Marked vs. unmarked word order

In our discussion so far we have focussed on the five clause patterns that could be distinguished when taking merely into consideration the obligatory **functional constituents** that occur in these patterns. By a functional constituent we mean a constituent when it is considered from a syntagmatic point of view, in other words, what role it plays in the next higher constituent. For instance, if a constituent is called a subject, it refers to the role it plays in the next higher constituent (the sentence).

The relative order of the functional constituents was not taken into account. Consequently, a discussion of any of the five clause patterns amounted to a discussion of a range of structures, where not only possible extensions by means of optional adverbial adjuncts were disregarded, but where the relative order of the obligatory functional constituents could vary. The range of structures we classified as intransitive, for example, includes not only sentences in which an unmarked word order is found (SU-V) but also sentences where the subject is extraposed or where there is subject-verb inversion. In this section we concern ourselves with the various subpatterns that occur with each of the clause patterns.[13]

Table 6.1: A typology of clause patterns

	TYPE 1:	TYPE 2:	TYPE 3:	TYPE 4:
intrans	SU-V	PRSU-V-SU		V-SU
intens	SU-V-CS	PRSU-V-CS-SU		V-SU-CS
mono	SU-V-OD	PRSU-V-OD-SU	SU-V-PROD-OD	V-SU-OD
ditrans	SU-V-OI-OD	PRSU-V-OI-OD-SU		V-SU-OI-OD
complex	SU-V-OD-CO		SU-V-PROD-CO-OD	

	TYPE 5:	TYPE 6:	TYPE 7:	TYPE 8:
intrans	V-PRSU-SU	SU-V-PRSU		
intens			CS-SU-V	CS-V-SU
mono			OD-SU-V	OD-V-SU
ditrans			OD-SU-V-OI	
complex			OD-SU-V-CO	OD-V-SU-CO

	TYPE 9:	TYPE 10:
intrans		
intens	CS-SU-V-CS	
mono		OD-SU-V-OD
ditrans		
complex		SU-V-CO-OD

Table 6.1 gives an overview of the subpatterns we encountered in our material. While there are as many as 26 subpatterns underlying our earlier classification of the clause patterns, it appears that these can be grouped into 10 different types of (sub)pattern.

Type-1 patterns show the unmarked word order that we typically find in declarative sentences: the subject precedes the verb and any objects and/or complements follow the verb. Type-2 patterns are patterns in which the subject is extraposed, while with type-3 patterns it is the direct object that is extraposed. Patterns of type 4, when they occur in matrix sentences, show an unmarked word order characteristic of interrogative sentences. On the other hand, when encountered in embedded sentences the word order is typically marked, showing subject-verb inversion. Type-5 patterns are similar to patterns of type 2 in that the subject is extraposed. Unlike type-2 patterns (which are found in declarative sentences), type-5 patterns occur in interrogative sentences. Type-6 patterns one would not readily expect since they rather counterintuitively show a provisional subject that follows the notional subject. With type-7 patterns we find that an object or a complement is preposed. Type-8 patterns show preposing of an object or a complement, as well as subject-verb inversion. Patterns of type 9 are typical of raised structures. Finally, the type-10 patterns show inversion in the order of the objects and/or complements. Below each of the clause patterns and its subpatterns is discussed in more detail.

The intransitive pattern

The majority of intransitive patterns consists of type-1 patterns. In 83.4% of the matrix sentences the subject precedes the verb, while in embedded sentences this holds true for 96.0% (cf. Table 6.2). Extraposition of the subject (patterns type 2 and type 5) occurs significantly more often in matrix sentences than in embedded sentences. A closer examination of occurrences where we find a provisional subject shows that they are not all instances of extraposed sentences, but that quite a number are existential sentences.

