

Finite Comment Clauses in Dutch: A Corpus-based Approach

Carla Schelfhout
Peter-Arno Coppen
Nelleke Oostdijk

Radboud University Nijmegen

The present paper presents the results of a corpus-based study of the form and distribution of finite comment clauses in Dutch. More specifically, it was investigated where in the sentence such clauses can occur. For the analysis of the data, a topological descriptive model was used. While in the literature an extraction analysis has been suggested in order to account for finite comment clauses in English and German, our findings lead us to challenge this type of analysis and argue that a parenthetical analysis is to be preferred.*

1. Introduction.

This study is part of a larger research program aimed at the automatic syntactic analysis of interruption constructions in Dutch. The type of interruption constructions we are aiming at, called INTERCALATIONS in Schelfhout et al. 2003a, is defined as the interruption of a running sentence by syntactic material that cannot be analyzed directly as (an) immediate constituent(s) of that sentence. After this interruption the sentence continues without experiencing syntactic or prosodic consequences of the intercalation. More specifically, intercalations seem to be set apart from the sentence with respect to prosody: the sentence prosody stops when the interruption is reached and continues at the point where it had stopped after the interruption. In addition, intercalations do not seem to have a syntactic influence on the clause, which can for instance be seen when they occur before the finite verb in Dutch. When an adjunct such as an adverbial occurs there, it causes inversion of verb and subject, but intercalations can occur between subject and verb without causing inversion. Examples are interjections, vocatives,

* Thanks are due to Antal van den Bosch and Hans van Halteren for their help in tagging the corpus, and to Toni Rietveld for his statistical advice.

reporting clauses and parenthesized clauses, but also nonrestrictive appositives, transparent free relatives (Wilder 1999; Schelfhout et al. to appear) as well as certain types of conjunction reductions that can be argued to be intercalations.

In this paper, we are concerned with FINITE COMMENT CLAUSES, or PARENTHETICALS; that is, interruptions, such as those in 1 and 2.¹

- (1) Dat is erg belangrijk, DENK IK,
that is very important think I
voor de ontwikkeling van onze theorie.
for the development of our theory
'That is very important, I think, for the development of our theory.'
- (2) Ze waren bang ZO LIJKT HET voor de gevolgen.
they were afraid so seems it of the consequences
'They were afraid, or so it seems, of the consequences.'

Our concern is mainly the analysis of finite comment clauses in Dutch. From time to time, however, we discuss English and German literature as well. It is our contention that reference to these other languages helps provide insight into the phenomenon we are investigating, while our conclusions with respect to Dutch largely carry over to these adjacent languages.

Since we aim at an analysis that can be used for Natural Language Processing (NLP) applications, it is important that we arrive at a description that accounts for real language use. In order to obtain information about the actual distribution of comment clauses we conducted a corpus study.

The present article is structured as follows. First, we describe our corpus study and the results. Next, we discuss how these results are interpreted when analyzing the comment clause in Dutch. Finally, our conclusions are demonstrated to be in line with the analysis of other interruption constructions as they have emerged from previous studies.

¹ We restrict ourselves to comment clauses that occur in sentence-internal (or medial) position. Where relevant to our argumentation, we will occasionally refer to comment clauses in sentence-final position.

2. Corpus Results.

Intercalations are often tacitly assumed to occur in any syntactic position. However, according to the discussion in the literature (see section 3), finite comment clauses are often expected in only one or two positions. In view of a search for the correct analysis of finite comment clauses this makes the question “Where do finite comment clauses occur?” a legitimate one.

In this investigation, we addressed this question by conducting a corpus study of the distribution of finite comment clauses in both written and spoken Dutch. A corpus was compiled comprising approximately 1.5 million words. The written component consists of approximately one million words with their origin in print. The 478 documents in this component were taken from the Internet. The spoken component consists of 930 files derived from the Spoken Dutch Corpus (Corpus Gesproken Nederlands or CGN; see Oostdijk 2000). The composition of the corpus is displayed in table 1.

Written		Spoken	
Essay	127,122	Lecture	62,810
Interview	126,376	Interview	62,510
News	123,140	News	80,121
Novel	255,503	Commentary	125,747
Short story	255,653	Private conversation	63,883
Scientific writing	125,846	Telephone conversation	63,205
Total	1,013,640	Total	458,276

Table 1. Corpus composition.

