Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is an author's version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/42755

Please be advised that this information was generated on 2019-06-27 and may be subject to change.

Interjections in Dutch: a corpus-based approach

Carla Schelfhout, Peter-Arno Coppen, Nelleke Oostdijk, Frans van der Slik;
Radboud University Nijmegen

Abstract

In this paper, the distribution of interjections over the clause is studied within the context of a topological framework. For this study authentic data were extracted from a corpus comprising one million words of written material and 174,000 words of spoken material. To explain this distribution we study the influence of the text type in which an interjection is found, the length of the interjection and its function. All three factors are shown to play a role. This confirms and amplifies hypotheses that hitherto had gone untested.

Samenvatting

Dit artikel bestudeert de distributie van interjecties over de zin in termen van een topologisch model. Hierbij wordt gebruik gemaakt van authentieke data, die voorkwamen in een corpus van één miljoen woorden geschreven materiaal en 174.000 woorden gesproken materiaal. Om deze distributie te verklaren, bestuderen we de invloed van het teksttype waarin de interjectie voorkomt, de lengte van de interjectie en de functie van interjecties. Alle drie factoren blijken een rol te spelen. Deze resultaten vormen een bevestiging van en aanvulling op totnogtoe onbewezen hypotheses.

1 Introduction

Interjections are undoubtedly the most-neglected word class in studies of Dutch, probably due to the prevailing view that they are peripheral to the language. The authoritative descriptive grammar of Dutch, *Algemene Nederlandse Spraakkunst* (ANS, 'General Grammar of Dutch', Haeseryn et al. 1997), for example, states that interjections are 'words that are usually or always outside the grammatical structure of the sentence' (*translation ours, CS et al.*).

The ANS divides the rather broad class of interjections into three groups. First, it makes a division between interjections that have a meaning and interjections that merely imitate a sound, like *miauw* 'meow', *wam* 'wham', *tsjoeketsjoek* 'choo choo'. Interjections with a meaning are subdivided into those that necessarily express emotion, like swearwords, *au* 'ouch' or *ocharme* 'oh, poor him', and those that do not. This group is further divided into announcements like *foei* 'shame on you' and *ja* 'yes', orders like *ho* 'stop' and *toe (nou)* 'come on', questions like *hè* 'right' and *nietwaar* 'isn't it', and social formulas like *goedemorgen* 'good morning', *pardon* 'pardon me' and *proost* 'cheers'.

Although interjections are assumed to be outside the grammatical structure of the sentence, they occur linearly within this grammatical structure of the sentence. The few studies of interjections, such as Van den Toorn (1968), Brummel (1978), Haegeman (1984), De Vriendt (1992) and Romijn (1998), assume that they can occur

almost anywhere, though probably not within major constituents.[1] However, these studies are either restricted to a few specific interjections or based on intuition, rather than on authentic data.

Is their assumption correct? And if interjections can indeed occur in any structural position, are they evenly distributed over these positions, or are certain positions preferred over others? Are there any differences in this respect between the text types in which interjections are used? Do various types of interjection (various in length, word or function) show different distributions? The answers to these questions are relevant both for the descriptive study of Dutch and for a comparison of interjections to other constructions which also are outside the grammatical structure of the sentence but occur linearly within the sentence, as for instance vocatives or disjunct commenting clauses.

We have carried out a corpus-based study into the distribution of interjections to answer these questions. The distribution was studied at the level of the clause, rather than the utterance. This study was based on a corpus which contains both written and spoken material. For practical reasons, we considered only interjections within the boundaries of the utterance; interjections in utterance-initial or utterance-final position were excluded. We will show (1) that interjections are not evenly distributed over the clause, and (2) how the length of the interjections and the text type in which they occur influence their distribution. Finally we compared interjections with another type of interrupting constructions, parentheticals, to determine whether function influences their distribution.

2 The corpus material

There are several considerations for the design of a corpus for a study of the distribution of interjections. The corpus must be large enough to provide sufficient material for a statistically reliable analysis, yet small enough to keep its compilation and analysis feasible. It must comprise enough different text types to be representative of more than one language variety and, of course, it must be suitable to check our hypotheses.