> On the far side of the little square there was a wall with a pierced decorative border in brickwork.
>
> It is something *in which there is a kind of softness*, Appleby told himself;
>
> I think *there's a flaw in the plausibility of your notion* too.
>
> It will be seen that in fact they correspond rather closely in many respects.
>
> It is generally considered that the two processes have the same basis.
>
> It would be Mrs Martineau who would be chiefly horrified *if, say, it were suggested that the Holman hunts be detached from the*

> *walls for despatch to a museum, or even that a little more room be made here for simple moving about*

Table 6.2: Overall distribution of SV subpatterns

	PATTERNS					
status	TYPE 1	TYPE 2	TYPE 4	TYPE 5	TYPE 6	total
matrix	1880	237	123	13	1	2254
embedded	2909	95	22	3	-	3029
total	4789	332	145	16	1	5283

The unique occurrence in the corpus of the pattern SU-V-PRSU was found for the sentence:

> And what else is there for me to do?

The intensive pattern

Table 6.3 lists the subpatterns that were encountered in the case of intensive complementation. As with the intransitive pattern the occurrences in which we find an unmarked word order are the most frequent by far.[14]

Table 6.3: Overall distribution of SVC subpatterns

	PATTERNS						
status	TYPE 1	TYPE 2	TYPE 4	TYPE 7	TYPE 8	TYPE 9	total
matrix	1557	124	66	13	56	34	1850
embedded	1240	33	3	39	5	-	1320
total	2797	157	69	52	61	34	3170

We find occurrences of provisional subjects with existential sentences (e.g. *There he was, sitting at the table with the reading light on and a book in front of him, dead to the wide* [sic!]) and rather more frequently with extraposed sentences (e.g. *It was evident that Avril saw nothing very amazing in the whole business*).

In a small number of instances the subject complement is preposed. Preposing of the subject complement occurs significantly more frequently in embedded sentences than in matrix sentences. Matrix sentences in which such preposing occurs are almost always exclamatory sentences:

> Here they are now.
>
> How deep this pool is!

Embedded sentences in which the subject complement is preposed are generally *wh*-clauses:

> He advises the posturing but intelligent Hector Hushabye to learn his business as an Englishman, and when asked *what that business is* he replies: ‘navigation’.
>
> The question of culture has been of fundamental importance ever since during the 1930s the American anthropologists Margaret Mead and Ruth Benedict began their studies of primitive cultures to show *how very variable ‘human nature’ can be under different circumstances*;

Subject-verb inversion in embedded sentences is typical of zero-subordinate adverbial clauses:

> The stripling long-livers who bear the brunt of dealing with the visitors would even command our sympathy, *were they not condemned by their own rules to a wholly disagreeable assiduity in snubbing and squashing*.
>
> *Had the genetic code been the overlapping type* it would have been predicted that changes in single DNA bases would result in changes in more than one amino acid.

The monotransitive pattern

The relative order in which the obligatory functional constituents occur with monotransitive sentences appears rather inflexible (cf. Table 6.4). In declarative sentences the subject precedes the verb, while the direct object follows the verb. In yes-no interrogative sentences the verb precedes the subject, in *wh*-interrogatives the direct object occurs sentence initially and is followed by the verb and the subject. The significantly large number of instances in embedded sentences where the direct object occurs sentence-initially (OD-SU-V) must be attributed to the fact that these are generally instances of nominal (or relative) *wh*-clauses. For example:

> They wouldn’t all go to the pictures, if that’s *what you mean*.
>
> And *what Don Juan is saying now* has again relevance to the comedy proper.

In embedded sentences the pattern OD-V-SU is generally found in stretches of direct speech. Examination of the corpus yields occurrences like the following:

> ‘*What am I thinking now?*’ he demanded and composed his features as he concentrated.
>
> Thus the account of the priests and the bells was taken from an article in a German paper which asked in effect: ‘*What next will they believe of us?*’

In very few instances of this pattern do we find extraposition of the subject. In all there are 4 occurrences in the corpus:

> It seemed to amuse Martine to give this answer.
>
> It would require a rather foolish sub-editor to let pass the information that British housewives were indignant at the introduction of a second meatless day if in Germany there were three or four.
>
> Consequently, the code seems to be of the non-overlapping triplet type read in a regular manner from one end to the other, *since it requires three single moves of the register to restore sense in the transcription*.
>
> I mean *it has crossed your mind I might like to know how long I’d be working two shifts a day and no breaks?*

Similarly, extraposition of the direct object hardly ever occurs:

> I mean, *I take it that we haven't just taken part in the greatest clanger of our joint careers*?
>
> So he has taken it into his head to show that, when he wants to, he can carry any woman off her feet.
>
> You don't need to try *to keep it from me that something's up*.