We conducted a qualitative investigation into the variation within comment clauses. This implies that we started from the canonical examples as discussed in the literature and then looked at randomly selected parts from the corpus to spot similar constructions. We then decided whether they were indeed comparable constructions by, among other things, determining if they could be replaced by canonical finite comment clauses or not. It turned out that finite comment clauses appear in two forms:²

² Another construction can be found in which the subject occurs initially with a finite verb following. This verb expresses an opinion; for example, *The train*

1. A main verb expressing an opinion (*denken* ‘think’, *veronderstellen* ‘suppose’, etc.), followed by a subject, one or more optional modifiers, and possibly preceded by the adverbial *zo* ‘so’.
2. A finite copula (*zijn* ‘be’, *lijken* ‘seem’, etc.), followed by a subject or an indirect object, one or more optional modifiers, and possibly preceded by the adverbial *zo* ‘so’.

The variation found in comment clauses is exemplified in examples 3–6 below, which are all derived from the corpus.

- (3) Een doffe tik van metaal op metaal, dat was DENK IK
 a dull tap of metal on metal that was think I
 de beste omschrijving.
 the best description
 ‘A dull tap of metal on metal, that was the best description, I think.’
- (4) Het was,
 it was
 ZO HERINNEREN ZIJN VRIENDINNEN EN MINNARESSEN ZICH,
 so remember his girlfriends and lovers PRT
 alsof hij geen innerlijk bezat
 like he no inner-self had
 ‘It was, his girlfriends and lovers remember, as if he did not possess an inner self.’

stops between, I think, Tilburg and Breda. In this case, adding modifiers or adverbial *zo* is impossible. This construction should not be confused with the one we are concerned with in this article.

- (5) 't Is heel wat werk LIJKT ME als ik het 'ns zo hoor.
 it is quite some work seems me if I it once so hear
 'It's quite a lot of work, I guess, judging from what you say.'
- (6) Jozua trekt als een dolle stier door het beloofde land
 Joshua travels like a wild bull through the promised land
 om het te ontdoen, ZO LIJKT HET, van de oorspronkelijke
 in-order it to strip so seems it of the original
 bewoners
 inhabitants
 zodat het volk van Israel er onbekommerd kan leven.
 so-that the people of Israel there carefree can live
 'Joshua rages through the promised land like a wild bull in order to
 strip it, so it seems, of the original inhabitants, so that the people of
 Israel could live there carefree.'

With the insights gained into the nature of comment clauses, we semi-automatically searched the material for comment clauses. The CGN material was already pre-processed in so far that it was split up into sentences in which each word was given a word class tag and the appropriate lemma. We used the CGN tools and procedures to repeat this for the written part of the corpus, so that both kinds of material had a comparable annotation. This allowed us to write a Perl program selecting all sentences that contained a verb of the interesting kind and/or the word *zo* 'so' not occurring in the very first position of the clause. Of course, this resulted in too many sentences being selected, but the real finite comment clauses were then selected by hand. The results are summarized in table 2.³ The first column records the number of instances found. In the second column the numbers have been standardized and represent the number of instances per 10,000 words.

³ The sentences themselves can be retrieved at <http://lands.let.kun.nl/~schelfht/>.

	Number of comment clauses	Comment clauses per 10,000 words
Spoken	195	4.3
Written	76	0.7
Total	271	1.8

Table 2. Instances of comment clauses found in the corpus.

Next we investigated the positions where the comment clauses occurred. To this end we analyzed the corpus sentences according to the standard topological model used in Dutch traditional grammar (as described in the authoritative Dutch grammar *Algemene Nederlandse Spraakkunst* (ANS; Haeseryn et al. 1997). This model distinguishes two verbal poles in the Dutch sentence (the left and right bracket, LB and RB), with a middle field (MI) in between. The left bracket contains the finite verb in main clauses and the subordinator in subordinate clauses, while the right bracket contains remaining verbs (if any) in main clauses and all verbs in subordinate clauses. Preceding the first pole the PRE field can be found, which is used for topicalized elements, possibly preceded by a left dislocation field (LD). Following the second verbal pole is the POST field (for extraposed elements), possibly followed by a right dislocation field (RD).⁴ Any of these fields can be empty. Some examples are displayed in table 3.