We expect to find differences in the use of interjections in spoken and written language. In particular, we expect interjections in spoken language to be more frequent, show a wider distribution and greater variation in types than in written language. For written language, we expect a difference between text types which somehow reflect spoken language (interviews, the dialogues in fiction texts) and text types which do not. In spoken language, it is conceivable that differences appear

[1] Except Van den Toorn (1968), who claims that interjections almost always precede or follow the clause.

between public, often prepared text types like lectures, and private, spontaneous text types like private conversations. In spoken language, also differences between monologues and dialogues can be expected. On the basis of these considerations, we compiled a corpus with the following design:[2]

Table 1: The corpus design

Written text type	**Number of words**	**Spoken text type**	**Number of words**
Novel	255,503	Lecture	62,810
Short story	255,653	News	36,143
Interview	126,376	Sports commentary	4,209
News	123,140	Interview	7,101
Essay	127,122	Private conversation	63,883
Scientific writing	125,846		
Total	1,013,640	Total	174,146

The written corpus was taken from the Internet. The linguistic quality of material found on the Internet varies greatly, but the text types selected for the corpus design usually comprise carefully written material. The spoken corpus was taken from the Spoken Dutch Corpus (*Corpus Gesproken Nederlands,* CGN; Oostdijk 2000). The written corpus was annotated by means of the same tagger that was used for the CGN, so that both parts had comparable annotation. All utterances containing at least one interjection were extracted from the tagged corpus with an automatic search program. Instances in utterance-initial or utterance-final position were discarded. If an utterance contained more than one interjection, each instance was analyzed separately, so that the total number of instances is greater than the total number of utterances.[3] As it is very difficult to decide objectively when two interjections are used adjacent to each other and when there is one multiword interjection, we consider a string of interjections, like 'ja ja' *yes yes*, to be one interjection in all cases. In all, there are 939 instances, distributed over the text types as follows:

[2] In fact, this design was an intermediate stage in the compilation of a corpus with a spoken component of 500,000 words, carefully divided over the spoken text types. However, at the moment that the research was carried out, the CGN had not yet been completed, so that the number of words available for each text type differs. As there is still a balance between prepared material on the one hand and spontaneous material on the other hand, we feel this should not bias our results.

[3] For practical reasons, *uh* was not regarded as interjection in the spoken material.

Table 2: Distribution of instances over the text types with a normalization per hundred thousand words

	text type	**# interjections**	**# interjections per 100,000 words**
Written material	Novel	56	22
	Short story	122	48
	Interview	31	25
	News	1	1
	Essay	13	10
	Scientific writing	1	1
	Written total	*224*	*22*
Spoken material	Lecture	130	207
	News	3	8
	Sports commentary	11	261
	Interview	70	986
	Private conversation	501	784
	Spoken total	*715*	*411*
	Total	939	79

As expected, interjections occur more frequently in spoken than in written language. In written language, they occur more often in text types which reflect spoken language, viz. novels, short stories and interviews. In spoken language, they occur less often in monologues (lectures, news, sports commentaries) than in dialogues (interviews, private conversations). This all agrees with the expectations. The total number of different interjections (types) is 86, 49 of which are used only once (*hapax legomena*). The most frequent interjections are listed in Table 3.

Table 3: Interjections occurring 10 times or more in the corpus

Interjection	**Translation**	**Frequency**
ja	*yes*	490
hè	*right*	101
nee	*no*	69
hoor	*really*	51
nou ja	*well*	22
nou	*well*	18
verdomme	*damn*	15

ach	*oh well*	12

In order to study the distribution of interjections over the clause, we need an analysis model for the clause. This will be presented in the next section.

3 Analysis model

To investigate the positions at which interjections occur, we follow the standard topological model of Dutch traditional grammar as described in the ANS. This model is relatively theory-neutral, so that the results can be transposed to several types of syntactic analysis. In this model, the clause is organized around its verbal positions, V1 and V2, with the middle field MI in between. Topicalized elements appear in TOP, extraposed elements are placed in EX. Dislocated elements occur either clause-initially (LD) or clause-finally (RD).

- LD: the Left Dislocation field. This is the position for left-dislocated elements.
- TOP: the topicalization field, which is the canonical position for subjects and topicalized elements.
- V1: the first verb field. In main clauses this field contains the finite verb, in subordinate clauses, the subordinator.
- MI: the middle field, between the two verbal fields.
- V2: the verbal cluster field; it contains all non-finite verbal elements in a main clause and all verbal elements in a subordinate clause.
- EX: the extraposition field, for extraposed elements.
- RD: the Right Dislocation field, for right-dislocated elements.

As defined, only the middle and extraposition fields can contain more than one major constituent, i.e. clause-level constituents. These major constituents can be composed of minor constituents, for instance an adverbial PP is composed of a preposition and a noun phrase, but it is the PP as a whole which fulfills a role at clause level. Table 4 illustrates the topological analysis.

