Positions of parentheticals and interjections

A corpus-based approach

Carla Schelfhout, Peter-Arno Coppen and Nelleke Oostdijk
University of Nijmegen

1. Intercalations and their positions

Sentences in Dutch can be interrupted in various ways. We will use the term *intercalation* for interruptions by the speaker/writer that occur within the boundaries of a sentence without influencing this sentence (cf. de Groot 1949). The sentence (the *host sentence* or *host*) continues after the intercalation as if the intercalation were not there; the syntactic structure and the intonation pattern of the host sentence remain unaltered. This description of intercalations can be summarized in the following formula:

> ⟨I⟩ is an intercalation iff ⟨AIB⟩ is such that ⟨A⟩, ⟨I⟩ and ⟨B⟩ are not empty and both ⟨AB⟩ and ⟨AIB⟩ are Dutch sentences in which the intonation pattern and syntactic structure of ⟨AB⟩ are the same.

This definition covers a broad range of constructions, from appositives and conjunction reductions to interjections and reporting clauses (see Schelfhout et al. to appear). In this paper, we will focus on one group of intercalations, viz. those intercalations that can move about freely in the sentence and are hence called *free intercalations*. They include reporting clauses, vocatives, comment clauses, interjections, and parentheticals. These last four types of structure can be exemplified as variations (1b–e) on the reporting clause in sentence (1a):[1]

(1) a. *"Ik geloof," zei Annie, "dat je wat moet uitleggen."*
I believe said Annie that you something have.to explain
'"I think," Annie said, "that you ought to explain something."'

b. *Ik geloof John, dat je wat moet uitleggen.*
I believe John that you something have.to explain
'I think, John, that you ought to explain something.'

Linguistics in the Netherlands 2003, **155–166.**
ISSN 0929–7332 / E-ISSN 1569–9919

c. *Ik geloof — maar wie maalt er om mijn mening? — dat je*
I believe but who cares THERE about my opinion that you
wat …
something
'I think — but who cares about my opinion? — that you ought to explain something.'

d. *Ik geloof verdorie dat je wat moet uitleggen.*
I believe damn that you something have.to explain
'I think damn that you ought to explain something.'

e. *Ik geloof, vrees ik, dat je wat moet uitleggen.*
I believe fear I that you something have.to explain
'I think, I'm afraid, that you ought to explain something.'

In written material, free intercalations are commonly surrounded by stylistic markers such as parentheses, dashes, commas, or quotes.

In the literature not much is said about the positions at which free intercalations can occur. The tacit assumption seems to be that they can occur everywhere. If this assumption is true, we should expect an equal distribution of free intercalations over all available positions or at least a similar distribution for all free intercalations. In order to test these hypotheses, we decided to carry out a corpus-based study. We compiled a corpus that comprises both written and spoken language data. The corpus was used to obtain information about the distribution of intercalations. For practical reasons, we restricted ourselves to two types of intercalation: parentheticals and interjections, as exemplified in (1e) and (1d). Parentheticals were included in the study because they can be structurally rather complex and therefore — from a syntactic point of view — are interesting to study. However, parentheticals are not all that frequent. Therefore, we decided to include the interjections: these are items with a high frequency and including them will make it possible to draw reliable conclusions.

2. Parentheticals and interjections

Parentheticals can be divided into two groups: those containing a verb that expresses the subject's opinion and those containing a copula.[2] In both groups, the finite verb can be preceded by the particle *zo* (so): wherever *zo* is present, it can be left out, and wherever *zo* is not present it can be added without affecting the grammatical acceptability or changing the meaning of the parenthetical. Parentheticals of the first group are very similar to reporting clauses (ex. 1a; see e.g. Collins and Branigan 1997, and Schelfhout 1999). Examples of parentheticals that use an opinion-expressing verb or a copula are (2) and (3), respectively.

(2) *uh dat is vind ik gewoon heel mooi.*
uh that is think I just very beautiful
'That is just very beautiful, I think.'

(3) *'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.'

In the literature, the nature of parentheticals is subject to debate: are they parenthetic constructions or are they main sentences, the direct object of which is realized by a sentence that has been split up as a result of extraction or scrambling? An overview of the discussion is given in Schelfhout et al. (2003). In this same article, results are presented that support the analysis as parenthetic constructions. In the present paper, we will therefore assume this analysis.

Interjections are a rest category. This explains why the definition of interjections is rather broad and vague. We have roughly followed the description as developed in the authoritative Dutch grammar *Algemene Nederlandse Spraakkunst* (ANS; Haeseryn et al. 1997). We have excluded those interjections that imitate a sound, like *ratatatata* or *kukeleku*. These interjections *do* affect the host sentence, and therefore do not qualify as intercalations.