⁴ The terms used in the ANS are left dislocation field, topicalization field, first verbal pole, middle field, second verbal pole, extraposition field, and right dislocation field. For the term “middle field,” the term “inner field” could also be used. In German, the terms *Vorfeld*, *Mittelfeld*, and *Nachfeld* are in use for prefield, middle field, and postfield. There is no straightforward relationship between this topological analysis and a generative analysis, if only because no standard generative analysis exists for Dutch sentences. Moreover, even within specific generative theories the relationship between sentence position and topological position varies. For instance, a finite verb in a main clause (the LB position) may be located in the head of IP or the head of CP, depending on the generative theory, or even depending on the type of sentence. The position of adjuncts is even more problematic in generative analyses. Although complements generally occur to the left of the main verb as a result of some raising process, or to the right if the verb is moved, it is not at all clear how recent generative theories allowing only left adjunction can account for adjuncts occurring to the right of the verb (in the POST field). Therefore, given the

LD	PRE	LB	MI	RB	POST	RD
		snap understand	je dat you that			
	Ik I	heb have		gevloekt sworn	als een ketter like a trooper	
Elvis, Elvis	die him	zou would	ik graag I very	willen horen want hear		
		omdat because	ik hem I him	haat, hate		die rotzak that jerk

Table 3. Examples of a topological analysis.

Within the topological framework, finite comment clauses could theoretically occur between consecutive fields or within a certain field (except for the LB field which by definition can only contain one element). In addition, they can occur between two (coordinate or subordinate) clauses, a position indicated by “#.” In table 4, we exemplify the positions between the left dislocation field and the PRE field (LD-PRE for short) and within the middle field (MI).⁵

LD		PRE	LB	MI	RB
Prince, Prince	MEEN IK think I	die him	zou would	ik graag I very-much	willen ontmoeten want meet
Prince, Prince		die him	zou would	Ik MEEN IK graag I think I very-much	willen ontmoeten want meet

Table 4. Comment clauses in the positions LD-PRE and MI.

The distribution of comment clauses in written and spoken data is shown in table 5 on the next page.⁶ The distribution is displayed in terms

current state of generative theory, it seems impossible to relate topological analyses straightforwardly to generative analyses. However, such a relationship is not crucial to the line of reasoning in this paper.

⁵ Non-used peripheral fields (LD, POST, RD) are not shown below for reasons of space.

⁶ If an intercalation appears between two non-neighboring fields, we have transparency. Since the intervening field(s) is (are) empty, we are unable to decide where exactly the intercalation occurs. These instances are excluded from the

of both the absolute numbers and the (relative) proportion of the occurrences in various positions of the total number of occurrences. Note that intercalations appearing within a field rarely occur within a major constituent (the few instances we found all occurred in the spoken component of our corpus).

To establish whether this distribution was the result of coincidence or really reflected certain preferences, we performed a likelihood ratio test.⁷ To ensure the independence of each instance we decided to use only one sentence per file, so we removed sentences from files that had already delivered another sentence. After this operation only 57 sentences from written material and 106 sentences from spoken material remained. These numbers were counted back to instances per million words to ensure comparability. The value of the likelihood ratio statistics was 42.701 (df = 9), $p < 0.01$. In view of the low numbers, the significance of this statistical test can be assumed to reflect a significant difference.

Position	Written		Spoken	
	Number	%	Number	%
LD	0	0	0	0
LD-PRE	5	6.6	1	0.5
PRE	3	3.9	5	2.7
PRE-LB	7	9.2	8	4.3
LB-MI	21	27.6	29	15.6
MI	11	14.5	84	45.2
MI-RB	2	2.6	2	1.1
RB	0	0	0	0
RB-POST	1	1.3	8	4.3
POST	0	0	7	3.8
POST-RD	0	0	0	0
RD	1	1.3	3	1.6
#	25	32.9	39	21
Total	76	99.9	186	100.1

Table 5. The distribution of comment clauses in corpus data.

further research. This was the case for nine out of 195 spoken comment clauses; transparency did not occur with the written comment clauses.

⁷ Performing a chi-square test here was impossible as the expected values of frequencies in a number of cells was less than 5.