Table 4: Examples of analyses of clauses in a topological model

LD	TOP	V1	MI	V2	EX	RD
Jan, John	die him	kan can	ik wel I PRT	schieten! shoot		
	Ik I	heb have	die man that man	gezien seen	met de hoed with the hat	

		dat that	de zon the sun	kan schijnen can shine		
		Verdwijn! Get_out				
	Gisteren yesterday	is has	hij directeur he manager	geworden, become		de slijmbal. the toady

An interjection can occur either within a certain topological field,[4] e.g. within the middle field, or between two topological fields, e.g. between the first verb field and the middle field. If a certain topological field is not occupied, the position of an adjacent interjection is unclear. These instances are encoded as TRANSPARENT. In Table 5 some examples are given.[5]

Table 5: Examples of interjections in two topological positions and in a transparent position

Position	**TOP**	**V1**		**MI**		**V2**
MI	Jan John	kan could		ik VERDORIE wel I darn PRT		schieten! shoot
V1-MI	Ik I	heb have	VERDORIE darn	de man met de hond the man with the dog		gezien. seen
TRANSP ARENT	Hij he	is is	VERDORIE? darn	VERDORIE? darn	VERDORIE? darn	verdwaald lost

There are 9 transparent cases in the written material and 18 in the spoken material. These cases are discarded when statistical analyses of the position are carried out.

Some confusion may arise when an interjection occurs in an embedded clause, as for instance in 1:

1. Ik heb gezien dat VERDORIE de zon schijnt.
   I have seen that darn the sun shines
   'I noticed that the sun is shining, darn it!'

Example 1 contains two clauses: the matrix clause 'ik heb gezien dat verdorie de zon schijnt' and the embedded clause 'dat verdorie de zon schijnt'. Their topological analyses are depicted in Table 6.

[4] Except the V1 field which by definition can contain only one word.
[5] From here onwards, peripheral fields (LD, EX, RD) that remain empty are no longer shown for reasons of space.

Table 6: An analysis of example 1 at matrix clause level and embedded clause level

TOP	V1		MI	V2	EX
Ik I	heb have			gezien seen	dat VERDORIE de zon schijnt. that darn the sun shines
	dat that	VERDORIE darn	de zon the sun	schijnt shines	

At matrix clause level, the subordinate clause 'dat verdorie de zon schijnt' is in the extraposition field. But does that imply that the interjection 'verdorie' is in the extraposition field? In the embedded clause 'verdorie' is between V1 and the middle field. We decided to determine the position in the lowest clause rather than the matrix clause, hence in example 1 the position of 'verdorie' is V1-MI. This decision has consequences for the analysis of interjections at clause boundaries, e.g. the position between 'gezien' and 'dat' in example 1'.

1'. Ik heb gezien, TJA, dat de zon schijnt.
I have seen well that the sun shines
'I noticed, well, that the sun is shining.'

In example 1', the interjection occurs at the boundary between the main clause and the subordinate clause. It would be inconsistent to use the matrix-clause code V2-EX for example 1' while we do not use the matrix-clause code EX for example 1. Furthermore when two clauses are coordinated and an interjection occurs between them, it is unclear what position this interjection should be assigned to. Our decision to assign instances to the lowest clause requires a label for between-clause instances. We chose the special sign #.

4 Methodology

Each instance was annotated with respect to six variables: written or spoken, text type, position in the clause, type of interjection (e.g. *ja* 'yes', *nee* 'no', *verdorie* 'darn'), number of syllables, and whether or not it interrupts a major constituent. To determine whether these nominal variables reflect meaningful differences between groups of instances and whether they are related to each other, we used the program AnswerTree (SPSS, www.spss.com/answertree).

This program takes annotated data and automatically divides them into groups, using chi-square tests to determine whether the differences between those groups are significant. The program identifies the group division with the largest significant difference. The user can also define groups manually and check whether the

differences between them are significant or not; we set α to .05. The result looks like a tree structure in which the root is the entire set of data and the leaves are subgroups. The example in Figure 1 reflects the distribution over the clause of interjections in written and spoken material.