3. Methodology

3.1 The corpus

The corpus is used to study intercalations in actual language use. By comparing data from the spoken component with data from the written component, we intend to find out whether there is any difference in the use of intercalations that relates to the mode of speech. We expect that in spoken language intercalations will be more frequent and show a wider distribution than in written language. The written component in our corpus is constituted by some one million words with their origin in print. The spoken component is derived from the Spoken Dutch Corpus (*Corpus Gesproken Nederlands* or CGN; see Oostdijk 2000). We started with roughly 175,000 words spoken data, but as the parentheticals turned out to be infrequent indeed, we felt a need to extend the spoken part of the corpus to roughly 460,000 words in order to obtain enough instances for a reliable analysis. The extension was not used for retrieving more interjections, as we had already found enough instances of interjections in the original spoken material. The composition of the corpus is displayed in Table 1.

First, all material was automatically tagged for part-of-speech information by means of the CGN tagger.[3] We then semi-automatically retrieved all parentheticals and interjections. The results are summarized in Table 2. Apart from the absolute

Table 1. Corpus composition

WRITTEN		SPOKEN interjections		parentheticals
essay	127,122	lecture	62,810	62,810
interview	126,376	interview	7,101	62,510
news	127,578	news	36,143	80,121
novel	255,503	commentary	4,209	125,747
short story	255,653	private conversation	63,883	63,883
scientific writing	127,441	telephone conversation	0	63,205
Total	1,019,673	Total	174,146	458,276

Table 2. Instances of parentheticals and interjections found in the corpus

	written	spoken	overall
parentheticals	76	195	271
parentheticals per 10,000 words	0.7	4.3	1.8
interjections	159	496	655
interjections per 10,000 words	1.6	28.5	5.5

number of instances found in the data, we also include the standardized frequencies, i.e. the number of instances per 10,000 words.

These findings indeed confirm that interjections are more frequent than parentheticals while free intercalations (at least parentheticals and interjections) are more frequent in spoken language than in written language.

3.2 Descriptive model

Our descriptive analysis of the data is based on the ANS. Clause structure is described in terms of structuralistic fields, which are defined in terms of their relative position to the verbal elements that can occur in a clause. There are two verbal poles in Dutch clauses. The first verbal pole contains the finite verb in a main clause and the subordinator in a subordinate clause; the second verbal pole contains all non-finite verbs of the main clause and all verbs in the subordinate clause. The following positions are distinguished:

– LD: the left dislocation field. This is the position for left-dislocated elements.
– TOP: the topicalisation field, which is the canonical position for subjects and topicalised elements.

– P1: the first verbal pole, which contains the finite verb in main clauses. In subordinate clauses, this is the position for the subordinator.
– MI: the middle field, the part of the clause between the two verbal fields.
– P2: the second verbal pole or verbal cluster; all non-finite verbal elements in a main clause and all verbal elements in a subordinate clause occur here.
– EX: the extraposition field. It is used for extraposed elements like postposed PPs.
– RD: the right dislocation field. This is the position for right-dislocated elements.

In a clause, each of these fields can be empty. In Table 3 below, some examples are displayed.

If we encode the positions of parentheticals in a clause within this framework, we have 12 logical possibilities: within each of the seven fields (indicated by 'name of field'), except the P1 field that by definition can only contain a single word, or between any two adjacent fields (indicated by 'name of field to the left'- 'name of field to the right').[4] In Table 4, the positions LD-TOP and MI are exemplified.[5] There is one further possibility: it is possible for a parenthetical to occur between two (coordinate or subordinate) clauses.[6] This position is indicated by #.

If the hypothesis that intercalations can occur anywhere in a sentence is correct, we should expect a similar distribution of parentheticals and interjections over all 13 positions.

Table 3. Examples of a structuralistic analysis

LD	TOP	P1	MI	P2	EX	RD
		heb *have*	jij Marie *you Mary*	gezien? *seen*		
	Ik *I*	heb *have*	de man *the man*	gezien *seen*	met de hond *with the dog*	
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 4. Parentheticals at the positions LD-TOP and MI

LD		TOP	P1	MI	P2
Elvis *Elvis*	meen ik *think I*	dat *that*	is *is*	mijn favoriet *my favourite*	
			omdat *as*	ik hem meen ik ergens van *I him think I somewhere of*	ken *know*

4. Results

The distribution of parentheticals and interjections in written and spoken data is shown in Tables 5 and 6 respectively.[7] The results show that parentheticals and interjections occur most frequently between fields ($p<0.01$ for TOP-P1, P1-MI and #, in spoken language data also P2-EX), but in spoken language data occurrences within a field occur significantly more frequently as well ($p<0.01$ for MI, EX and RD). Intercalations within a field rarely occur within a major constituent in this field (and if so, only in spoken language). They prefer positions between major constituents.