Only five out of 271 instances of comment clauses use a copula, three of these occur in written data and two in spoken data. They occur in four common positions in the sentence.⁸ This number is too small to draw reliable conclusions, but there does not seem to be a reason to assume a different analysis for comment clauses with a copula. Only in three cases does a comment clause occur within a major constituent. In all three cases it occurs between a preposition and an NP in spoken language, where we have an indication that the speaker is confused or hesitant (for example, repeating the preposition or saying *uh ...*).

The literature does not provide very explicit claims about the positions in which intercalations can appear, but the general (tacit) assumption seems to be that they can occur anywhere. If this were true, we would expect a regular distribution, but the distribution of comment clauses in table 5 is far from regular. Both in written and in spoken data, three positions together cover more than 75% of the cases; namely, the positions LB-MI, MI, and #. The position between clauses and the position between the finite verb and the middle field are clear enough, but the position MI is a rather broad category. The middle field can contain various elements. In order to obtain a more accurate description of the distribution, we will first develop a more specific description of the instances in the middle field.

The order of the elements in the middle field in Dutch is discussed extensively in the ANS (chapter 20.4/5) and nicely summarized in Haeseryn 1998. The elements in the middle field are ordered on the basis of their information value (the higher the information value, the further to the right an element occurs), their relation to the main verb (elements closely related to the verb, like predicates, occur closer to the right bracket), and their complexity (the heavier an element is, the further to the right it occurs). The ANS describes a division over three subfields for which we developed the following paradigm. The first part of the middle field is the canonical position for subjects, clitics, and particles (see Gerrits 2001). We call this the pre-middle field or PREMI. The last part of the middle field contains predicates, R-particles (also known as stranded prepositions), or resultatives (see Van Dreumel 2000). We call this the post-middle field or POSTMI. All other elements are in the middle-middle field or MIMI. Each of the three subfields can be empty. Examples are given in table 6.

⁸ RB-POST, PRE, LB-MI, and twice #.

PRE	LB	PREMI	MIMI	POSTMI	RB
Daar There	peins think	ik I	niet not	over! about	
Ik I	had had	't 'm nog wel it him PRT PRT	zo duidelijk so clearly		uitgelegd. explained
	omdat because	we we	het hek the fence	groen green	moesten verven had-to paint
De man The man	heeft has		de hond the dog		geslagen beaten

Table 6. Examples of a topological analysis with a refined MI field.

We reanalyzed all instances of comment clauses whose positions were encoded as LB-MI, MI, or MI-RB into the appropriate position within the middle field in both written and spoken language. Then we combined all possible positions into either positions following a certain field or positions within a certain field. For example, LB-MIMI was mapped on LB-any following field. Finally, we represented the distribution around the middle field in terms of the proportion of occurrences relative to the total number of occurrences around the middle field, so we did not zoom in on the percentages of the total distribution. The results are given in table 7.

Position	Written		Spoken	
	#	%	#	%
LB-(PREMI/MIMI/POSTMI)	21	61.8	29	25.2
PREMI	2	5.9	20	17.4
PREMI-(MIMI/POSTMI/RB)	10	29.4	32	27.8
MIMI	1	2.9	30	26.1
MIMI-(POSTMI/RB)	0	0	4	3.5
POSTMI	0	0	0	0
POSTMI-RB	0	0	0	0
Total	34	100	115	100

Table 7. The distribution of comment clauses around MI.

We tested the significance of these results in the same way as we did with the total number of occurrences as described above. Now the value of the Likelihood Ratio test was 20.134 ($df = 5$), $p < 0.01$.

In the written data, we see a sharp decrease in frequency of use from left to right until the MIMI field, and no comment clauses are used following this field. The exception to the gradual decrease is the PREMI field, which is used less often than would fit the line of decrease. In spoken data, the positions early in the middle field are all used in roughly 25% of the cases with a slight decrease in the PREMI field, and then the frequency shows a sharp decrease. Only a few instances follow the MIMI field (in fact they all occur between MIMI and POSTMI), and no instances occur in POSTMI or between POSTMI and RB. In sum, a comment clause following MIMI is almost impossible, and in written language there is a strong preference for positions preceding MIMI.