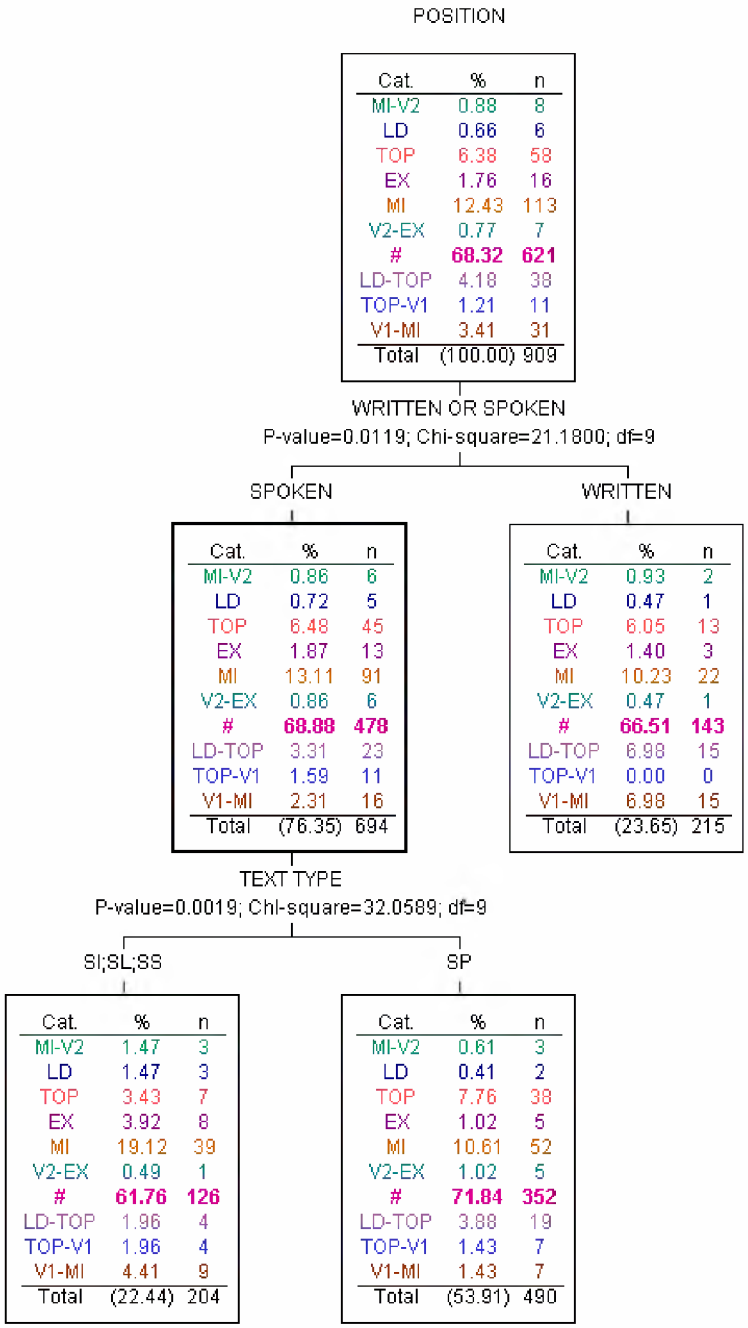

Figure 1: The distribution of the clause of interjections in spoken and written material

In Figure 1 the root contains all input data, which are divided over two leaves. The percentages given with the number of instances at each position apply to one leaf, not to the root. Percentages within one leaf always add up to 100. However, the percentages between parentheses at the lowest row of any leaf indicate the percentage of the root. The difference between leaves is determined from the differences between the percentages, to account for differences in group size. In Figure 1, the difference between spoken and written material is found to be significant (p = 0.01, chi-square = 21.2, df = 9). This seems to be the result of many smaller differences between the frequencies of occurrence in certain positions.

The leaves are then analyzed as new roots to determine whether there also are differences among the distributions of interjections in the various text types within the spoken or written material. For written material, there is no significant subdivision to be made. For spoken material, there is a significant difference between private conversations on the one hand and lectures, interviews and sports commentaries on the other hand (p < 0.005, chi-square = 32.1, df = 9).[6] The differences are large in the positions between clauses and within the middle field.

The instances in a transparent position are disregarded since they cannot be assigned exactly. In addition, the three instances in the category 'spoken, news' are excluded because we wanted to examine differences in distribution depending on the text type in which an interjection occurs. Three instances in a given category are insufficient for conclusions. Therefore, the total number of instances at the top of the tree is 909, not 939. For meaningful results, we often had to disregard instances that occurred only sporadically and whose classification in a certain group therefore seemed merely accidental. These cases are indicated at the relevant places in the discussion.

5 Results

The first hypothesis to check is the standard assumption in the literature that interjections can occur in all positions, but not within major constituents. Related questions are whether there are preferred positions for interjections, whether the prosody of the clause influences the distribution of interjections over the clause and whether various types of interjection behave differently. This section presents relevant data and discusses their initial implications; an extensive discussion of remaining issues appears in Section 6.