5. Discussion

When we look at the results, the first thing to note is that some positions are hardly ever used: there are very few intercalations in LD, MI-P2, P2 and EX-RD in the corpus and intercalations rarely occur within major constituents. In the case of LD and EX-RD we cannot draw any conclusions from this observation, as these fields are hardly ever occupied. Not many sentences contain a left-dislocated element or contain both an extraposed and a right-dislocated element, which implies that intercalations cannot occur here either. In the case of MI-P2 and P2, Dutch language users seem to find it difficult to place an intercalation at these positions. P2 may be empty, or it may be occupied by a main verb or one or more modals and a verb. An intercalation within P2 can only occur in the last case. Apparently the cohesion of modals with their main verbs is strong enough to prevent interruption. This, however, cannot explain the negligible use of the MI-P2 position. Another factor here might be that elements occurring more to the right in a sentence tend to carry more information.[8] This makes an interruption inappropriate, as it would detract the listener's attention from the main message. The rare use of positions within major constituents suggests, that XPs in general are resistant to penetration by intercalations.

When it comes to the positions that *are* used, we find that there is by no means an equal distribution of intercalations over the available positions. In written language, interjections strongly prefer the position between clauses. Parentheticals have the same preference for the # position but an almost similar preference for the P1-MI position. Other positions are available, but their actual use is negligible. In spoken language, the positions within MI join # and P1-MI as preferred positions; the MI-position is even the first preference for parentheticals.

While at this stage an explanation of the different distribution of parentheticals and interjections can only be highly speculative, we want to offer the following explanation for consideration: one possible explanation for the different distribution may be found in the fact that parentheticals and interjections serve different functions. Interjections are often used to keep the ground while speaking, and of

Table 5. Distribution of parentheticals and interjections in written language data

position	parentheticals		interjections	
	#	%	#	%
LD	0	0	0	0
LD-TOP	5	6.6	12	8.1
TOP	3	3.9	3	2
TOP-P1	**7**	**9.2**	**0**	**0**
P1-MI	**21**	**27.6**	14	9.4
MI	11	14.5	11	7.4
MI-P2	2	2.6	1	0.7
P2	0	0	0	0
P2-EX	1	1.3	0	0
EX	0	0	1	0.7
EX-RD	0	0	0	0
RD	1	1.3	0	0
#	**25**	**32.9**	**107**	**71.8**
Total	76	100	149	100

Table 6. Distribution of parentheticals and interjections in spoken language data

position	parentheticals		interjections	
	#	%	#	%
LD	0	0	3	0.6
LD-TOP	**1**	**0.5**	20	4.1
TOP	5	2.7	16	3.3
TOP-P1	**8**	**4.3**	9	1.8
P1-MI	**29**	**15.6**	**6**	**1.2**
MI	**84**	**45.2**	**52**	**10.7**
MI-P2	2	1.1	6	1.2
P2	0	0	0	0
P2-EX	**8**	**4.3**	3	0.6
EX	**7**	**3.8**	6	1.2
EX-RD	0	0	0	0
RD	**3**	**1.6**	1	0.2
#	**39**	**21**	**365**	**74.9**
Total	186	100	487	100

course the moment following a complete utterance is more prone to interruption by the interlocutor than a moment when the speaker obviously has not yet finished; therefore, making a sound to keep the turn while thinking is more useful between sentences than within them. An example is given in (4):

(4) *Ik kon wel janken verdomme, maar hield me groot.*
I could PRT cry damn but held me big
'I felt like crying, damn, but I kept up appearances.'

Parentheticals serve a different purpose: speakers use parentheticals to warn the listeners that they are not communicating a fact, but only an opinion. This is corroborated by the fact that of all copulas, only the ones with a modal meaning aspect can be used in a parenthetical. A suitable position for such meta-statements might be the position immediately preceding the main message, as exemplified in (5).[9]

(5) *nou dat was denk ik twee jaar geleden of zo.*
well that was think I two years ago or so
'Well, that was two years ago or something, I think.'

If our assumption that parentheticals prefer the position before the main message is correct, we should expect that they will prefer the positions at the beginning of the middle field, as it is generally assumed that the further you proceed into the middle field, the higher the information value of the elements is. In order to find out whether this holds true, we divide the middle field into three subfields by splitting off the peripheral elements, which can easily be identified. The first part of the middle field is the canonical position for subjects, clitics and particles (cf. Gerrits 2001); we will 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 (cf. Van Dreumel 2000); we will 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 7.