Now that we have a clear picture of the distribution of comment clauses, we may ask why this distribution is as it is. There are probably three factors influencing the distribution: prosody, syntax, and semantics. With respect to prosody we must note that finite comment clauses are several syllables long and that they are set apart from the intonation pattern of the clause. The intonation stops when the finite comment clause begins. This clause then has its own pattern, and after the finite comment clause the intonation of the host clause continues where it had stopped. It can be expected that this intonation pattern becomes difficult when the preceding or following part of the host clause is only one or a few syllables long. Furthermore, the PREMI field is a difficult position from a prosodic viewpoint, since it often contains clitics and particles that are intonationally strongly bound together. By contrast, a position that has a pause in the intonation, such as the position between clauses or following/preceding the left and right dislocation field, is very suitable for interruption.

Syntactically we see that elements within POSTMI are often closely related to the verb; they are often nonverbal parts of verbal expressions, predicates, and resultatives. Apparently, the fact that comment clauses rarely occur in POSTMI, between POSTMI and RB, or within RB indicates a strong relationship between the verbs and their nearest complements that cannot be disrupted. The semantic factor of interest might be the tendency of new information to occur further to the right in the clause. It could be the case that new information cannot be interrupted by parenthetical material, although further research is necessary to find out whether this is indeed the case.

Another point observed is that comment clauses rarely occur within a major constituent. In fact, our corpus instances do not occur within NPs, APs, or ADVPs, and hardly ever within PPs. The position preferences may be explained by adding some verbal projection to this list. Note that MIMI, POSTMI, and RB together form VP (Van Zonneveld 1994) or some functional projection like IP (Sybesma 2002) in a generative framework. Apparently, the coherence of major constituent-XPs and VPs is strong enough to prevent interruptions, but gets weaker as higher levels in the sentence are reached. There are no comment clauses in NPs, while there are a few comment clauses in PPs (between the preposition and its complement), and a few more in MIMI. If this assumption is correct, we would expect that as we move higher in a generative syntactic tree (which equals moving more to the left in a topological analysis) the more comment clauses we find. This expectation does not come true in table 5. The non-occurrence of comment clauses *within* the fields LD, PRE, and LB can be explained by the fact that these fields usually contain only one constituent (LB even contains only one word), and as we saw above major constituent-XPs are hard to penetrate. The reason that relatively few comment clauses appear *between* these first sentence fields may have to do with the scope of the comment clause. That is, the comment clause usually modifies not the known information in a discourse but the new information as provided by this sentence, and it is often the VP that contains the new information. Therefore, positions preceding or following the VP may be a more appropriate alternative to the syntactically most straightforward clause boundaries than positions preceding or following elements with a low information value, such as topics or modals. Another explanation could be that the LD field and the PRE field are simply less often used than the other fields, and consequently finite comment clauses have fewer opportunities to occur in these positions. Again, of course, this is an area that requires more research.

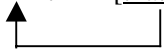
An analysis as sketched above should hold for both written and spoken language, but table 5 shows differences between the distribution of comment clauses in written versus spoken language. These differences mainly pertain to the order of the preferred positions. In written data, # is most often used, followed first by LB-MI, and then by MI. By contrast, in spoken data MI is the favorite position, followed by # and LB-MI. In other words, while MI moves from third preference to first preference, the order of the other preferred positions remains the same. A tentative explanation is that the restrictions on XP-interruption are less strong in

spoken language. It is a well-known fact that spoken language offers more freedom than written language in many respects. Another explanation might be that MI-internal comment clauses tend to comment on an MI-internal XP, whereas the other positions tend to comment on the entire VP or CP. Commenting on only a part of the message could be more frequent in spoken language than in written language, as it is uttered at the moment when the speaker realizes that his utterance needs modification. In written language, self-comment is more controlled since it is often a deliberate warning that the message is not a fact but merely an opinion. This hypothesis could be the subject of further research, although we must be aware that it will be hard to develop objective criteria for the question as to what exactly the comment clause is commenting on, especially in spoken language. Nevertheless, the differences in distribution of comment clauses in written and spoken language seem either negligible or explicable.

3. Analyses of Finite Comment Clauses.

The analyses of finite comment clauses that have been given in the literature can be divided into two groups, parenthetical analyses and extraction analyses. A parenthetical analysis implies that an independent chunk is inserted into a matrix clause to which it bears no syntactic relationship. By contrast, an extraction analysis assumes that the comment clause is in fact the main clause. The surrounding clause is the direct object of the verb in the comment clause, which is discontinuous because one or more of its parts were extracted out of it. A rough example is provided in 7.