[6] We used abbreviations for the text types in Figure 1; SP = private conversation, SL = lecture, SI = interviews and SS = sports commentary.

1.1 The distribution of interjections over the clause

Table 7 summarizes the distribution of interjections over the positions in the clause in spoken and written material.[7] Apart from the hypothesis that interjections can occur in almost any position except within major constituents, no hypotheses can be derived from the literature about which positions are preferred or avoided. We will study the distribution of interjections in the topological framework as discussed in Section 3: e.g. LD means 'within the left dislocation field', LD-TOP means 'between the left dislocation field and the topicalization field', TOP means 'within the topicalization field' and so on. The sign # indicates the between-clause position and TRANSPARENT means that the position cannot be determined exactly.

Table 7: The distribution of interjections over the clause in the topological framework in written and spoken material

Position	**# interjections in written material**	**% interjections in written material**	**# interjections in spoken material**	**% interjections in spoken material**
LD	1	0.4	6	0.8
LD-TOP	15	6.7	23	3.2
TOP	13	5.8	45	6.3
TOP-V1	0	0.0	11	1.5
V1-MI	15	6.7	16	2.2
MI	22	9.8	92	12.9
MI-V2	2	0.9	6	0.8
V2	0	0.0	0	0.0
V2-EX	1	0.4	6	0.8
EX	3	1.3	13	1.8
EX-RD	0	0.0	0	0.0
RD	0	0.0	0	0.0
#	143	63.8	479	67.0
TRANSPARENT	9	4.0	18	2.5
Totals	224	100	715	100

[7] In Section 3 we suggested that LD, TOP and RD could only contain one major constituent. Since interjections are supposed not to interrupt major constituents, this would imply that EX and MI are the only topological fields in which interjections can occur. LD and TOP are also used, however. In these cases interjections are used when people repeat themselves, correct themselves or want to bridge the gap while they think about the continuation.

The position #, clause boundary, is most frequent: about 65% of the interjections occur in this position. The middle field, MI, is the most frequent clause-internal position, roughly 10% of the cases. The positions LD-TOP, TOP and V1-MI account for about 5% of the cases each, and the remaining positions are only used in a few percent of the cases. Some positions, viz. V2, EX-RD and RD are not used at all. The absence of instances in EX-RD and RD is easily explained, since these positions are rarely available; the right-dislocation field RD is rare. The absence of instances in the second verb field, V2, cannot be explained in the same way; this field is relatively frequent and can contain more than one word. The absence of interjections here apparently indicates that the coherence of the elements in the verbal cluster is too strong for interruption by an interjection.

Although the distribution of written and spoken material only differs a few percent per position, the difference is significant (p = 0.01, chi-square 21.5, df = 9; cf. Figure 1). The distribution of interjections in the spoken material is significantly different between private conversations on the one hand and sports commentary, lectures and interviews on the other hand (p < 0.005, chi-square = 32.1, df = 9).[8] This contrast suggests a stylistic difference between public and private speech.

It appears that the general assumption about major constituents is correct: interjections rarely interrupt them. Of the 939 instances we found, only 25 interrupted a major constituent, all in the spoken material. Another 80 instances, 17 in the written material and 63 in the spoken material, occurred in situations of self-correction, repetition of a word or a string of words, restarts and the like. In these situations it is debatable whether they are interrupting a major constituent or not, but in our view they are not.

Although Table 7 suggests that interjections show preferences for certain positions, we have to bear in mind that this would only be true if all positions were equally available for interruption by an interjection. This is obviously not the case, however. A certain topological position is only available for an interjection when this particular topological field is occupied, e.g. in the clause 'hij is verdwaald' *he got lost*, as exemplified in Table 5, the positions EX, EX-RD and RD are not available. It seems straightforward that the field RD will be used less frequently than the field MI, and hence that interjections can occur less often in the former field than in the latter. To account for this difference, the frequencies of interjections in certain fields should be divided by the frequencies of use of those fields.

However, to our knowledge, there is no large corpus of written or spoken Dutch analyzed according to the topological framework, so the relative frequencies of

[8] We removed the three instances from the category news items since this number is too low to base conclusions about the entire category on it. Besides, we also removed the transparent instances.

topological fields are unknown. Therefore we must be careful with conclusions about high frequencies of interjections at certain positions. They might reflect a larger availability of these positions rather than a prosodic, stylistic or other reason to prefer these positions. However, this caveat holds both for written and for spoken material and for all types of interjection; therefore conclusions about the differences between groups of instances can be drawn safely from this position analysis.