We reanalysed all instances of parentheticals that were originally encoded as P1-MI, MI or MI-CL. Then we combined all possible positions into either positions following a certain field or positions within a certain field. We present the distribution of parentheticals around the middle field as percentages of only instances around the middle field; we did not calculate the percentages of all instances. The results are given in Table 8; all results are significant at $p < 0.01$.

Table 7. Examples of a structuralistic analysis with a refined MI-field.

TOP	P1	PREMI	MIMI	POSTMI	P2
Daar	peins	ik	niet	over!	
there	*think*	*I*	*not*	*about*	
Ik	had	't 'm nog wel	zo duidelijk		uitgelegd.
I	*had*	*it him PRT PRT*	*so clearly*		*explained*
	omdat	we	het hek	groen	moesten verven
	because	*we*	*the fence*	*green*	*had.to paint*

Table 8. Distribution of parentheticals over the Middle Field

MI-position	written		spoken	
	#	%	#	%
P1 — (PREMI/MIMI/POSTMI)	**21**	**61.8**	**29**	**25.2**
PREMI	2	5.9	20	17.4
PREMI — (MIMI/POSTMI/P2)	10	29.4	32	27.8
MIMI	**1**	**2.9**	**30**	**26.1**
MIMI — (POSTMI/P2)	0	0	4	3.5
POSTMI	0	0	0	0
POSTMI — P2	0	0	0	0
Total	34	100	115	100

In written language, we see a gradual decrease in frequency from the beginning to the end of the middle field, as we expected. The one exception is the PREMI field, in which occurrences are relatively rare. The reason for this exception may be twofold: first, the number of sentences that fill the PREMI field is less than the number of sentences that contain a P1 and a MIMI field; and if a sentence has a PREMI field, it may contain only one element. Second, if more elements occur in the PREMI field, they often form a prosodic unit that is strongly bound to P1; this makes interruption more difficult.

In spoken language, the distribution over the first four positions is more or less equal, apart from a dip in the PREMI field; then we see an abrupt drop in frequency of use. Unlike written language, the assumption that an interruption in the middle field becomes less likely as it occurs further to the right does not hold generally in spoken language. The general conclusion is that positions following MIMI are truly more difficult to use than positions earlier in the middle field. Of course, POSTMI is not used in every sentence, so this might explain part of the frequency drop, but it does not explain all of it. The explanation might have to do with a strong relationship of elements in POSTMI with P2; especially predicates and resultatives are strongly bound to the main verb. As P2 is impenetrable as well, there is apparently a strong cohesion between the predicate and the main verb that makes an interruption very hard.

The distribution of interjections over the middle field, as depicted in Table 9, is comparable to parentheticals as far as written language is concerned, but for spoken language the frequency of use is increasing instead of decreasing right up to the point where there is the boundary following MIMI. The results are significant at $p<0.01$ for spoken material and $p<0.002$ for written material. Apparently the increase of information value through the middle field hampers interruption in written language. In spoken language, there is no such effect. The boundary

Table 9. Distribution of interjections over the Middle Field

MI-position	**written**		**spoken**	
	#	%	#	%
P1 — (PREMI/MIMI/POSTMI)	**14**	**53.8**	**6**	**9.4**
PREMI	2	7.7	1	1.6
PREMI — (MIMI/POSTMI/P2)	8	30.8	20	31.3
MIMI	**2**	**7.7**	**31**	**48.4**
MIMI — (POSTMI/P2)	**0**	**0**	**6**	**9.4**
POSTMI	0	0	0	0
POSTMI — P2	0	0	0	0
Total	26	100	64	100

following MIMI, however, seems to hold in both language types and for both intercalation types.

Finally, two caveats are in order about our methodology. First, the fact that certain positions are not used in a corpus does not prove that these positions are not ever used. Further research into those positions is necessary. Earlier research into the reporting clause (Schelfhout 1999), which is closely related to parentheticals, has already indicated that these constructions are truly rare at the positions MI-P2 and P2. A second point relates to the preferred positions. These preferences are now based on the absolute distribution figures. However, these figures can only be indicative of preference if we assume that all fields and positions are equally frequently used. This, of course, need not be the case. 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 to be marked structures. Moreover, TOP can be expected to be less frequent than P1 and MI, because of its absence in subordinate clauses. When this is taken into account, the TOP and TOP-P1 positions might receive higher peaks relative to P1-MI. Unfortunately, at present no Dutch corpus is available which is annotated according to the descriptive model employed in the ANS. Hence, we do not have any figures about the relative use of fields. The only parser that delivers surface structure analyses for Dutch sentences that conform to the ANS is the AMAZON-parser (see Coppen 2002); perhaps one day we will have our present corpus annotated by AMAZON but until then, absolute distribution figures of intercalations are the best we can get.