Table 6.4: Overall distribution of SVO subpatterns

	PATTERNS						
status	TYPE 1	TYPE 2	TYPE 4	TYPE 7	TYPE 8	other	total
matrix	1662	2	94	9	44	3	1814
embedded	2225	2	2	153	3	4	2389
total	3887	4	96	162	47	7	4203

The ditransitive pattern

Ditransitive sentences have a relatively fixed word order: 92.7% of the sentences conform to the unmarked type-1 pattern, while only some 7.3% displays a deviant word order (cf. Table 6.5).

Table 6.5: Overall distribution of SVOO subpatterns

	PATTERNS				
status	TYPE 1	TYPE 2	TYPE 4	TYPE 7	total
matrix	86	1	-	3	90
embedded	54	-	1	6	61
total	140	1	1	9	151

Extraposition of the subject with ditransitive sentences is rare. In the corpus only a single occurrence is found:

> It took you long enough to find him out, cretin or not.

Also subject-verb inversion appears to be a rare phenomenon where ditransitive sentences are concerned. The one instance that we find in the corpus is found in an embedded sentence and is an interrogative:

> 'If that were true it really would be the last thing any of us here would know about,' she said coldly, and added, to soften the snub, '*can we offer you some tea?*'

Preposing of the direct object in matrix sentences typically occurs in interrogative sentences. For example:

> Whatever has Martine been telling you?

Embedded sentences in which the direct object has been preposed are generally *wh*-clauses:

> I thought that might have been *what security told you when they sent for you this afternoon*?

> It was on the receipt *which Mayo gave Luke in return for the devices.*

The complex transitive pattern

The complex transitive sentences encountered in the corpus display a relatively high degree of variation as far as their word order is concerned. Only 74.2% of the sentences have an unmarked word order (cf. Table 6.6).

Table 6.6: Overall distribution of SVOC subpatterns

	PATTERNS					
status	TYPE 1	TYPE 3	TYPE 7	TYPE 8	TYPE 10	total
matrix	64	8	3	1	6	82
embedded	94	7	23	-	7	131
total	158	15	26	1	13	213

Interesting is the fact that unlike with any of the other patterns extraposition of the subject does not occur (i.e. no occurrences were found in the corpus), nor do we find subject-verb inversion without preposing of the object or complement.

While with matrix sentences preposing of the direct object is not very frequent, with embedded sentences we find that in 17.6% of the complex transitive sentences the direct object has been preposed. Generally the direct object in such cases is realized by a *wh*-element:

> That's *what we used to call telepathy*.
>
> When Cleopatra contrives the treacherous assassination of a prisoner *whom he regards as a guest*, he is angry but not surprised, and her attempt to justify herself educes from him one of the central speeches of the day.

Extraposition of the direct object occurs relatively more frequently in matrix sentences than in embedded sentences (9.8% vs. 5.3% respectively). Among the occurrences in the corpus we find:

> I think it best to speak to Charles first.
>
> We must think it strange that one who began his career as professional critic of one art after another, and ended it as himself the greatest living exponent of a major artistic form, should be almost uniformly unconvincing in the presentation of artists of any sort.
>
> Interesting men, one was considered to feel, represented a category *she judged it unnecessary to approve of*.
>
> But from this it also follows that those *who do not find it easy to associate themselves with groups* are unlikely to be changed by them, as was seen in the case of the neurotic members of the Bennington community.

A phenomenon that we do not encounter with any of the other patterns is that of inversion involving the object and the complement, without the help of a provisional object.[15] Examples of such inversion are:

> 'Major Barbara' uses as a springboard for its action that theme of a father making the acquaintance of his grown-up children which

> had proved effective in 'You never can tell'.
>
> Those who depreciate the arrival of Richard III at Bosworth Field on a live horse and who are unimpressed by Jack Tanner's expensive car may judge Jejune the London taxicab which here jerks on and off the stage in the first act.
>
> At his death he took pleasure in emphasizing its prosperous, bourgeois, curiously unaesthetic tone by *proposing to the British nation as place of pilgrimage his extremely commonplace house with its extremely commonplace contents.*
>
> Nor need the response be immediate; for memory makes possible a great variety of conditioned responses and a tendency to respond long after the exposure to the original stimulus.