1.2 Does the prosody of the clause influence the distribution of interjections?

The influence of prosody on the distribution of interjections is difficult to predict, since many factors play a role in the prosody of a clause. Unfortunately the prosody cannot be observed in the data; obviously the written material is not prosodically annotated, and only a small proportion of the CGN received a prosodic annotation. We restricted our analysis to one factor only: it seems likely that the length of an interjection could influence the positions at which it can occur. We expect that the longer an interjection is, the more difficult it becomes to integrate into the intonation contour of the clause. The number of syllables proved a good indicator: we found a significant difference ($p < 0.005$, chi-square 37.5, df = 9) between interjections with a length of one syllable (*monosyllabic* interjections) and interjections with a length of two, three and four syllables (*polysyllabic* interjections).[9]

Table 8: The distribution of monosyllabic and polysyllabic interjections over the topological positions

Position	**One syllable**		**More syllables**	
	#	**%**	**#**	**%**
LD	6	0.79	1	0.65
LD-TOP	34	4.49	4	2.61
TOP	54	7.13	4	2.61
TOP-V1	8	1.06	3	1.96
V1-MI	15	1.98	16	10.46
MI	89	11.76	25	16.34
MI-V2	6	0.79	2	1.31
V2-EX	5	0.66	2	1.31
EX	14	1.85	2	1.31
#	526	69.48	94	61.44
Totals	757	100	153	100

[9] There are no interjections with a length of five syllables and two interjections with a length of six syllables in the data; these two instances were removed for this test for reasons of parsimony. Instances at a transparent position were also removed.

Most instances are monosyllabic interjections; there are 757 monosyllabic interjections (most of which are *ja* or *hè*, cf. Table 3), while there are 153 polysyllabic interjections. The main difference between monosyllabic and polysyllabic interjections is that monosyllabic interjections occur more often at clause boundaries and less often at the positions within and preceding the middle field, MI and V1-MI. The differences become clearer when we divide the groups of monosyllabic and polysyllabic interjections by written or spoken text.

Table 9: The distribution over the clause of monosyllabic interjections in spoken and written material

Position	In spoken material		In written material	
	#	%	#	%
LD	5	0.81	1	0.71
LD-TOP	21	3.41	13	9.22
TOP	42	6.82	12	8.51
TOP-V1	8	1.30	0	0.00
V1-MI	15	2.44	0	0.00
MI	81	13.15	8	5.67
MI-V2	6	0.97	0	0.00
V2-EX	5	0.81	0	0.00
EX	12	1.95	2	1.42
#	421	68.34	105	74.47
Totals	616	100	141	100

Table 10: The distribution over the clause of polysyllabic interjections in spoken and written material

Position	In spoken material		In written material	
	#	%	#	%
LD	1	1.23	0	0.00
LD-TOP	2	2.47	2	2.78
TOP	3	3.70	1	1.33
TOP-V1	3	3.70	0	0.00
V1-MI	1	1.23	15	20.83
MI	11	13.58	14	19.44
MI-V2	0	0.00	2	2.78
V2-EX	1	1.23	1	1.39
EX	1	1.23	1	1.39
#	58	71.60	36	50.00
Totals	616	100	141	100

Comparing the distribution of monosyllabic and polysyllabic interjections in written material, polysyllabic interjections show a stronger preference for the positions V1–MI and MI, while monosyllabic interjections are more frequent between clauses. The differences between monosyllabic and polysyllabic interjections in spoken material, however, are now much narrower. Apparently the differences in position preference between monosyllabic and polysyllabic interjections on the whole are due almost exclusively to the differences in written material. This is remarkable, since one would expect the influence of prosody to be greater in spoken material than in written material. A closer examination reveals that 42 of the 72 polysyllabic interjections in the written material are swearwords. An alternative explanation for the significant difference between monosyllabic and polysyllabic interjections in written material might therefore be that swearwords have a distribution which is significantly different from the distribution of other types of interjection. We will return to this issue in Section 6.