6. Conclusion

In the present paper we set out to investigate whether it is true that intercalations can occur everywhere in a sentence. A corpus-based study shows that this is not the case: there is a restricted set of positions where intercalations can occur, and within this set, the positions at clause boundaries and the position between P1 and the middle field are strongly preferred. Positions following or within POSTMI and P2 are almost inaccessible, possibly because of the strong coherence between the verbal elements. Within these general restrictions, two different types of intercalation — parentheticals and interjections — show different distributions, which suggests that other types may have different distribution patterns as well. One explanation for this difference may be the fact that different intercalations have different functions. Further research into form, distribution and function of other types of intercalation may bring further clarification in this matter.

Notes

1. Examples (1a), (2), (3), (4), and (5) in the text are authentic examples; the examples in the tables are made up.

2. Although there are nine copulas in Dutch, only six can occur in a parenthetical: *blijken* (to appear), *lijken* (to seem), *schijnen* (to seem), *heten* (to be thought to), *dunken* (be of the opinion; archaic), *voorkomen* (to seem). These happen to be all copulas with modal content.

3. For this we would like to thank Antal van den Bosch and Hans van Halteren.

4. If an intercalation appears between two non-adjacent fields, we have transparency: as the intervening field(s) is/are empty, we are unable to decide where exactly the intercalation occurs. These instances are excluded from the further research. This was the case for none of the 76 written parentheticals, 9 of 195 spoken parentheticals, 10 of 159 written interjections and 9 of 496 spoken interjections.

5. From here onwards non-used peripheral fields (LD, EX, RD) are not shown for reasons of space.

6. Note that the clause boundary position can only be occupied by an intercalation if another clause precedes or follows it. By definition, intercalations only occur sentence-internally, so sentence-final or sentence-initial instances of parentheticals or interjections are not included in the present research.

7. In Tables 5, 6, 8 and 9 boldface indicates that the difference is found to be significant by a chi-square test.

8. Jansen (2002) questions the dominance of this so-called Left-Right Principle.

9. The position following the utterance would be suitable as well, of course. However, as according to our definition only sentence-internal variants are intercalations, we did not look further into the sentence-final occurrences of parentheticals and interjections.

References

Collins, C. and Branigan, P. (1997) 'Quotative Inversion'. *Natural Language and Linguistic Theory* 15, 1–41.

Coppen, P.-A. (2002) 'Het geheim van de oude dame: de Nijmeegse parser AMAZON'. *Nederlandse taalkunde* 4, 312–334.

Dreumel, S. van (2000) 'The Amazon Grammar and the Last Part of the Middle Field'. In F. van Eynde, I. Schuurman and N. Schelkens, eds., *Computational Linguistics in the Netherlands 1998; Selected Papers from the Ninth CLIN Meeting.* Rodopi, Amsterdam, 93–107.

Gerrits, A. (2001) 'Het begin van het middenveld'. Master's thesis, Dept. of Language and Speech, University of Nijmegen.

Haeseryn, W., ed., (1997) *Algemene Nederlandse Spraakkunst.* Martinus Nijhoff/Wolters Plantyn, Groningen/Deurne.

Jansen, F. (2002) 'Preposed adverbial phrases in Dutch texts. Evidence for the Left-Right Principle or for Linear Modification?' In H. Broekhuis and P. Fikkert, eds., *Linguistics in the Netherlands 2002.* John Benjamins Publishing Company, Amsterdam, 97–106.

Oostdijk, N. (2000) 'Building a corpus of Spoken Dutch'. In P. Monachesi, ed., *Computational Linguistics in the Netherlands: Selected Papers from the Tenth CLIN Meeting.* Universiteit Utrecht, Utrecht, 147–159.

Schelfhout, C. (1999) 'DIP-constructies in AMAZON: Een onderzoek naar plaats en vorm van de reporting clause in parenthetische directe rede-constructies'. Master's thesis, Dept. of Language and Speech, University of Nijmegen.

Schelfhout, C., Coppen, P.-A. and Oostdijk, N. (2003) 'Finite comment clauses in Dutch: a corpus-based approach'. Ms. University of Nijmegen.

Schelfhout, C., Coppen, P.-A. and Oostdijk, N. (to appear) 'Intercalaties? Dat zijn geloof ik van die tussendingen…' *Gramma/TTT.*