7. Conclusion

In this article we have presented an inventory of the most common clause patterns in Modern English. In this inventory we have looked at the relationships between the basic clause patterns that we distinguish and such features as the status of the sentence, the nature of the superordinate constituent, the form of the sentence, the actual order of the constituents in the sentence, and the occurrence of optional adverbial elements.

It has been shown that it cannot be maintained that the distribution of clause patterns with respect to the other variables is entirely random in the corpus under investigation. We shall be in a position in the near future to compare the results obtained in this study with those obtained from a larger corpus (the TOSCA corpus, cf. Oostdijk, 1991).

What we shall also want to do in the near future is to look at adverbial placement, which was beyond the scope of the present study. Also, the notion of marked word order needs to be elaborated. It was suggested, in Section 6, that specific orders of constituents which may be marked in matrix declarative sentences, may be unmarked in certain types of embedded sentence, or in matrix interrogative sentences. We did not have occasion to go into these aspects in the present study. However, it is clear that we need to gain more insight into the implications of such notions as end-focus, end-weight, topicalization, etc., with respect to the actual order of constituents in sentences.

Notes

1. The project is carried out jointly by Jan Aarts, Hans van Halteren and the authors. The present paper is a slightly revised version of a paper which has been published earlier as working paper no. 28 in the series *Dutch Working Papers in English Language and Linguisties* (DWPELL).
2. Although Quirk *et al*. (1972, 1985) from time to time do provide frequency information that has been derived from the Survey of English Usage, these are mostly isolated bits; cf. Quirk *et al.* 1985: 817, ‘In a collection of 858 *wh*-questions from the files of the Survey of English Usage, chiefly in surreptitiously recorded spoken material, 775 had a falling tone.’
3. Quirk *et al*. (1985: 817): ‘The construction with a preposition in final position ... **is less desirable** when the preposition is remote from its complement or when it is syntactically bound closer to the complement than the verb. **Awkward sentences** like *What time did you tell him to meet us at?* are generally avoided. A sentence like that **would probably be replaced by** *At what time did you tell him to meet us?* in formal style or, more generally, by *When did you tell him to meet us?* or prepositionless *What time did you tell him to meet us?* The **awkwardness reaches comic proportions** when several final particles co-occur: *What did you bring this book to be read out of up for?*’ (bold face added)
4. The tables and figures in this paper have been numbered according to sections they occur in.
5. Biber (1988) reports on difficulties of identifying certain constructions automatically in a corpus that has not been syntactically analyzed (so that you can only look at literal strings or, at best, tags). De Haan (1989) experienced similar difficulties in his initial identification of postmodifying relative clauses in the Nijmegen Corpus (when he started his project the Corpus had not yet been analyzed).
6. Major ‘standardized’ corpora for English, the Brown Corpus and the Lancaster-Oslo/Bergen (LOB) Corpus, contain 1 million words each.
7. This is not at all surprising since the descriptive framework for the analysis of the Nijmegen Corpus was largely based on Quirk *et al*. (1972). At the time the Nijmegen Corpus was compiled and analyzed the Comprehensive Grammar had not been published yet.