1.3 The distribution of various types of interjection over the various text types

When we look at the preferences of each type of interjection for a certain text type, we see that there is a significant difference ($p < 0.005$, chi-square = 260, df = 17) between spoken and written material. That holds at least for interjections that occur often enough in the material to yield statistically significant results; these are only the 18 types of interjection that occur five times or more in the data.[10]

Table 11: The distribution of 18 types of interjection over spoken and written material

Type of interjection	**Translation**	**In spoken material**		**In written material**	
		#	**%**	**#**	**%**
ach	*oh well*	3	0.45	9	5.17
godverdomme	*goddamn*	0	0.00	5	2.87
goed	*okay*	5	0.75	0	0.00
hè	*right*	100	15.06	1	0.57
hoor	*really*	38	5.72	13	7.47
ja	*yes*	425	64.01	64	36.78
ja goed	*yes okay*	9	1.36	0	0.00
ja hoor	*oh yes*	0	0.00	7	4.02
ja nou	*definitely*	5	0.75	0	0.00

[10] These 18 types together cover 841 instances. In order to study the distribution over text types, we removed the 3 instances in spoken news; this number was too small to draw reliable conclusions from these instances.

nee	*no*	31	4.67	38	21.84
nee hoor	*oh no*	1	0.15	5	2.87
nou	*well*	16	2.41	2	1.15
nou ja	*well*	17	2.56	5	2.87
oh	*oh*	4	0.60	2	1.15
pardon	*pardon me*	3	0.45	1	0.57
sorry	*sorry*	7	1.05	1	0.57
verdomme	*damn*	0	0.00	15	8.62
verdorie	*darn*	0	0.00	6	3.45
Totals		664	100	174	100

The main differences are that *hè* and *ja* have a strong preference for spoken material, whereas nee occurs much more frequently in written material. *verdorie*, *godverdomme* and *verdomme* occur exclusively in written material. The difference in use over the text types between *ja* and *nee* is remarkable, since intuitively these words are each other's counterparts, so that one would expect an equal distribution. A possible explanation, suggested by a look at the instances, is that *ja* is often used with a non-affirmative function in spoken material, such as back channel or as a filler or placeholder, a function which *nee* is less likely to fulfil. An example from the spoken corpus is given in 2.

2. en zo ben ik uh ja eigenlijk al op deze school beland waar ik (...)
   and so am I uh yes actually already on this school ended_up where I
   'and that's how I, well, ended up at this school where I...'

Since it is not necessary to prevent an interruption by the conversation partner in written material, this function is rare in writing; the only exception is fiction dialogues or interviews. This could explain why the differences between the frequencies of *ja* and *nee* are smaller in written material than in spoken material. This explanation seems even more likely when we examine the spoken material, in Table 12.[11]

Table 12: The distribution of 14 types of interjection over four spoken text types

Interjection	**Private conversation**		**Lecture, interview, sports commentaries**	
	#	**%**	**#**	**%**
ach	2	0.43	1	0.51

[11] The types which do not occur in the spoken material are not depicted in Table 12.

goed	3	0.64	2	1.02
hè	32	6.85	68	34.52
hoor	29	6.21	9	4.57
ja	337	72.16	88	44.67
ja goed	9	1.93	0	0.00
ja nou	2	0.43	3	1.52
nee	22	4.71	9	4.57
nee hoor	1	0.21	0	0.00
nou	14	3.00	2	1.02
nou ja	10	2.14	7	3.55
oh	4	0.86	0	0.00
pardon	0	0.00	3	1.52
sorry	2	0.43	5	2.54
Totals	467	100	197	100

The distribution of interjections in the spoken material reveals two groups: the private conversations on the one hand and lectures, interviews and sports commentaries on the other ($p < 0.005$, chi-square = 112, df = 13). The main difference is that in private conversations the word *ja* occurs more frequently than in public material. The use of *ja* as a filler or placeholder is more often necessary in private conversations than in sports commentaries or lectures, where an interruption by the conversation partner is unlikely.

6 Discussion

The results presented in Section 5 confirm the standard assumption in the literature: interjections can occur in almost all structural positions, but they rarely interrupt a major constituent. Their distribution differs between spoken and written material, and in the spoken material it also differs between public monologues on the one hand and private dialogues on the other hand. This probably reflects a stylistic difference between those text types. Various interjections show clear preferences for text types, presumably also due to stylistic differences. Whether the length of an interjection influences its distribution remains unclear; at the end of this section we will present additional research into this question. First, we will discuss two other results of Section 5: the very frequent use of the position between clauses, #, and the almost entire absence of swearwords in spoken language.

1.1 The preference of interjections for the position between clauses

The frequency of the between-clause position # is so high, 65%, that it can hardly be explained by the frequent availability of this position. Apparently there is a reason to prefer interjections at positions between clauses over positions within clauses. The explanation may be related to the prosody of the clause and the extra-grammatical nature of interjections. Clause boundaries seem more suitable for extra-grammatical constructions than clause-internal positions; besides, these boundaries often show a prosodic pause, which may make it easier to interject a clause-external element. The fact that the second preference for interjections is the position within the middle field seems to falsify this hypothesis, but many of these instances indicate hesitation, self-correction and the like. An example from the spoken corpus is given in 3:

3. dat zal over uh zo ongeveer uh nou_ja zesenhalve minuut zijn.
   that will over uh so about uh well six_and_a_half minute be
   'That will be in, well, about six minutes and a half.'