(7) I think [that idea is stupid]_{DO}



That idea, I think, is stupid.

The parenthetical analysis has been defended by Reis (1996) for German and by Corver and Thiersch (2002) for Dutch. The extraction analysis has been argued for by Ross (1972), McCawley (1982) and Emonds (1973) for English, by Grewendorf (1988) and Staudacher (1990) for German, and again by Corver and Thiersch (2002) for Dutch. (Corver and Thiersch split up the group of comment clauses into two subgroups and give each of them a different analysis.)

It is remarkable that almost all these authors restrict themselves to comment clauses occurring in one position. Grewendorf and Staudacher discuss only comment clauses in prefinite position (the position PRE-LB in a topological framework) in German, and for English only the sentence-final and sentence-prefinal positions are discussed. Reis as well as Corver and Thiersch are the only ones who take comment clauses at several positions into account. As our corpus study shows, comment clauses can indeed occur in almost all positions in Dutch, although they have a strong preference for postfinite and clause peripheral positions. This does not confirm an analysis that allows only one position.

From our corpus study it appears that the comment clause in Dutch certainly does not occur everywhere. In fact, only a few positions (#, LB-MI, and the first positions within the middle field) are strongly preferred. The other positions used are between topological fields. Positions within major constituents and the position MI-RB are rarely used. In Schelfhout et al. 2003b, we compared these positions with the positions in which interjections preferably occur. It turned out that interjections demonstrate a strong preference for the position between clauses, but the preferred positions after this strong preference are comparable to those of finite comment clauses. In particular, we may note that interjections rarely interrupt a major constituent.

In earlier research (Schelfhout 1999), the position of reporting clauses in Dutch was described in roughly the same framework as was used for the present paper. The conclusion was that reporting clauses mainly occur between topological fields, but not between the middle field and the second verbal field. Schelfhout (2000) and Collins and Branigan (1997) argue for a parenthetical analysis of reporting clauses, and for interjections there is no alternative analysis to our knowledge. The similarities in the distribution of interjections, reporting clauses, and finite comment clauses therefore strongly suggest that a parenthetical analysis is applicable to comment clauses as well. Or, to put it differently, at present only parenthetical analyses are able to explain the distribution of comment clauses.

4. Discussion.

An overview of the distribution of finite comment clauses in Dutch and the analyses of finite comment clauses that have been put forward in the literature made clear that the distribution cannot be explained by the extraction analyses to date, but it can be explained by a parenthetical analysis. Accordingly, there are two possible approaches: we can either

adapt the extraction analysis or accept the parenthetical analysis. If an extraction analysis is to cover all instances as found in the corpus, it has to allow multiple extraction. Several elements must be moved out of the complement clause in several cycles to be raised to a position after the element that was raised earlier. However, there seems to be no theoretical basis for this type of approach.

At the same time, a parenthetical analysis also has its problems. The main question raised by a parenthetical analysis is why the parenthetical clause can be incomplete in itself. Usually parenthetical clauses are complete clauses, but in comment clauses the direct object role seems to be empty. It is I THINK, not I THINK SOMETHING. An extraction analysis does not have this problem, as the matrix clause bears the direct object role, but how does a parenthetical analysis deal with the apparent absence of an obligatory argument role?

The proposed solution is developed along the lines of the analysis of reporting clauses found in Schelfhout 2000. This paper follows the analysis of reporting clauses in English developed by Collins and Branigan (1997), which states that reporting clauses are parenthetically attached to the citation by use of an operator. This operator can optionally surface as the particle *so*, which always takes the first position in a reporting clause. This also explains the inversion in the reporting clause. Schelfhout (2000) notes that a number of reporting clauses gathered by corpus research was indeed introduced by the Dutch particle *zo* 'so'. The following test was conducted: in all clauses introduced by the particle *zo* it was left out, while in all clauses that were not introduced by the particle *zo* it was added in the first clause position. This did not change either the grammatical acceptability or the meaning of the clauses. Apparently, Dutch reporting clauses are comparable to English ones in this respect. There is an operator at the first clause position that might be phonologically empty but can be made explicit in the form of the particle *so/zo*. It is this operator that somehow absorbs or takes on the direct object role.