8. The functional constituents referred to here are: subject (SU), verb (V), subject complement (CS), direct object (OD), indirect object (OI) and object complement (CO).
9. As a consequence a count of, for example, the intensive pattern comprises a range of structures as diverse as SU-V-CS, CS-SU-V, V-SU-CS, A-SU-V-CS-A, CS-SU-V-A, etc.
10. Actually, there are 10,385 analysis trees in the LDB. These trees represent **utterances**. Quite a few utterances, however, are not realized by sentences, but by phrases or mark-up (e.g. speaker turns in the drama fragments and sports commentaries).
11. The category of the subordinate clause was introduced in the analysis of the Nijmegen corpus in order to make an explicit distinction between clauses introduced by subordinators and those that are not introduced by subordinators, e.g. *wh*-nominal clauses, or non-finite or verbless clauses without subordinators. In all cases the subordinate clause (SB) is analyzed as consisting of two constituents, viz. the subordinator (SUB) and a finite clause (SF) or a non-finite (SN) or elliptical (ELL) clause.
12. The same observation was also made by De Haan (1989b).
13. Any extensions by means of optional adverbials are disregarded.
14. In matrix sentences the word orders SU-V-CS, V-SU-CS and CS-V-SU can be considered unmarked for declarative sentences, interrogative sentences and *wh*-interrogatives respectively. In embedded sentences only SU-V-CS is considered to be unmarked.
15. With ditransitives inversion of the two objects as such does not occur since the descriptive approach that was adopted assumes that whenever there are two consecutive objects the first of these is the indirect object and the second the direct object. In sentences like ‘I gave the book to John’ *the book* is taken to be the direct object, while *to John* is looked upon as an adverbial.

References

Aarts, F. (1971): ‘On the distribution of noun phrase types in English clause-structure’ in: *Lingua*, 26: 281-293.

Aarts, F. and J. Aarts (1982): *English syntactic structures*. Oxford: Pergamon Press Ltd.

Aarts, J. and W. Meijs (eds.) (1986): *Corpus linguistics II. New studies in the analysis and exploitation of computer corpora*. Amsterdam: Rodopi.

Biber, D. (1988): *Variation across speech and writing*. Cambridge: Cambridge University Press.

Chafe, W. (1976): 'Givenness, contrastiveness, definiteness, subjects, topics and point of view' in: Li, C.N. (ed.) (1976): 25-55.

Ek, J.A. van (1966): *Four complementary structures of complementation in contemporary British English*. Groningen: Wolters.

Ellegård, A. (1978): *The syntactic structures in English texts. A computer-based study of four kinds of text in the Brown University Corpus*. Gothenburg: Acta Universitatis Gothoburgensis.

Erdman, P. (1976): *THERE sentences in English. A relational study based on a corpus of written texts*. Munich: Tuduv Verlaggesellschaft mbH.

Gleason, H.A. (1965): *Linguistics and English grammar*. New York: Holt, Rinehart and Winston Inc.

Haan, P. de (1987): 'Exploring the linguistic database: Noun phrase complexity and language variation', in: Meijs, W. (ed.) (1987): 151-165.

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

Haan, P. de (1989b): 'Structure frequency counts of Modern English: The set-up of a quantitative study', in: *Dutch Working Papers in English Language and Linguistics*, 13: 1-15.

Haan, P. de and R. van Hout (1986): 'Statistics and corpus analysis: A loglinear analysis of syntactic constraints on postmodifying clauses', in Aarts, J. and W. Meijs (eds.) (1986): 79-97.

Haan, P. de and R. van Hout (1988): 'Syntactic features of relative clauses in text corpora', in: *Dutch Working Papers in English Language and Linguistics*, 2: 1-28.

Halliday, M.AK. (1969): 'Options and functions in the English clause', in *Brno Studies in English*, 8: 81-88.

Halteren, H. van and Th. van den Heuvel (1990): *Linguistic exploitation of syntactic databases. The use of the Nijmegen Linguistic DataBase program*. Amsterdam: Rodopi.

Halteren, H. van and N. Oostdijk (1988): 'Using an analyzed corpus as a linguistic database', in Roper, J. (ed.) (1988): 171-181.

Hermerén, L. (1978): *On modality in English. A study on the semantics of the modals*. Lund: CWK Gleerup.

Hough III, G.A. (1971): *Structures of modification in contemporary American English*. The Hague: Mouton.

Huddleston, R. (1971): *The sentence in written English. A syntactic study based on an analysis of scientific texts.* Cambridge: Cambridge University Press.

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

Keulen, F. (1986): 'The Dutch computer corpus pilot project. Some experiences with the semi-automatic analysis of contemporary English', in: Aarts, J. and W. Meijs (eds.) (1986): 127-162.