This operator can also appear when the reporting clause or the finite comment clause occurs sentence finally, as in examples 8 and 9.

- (8) a. “Dat is erg belangrijk
that is very important
voor de ontwikkeling van deze theorie,” (ZO) ZEI HIJ.
for the development of this theory so said he
‘‘That is very important for the development of this theory,’’ (so)
he said.’
- b. Dat is erg belangrijk
that is very important
voor de ontwikkeling van deze theorie, ALTHANS, DAT ZEI HIJ.
for the development of this theory at-least that said he
‘That is very important for the development of this theory, or at
least, that’s what he said.’
- (9) a. Dat is erg belangrijk
that is very important
voor de ontwikkeling van deze theorie, (ZO) DENK IK.
for the development of this theory so think I
‘That is very important for the development of this theory, I think.’
- b. Dat is erg belangrijk
that is very important
voor de ontwikkeling van deze theorie, ALTHANS, DAT DENK IK.
for the development of this theory at-least that think I
‘That is very important for the development of this theory, or at
least, that’s what I think.’

The difference between the a and b examples illustrates that the operator *zo*, whether phonologically present or not, allows the direct object role to remain empty whereas this role must be fulfilled when the word *althans* ‘or at least’ enforces a new clause.

The same analysis seems to be applicable to finite comment clauses. When we apply the test described above to finite comment clauses, the same results are obtained. The operator *zo* can be present or phonologically empty without consequences for the syntactic acceptability or the meaning of the comment clause. Another similarity, as

discussed earlier, is the distribution of reporting clauses and finite comment clauses. Therefore, we conclude that the same analysis holds and that the main objection to a parenthetical analysis for finite comment clauses is sufficiently refuted.

Finally, two caveats are in order about our methodology. First, the fact that certain positions do not occur in a corpus does not prove that these positions can never be used. More research on those positions is necessary. However, the earlier study of reporting clauses (Schelfhout 1999) confirms that these constructions are truly rare in positions MI-RB and RB. A second point relates to the preferred positions. These preferences are now based on the absolute distribution figures, although these figures can only be indicative of preference if we assume that all fields and positions are available with equal frequency. Of course, this assumption is not necessarily correct. In fact, it is rather likely that some fields are used more frequently than others. Thus, the left and right dislocation fields are only occupied with what are considered marked structures. Moreover, PRE can be expected to be less frequent than LB and MI, because of its absence in embedded clauses. When this fact is taken into account, the PRE and PRE-LB positions might receive higher peaks relative to LB-MI. Unfortunately, no Dutch corpus is available that is annotated according to the topological model as described in the ANS. Accordingly, we do not have any figures about the relative use of fields, and absolute distribution figures of intercalations are the best we can attain at present.

5. Conclusion.

We have presented a corpus-based investigation of the distribution of finite comment clauses. In both written and spoken language it appeared that comment clauses can occur between most topological fields (except the MI-RB position), but they have a strong preference for clause boundaries or the positions following LB. This distribution is unexpected under the types of extraction analyses presented by several investigators, but is consistent with a parenthetical analysis. Under a parenthetical analysis, however, it has to be explained why the direct object role of the verb in the comment clause can be empty. In analogy to reporting clauses, the explanation is found in an operator that might be phonologically filled or empty. When filled it always occupies the first position in the comment clause and has the form of the particle *zo*.