Kruisinga, E. (1909-32): *A handbook of present-day English.* Groningen.

Lebrun, Y. (1965): *CAN and MAY in present-day English.* Brussels: Presses Universitaires de Bruxelles.

Leech, G.N. (1983): *Principles of pragmatics*. London: Longman.

Li, C.N. (ed.) (1976): *Subject and topic.* New York: Academic Press.

Li, C.N. and S.A. Thompson (1976): 'Subject and topic: a new typology of language', in Li, C.N. (ed.) (1976): 457-489.

Mair, C. (1990): *Infinitival complement clauses in English.* Cambridge: Cambridge University Press.

Meijs, W. (ed.): *Corpus linguistics and beyond.* Proceedings of the seventh international conference on English language research on computerized corpora. Amsterdam: Rodopi.

Meyer, C.F. (1992): *Apposition in contemporary English.* Cambridge: Cambridge University Press.

Olofsson, A. (1981): *Relative junctions in written American English.* Gothenburg: Acta Universitatis Gothoburgensis.

Oostdijk, N. (1991): *Corpus linguistics and the automatic analysis of English.* Amsterdam: Rodopi.

Poutsma, H. (1904-1926): *A grammar of late modern English.* Groningen: P. Noordhoff.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1972): *A grammar of contemporary English.* London: Longman.

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

Roper, J. (ed.) (1988): *Computers in linguistic and literary computing.* Proceedings of the thirteenth ALLC conference, University of East Anglia (Norwich) 1-4 April, 1986; under the direction of J. Hamesse & A. Zampolli (series eds.). Paris-Geneva: Champion-Slatkine.

Scheffer, J. (1975): *The progressive in English.* Amsterdam: North Holland.

Svartvik, J. (1966): *On voice in the English verb.* The Hague: Mouton.

Vestergaard, T. (1977): *Prepositional phrases and prepositional verbs. A study in grammatical function*. The Hague: Mouton.

Wekker, H. (1976): *The expression of future time in contemporary British English*. Amsterdam: North Holland.

Yotsukura, S. (1970): *The articles in English. A structural analysis of usage*. The Hague: Mouton.

Appendix A: The texts of the corpus

Allingham, M. (1965): *The mind readers*. London: Chatto and Windus (edition used: Penguin Books 1968: pp. 46-103). Text variety: crime fiction.

Innes, M. (1966): *The bloody wood*. London: Victor Gollancz (edition used: Penguin Books 1968: pp. 27-89). Text variety: crime fiction.

Stewart, J.I.M. (1963): *Eight modern writers*. (The Oxford history of English literature XII.) Oxford: Clarendon Press (pp. 122-183). Text variety: literary criticism.

Brown, J. (1963): *Techniques of persuasion*. Harmondsworth: Penguin Books (edition used: 1967, pp. 37-92). Text variety: popular scientific writing.

Paul, J. (1965): *Cell biology*. London: Heinemann Educational Books Ltd. (edition used: 1967, pp. 102-178). Text variety: scientific writing.

Livings, H. (1962): *Stop it, whoever you are*. Harmondsworth: Penguin Books (edition used: 1967, pp. 15-79). Text variety: drama.

Livings, H. (1963): *Nil carborundum*. Harmondsworth: Penguin Books (edition used: 1967, pp. 214-239). Text variety: drama.

Wimbledon final. BBC TV 1, 1968. A transcript of the comment to the match of Laver vs. Roche on 5-7-1968. Commentators: D. Maskell, J. Kramer and D. Coleman. Text variety: sports commentary.

Wightman cup. BBC TV 1968. A transcript of part of the comments to the singles matches of Christine Jones vs. Nancy Richey and Virginia Wade vs. Mary Ann Eisil, and of the doubles match of Winnie Shaw and Virginia Wade vs. Nancy Richey and Mary Ann Eisil. Commentators: P. West and D. Maskell. Text variety: sports commentary.

Correspondence: Erasmusplein 1
NL-6525 HT Nijmegen
e-mail: oostdijk@lett.kun.nl or dehaan@lett.kun.nl