REFERENCES

- Collins, Chris, and Phil Branigan. 1997. Quotative inversion. *Natural Language and Linguistic Theory* 15.1–41.
- Corver, Norbert, and Craig Thiersch. 2002. Remarks on parentheticals. *Progress in grammar*, ed. by Marc van Oostendorp and Elena Anagnostopoulou. Available at <http://meertens.library.uu.nl/progressingrammar/corver.pdf>.
- Van Dreumel, Simon. 2000. The Amazon grammar and the last part of the middle field. *Computational linguistics in the Netherlands 1998. Selected papers from the ninth CLIN meeting*, ed. by Frank Van Eynde, Ineke Schuurman and Ness Schelkens, 93–107. Amsterdam: Rodopi.
- Emonds, Joseph. 1973. Parenthetical clauses. You take the high node and I'll take the low node. *Papers from the comparative syntax festival. The differences between main and subordinate clauses, 12 April 1973. A paravolume to papers from the ninth regional meeting*, ed. by Claudia Corum, T. Cedric Smith-Stark, and An Weiser, 333–347. Chicago: Chicago Linguistic Society.
- Gerrits, Anouk. 2001. *Het begin van het middenveld*. Master's thesis, University of Nijmegen.
- Grewendorf, Günther. 1988. Aspekte der deutschen Syntax: Eine Rektions-Bindungs-Analyse. (*Studien zur deutschen Grammatik*, 33.) Tübingen: Gunter Narr Verlag.
- Haeseryn, Walter. 1998. Achteropplaatsing van elementen in de zin. Nederlands 200 jaar later. *Handelingen dertiende colloquium Neerlandicum*, ed. by Hugo Brems, 303–326. Woubrugge: IVN.
- Haeseryn, Walter, K. Romijn, G. Geerts, J. de Rooij, and M. C. van den Toorn (eds.). 1997. *Algemene Nederlandse Spraakkunst*. Groningen/Deurne: Martinus Nijhoff/Wolters Plantyn.
- McCawley, James D. 1982. Parentheticals and discontinuous constituent structure. *Linguistic Inquiry* 13.91–106.
- Oostdijk, Nelleke. 2000. The Spoken Dutch corpus: Outline and first evaluation. *Proceedings of the second international conference on language resources and evaluation (LREC)*, ed. by M. Gravididou, G. Caravannis, S. Markantonatou, S. Piperidis et al., 887–894. Athens: ELDA.
- Reis, Marga. 1996. Extractions from verb-second clauses in German? *On extraction and extraposition in German*, ed. by Uli Lutz and Jürgen Pafel, 45–88. Amsterdam: John Benjamins.
- Ross, John. 1972. Slifting. *The formal analysis of natural languages. Proceedings of the first international conference*, ed. by Maurice Gross, Morris Halle, and Marcel-Pau Schutzenberger, 133–169. The Hague: Mouton.
- Schelfhout, Carla. 1999. DIP-constructies in AMAZON: *Een onderzoek naar plaats en vorm van de reporting clause in parenthetische directe rede-constructies*. Master's thesis, University of Nijmegen.

- Schelfhout, Carla. 2000. Corpus-based analysis of parenthetical reporting clauses. *Computational linguistics in the Netherlands 1998. Selected papers from the ninth CLIN meeting*, ed. by Frank Van Eynde, Ineke Schuurman, and Ness Schelkens, 147–159. Amsterdam: Rodopi.
- Schelfhout, Carla, Peter-Arno Coppen, and Nelleke Oostdijk. 2003a. Intercalaties? Dat zijn geloof ik van die tussendingen ... *Gramma/TTT* 10.27–44.
- Schelfhout, Carla, Peter-Arno Coppen, and Nelleke Oostdijk. 2003b. Positions of parentheticals and interjections: A corpus-based approach. *Linguistics in the Netherlands 2003*, ed. by Leonie Cornips and Paula Fikkert, 155–166. Amsterdam: John Benjamins.
- Schelfhout, Carla, Peter-Arno Coppen, and Nelleke Oostdijk. To appear. Transparent free relatives. *Proceedings of CONSOLE XII*.
- Staudacher, Peter. 1990. Long movement from verb-second-complements in German. *Scrambling and barriers*, ed. by Gunther Grewendorf and Wolfgang Sternefeld, 319–339. Amsterdam: John Benjamins.
- Sybesma, Rint. 2002. *Syntaxis: een generatieve inleiding*. Bussum: Coutinho.
- Wilder, Chris. 1999. Transparent free relatives. *Proceedings of the seventeenth West Coast Conference on Formal Linguistics*, ed. by K. N. Shahin, S. Blake and E-S. Kim, 685–699. Cambridge: Cambridge University Press. (Also published in *ZAS Papers in Linguistics* [1998] 10.191–199.)
- Van Zonneveld, Ron. 1994. *Kleine syntaxis van het Nederlands*. Dordrecht: ICG Publications.

Radboud University Nijmegen
 Postbox 9103
 6500 HD Nijmegen
 The Netherlands
 [C.Schelfhout@let.ru.nl]