The need to keep the turn while thinking or to correct the previous word applies most frequently with content words that carry new information, which often occur in the middle field. This goes for spoken language as well as for fiction dialogues. In addition, interjections can occur as intensifiers in the written material, as in 4, which is derived from the corpus:

4. (...)de geëxecuteerde (...) karakteriserend als die stoere Belg ja zelfs
   the executed characterizing as that tough Belgian yes even
   toesprekend met (...)
   addressing with
   '... characterizing the executed one as that tough Belgian, yes even addressing him with...'

Thus there is a functional explanation for the occurrences in the middle field, which need not exclude the explanation for the preference for the between-clause position. Of course, further research into this hypothesis is necessary.

1.2 Swearwords

At first sight it seems surprising that swearwords occur more often in written than in spoken material. The most likely explanation is that all speakers in the CGN knew that their speech was being recorded. This might have made them more careful in the choice of their words, a well-known drawback of the legal obligation to ask people's

permission to record their speech in advance. This observation is reinforced by the fact that common taboo words like *fuck* 'fuck', *kut* 'cunt', *shit* 'shit' do not occur in the data anywhere. The three swearwords *verdorie* 'darn', *godverdomme* 'goddamn' and *verdomme* 'damn' have become a bit old-fashioned and less shocking than they once were. It makes them suitable for use in fiction dialogue, which is the most important source of these instances.

1.3 Prosodic and functional influence on the distribution of interjections

The results presented in Section 5 suggest that the position of an interjection is determined by its length, among other factors. However, about 3/5 of the polysyllabic interjections in the written material are swearwords. This makes it very difficult to decide whether it is really the length of an interjection which plays a role here, which would imply influence of the prosody, or the type of interjection, which would suggest that the function of the interjection plays a role. This issue cannot be resolved by dividing the group of polysyllabic interjections into two groups, viz. swearwords and other interjections, since too few instances would remain in each group to gain reliable results.

An alternative approach is to classify the instances on the basis of their meaning, as argued in the ANS and summarized in Section 1. If each instance in the corpus is classified into a group according to its meaning, we can test whether these groups have different distributions. The first problem is that many instances do not occur in their primary meaning but are used with a discourse function. *ja* 'yes' often acts as a placeholder, bridging the gap while the speaker is thinking, and *pardon* 'sorry' or *sorry* 'sorry' do not ask for clarification or repetition as the ANS suggests, but introduce a self-correction. The second problem is that some interjections can be used in different but related ways; so the classification becomes rather subjective. The same problem applies to classifying interjections not included in ANS' examples. The description is not formal enough to classify new instances objectively. Because of the serious risk of subjectivity, we did not apply this classification to our data. Consequently, we could not test whether a difference in meaning of an interjection is in any way related to a different distribution.

To determine whether the difference between monosyllabic and polysyllabic interjections in written material was due to their length or to the type of the polysyllabic interjections, we compared interjections with a different type of interruption construction. We chose the so-called parentheticals; constructions like *lijkt me* 'it seems to me', *dacht ik* 'I thought', *meen ik* 'I think'. The function of these constructions is to add a meta-comment to the surrounding clause, providing the information that it is just a personal opinion. This function is clearly different from the function of

interjections. What parentheticals have in common with interjections, however, is that they are not a grammatical part of the clause; both of them interrupt the clause (cf. Schelfhout et al. 2004). In addition, parentheticals consist of at least two words; consequently they are all at least two syllables long. Therefore we feel that a comparison of parentheticals and polysyllabic interjections could shed more light on the role of the function of an interruption versus the role of the length of the interruption in deciding its position.

We extracted the utterances containing one or more parentheticals from a corpus which is a superset of the corpus from which the interjections were extracted. The written part of the corpus is exactly the same, but the spoken part was extended to almost half a million words. We found 271 parentheticals in this material.[12] These instances were encoded in the same way as interjections with respect to the text types in which they originated, their position and their length. The swearwords were separated from the interjections to be able to see with what type of interruption they would cluster. The combined data were analyzed by AnswerTree.

We checked whether the different types of interruption resulted in a different pattern of position preference.[13] As parentheticals always contain at least two syllables, interjections consisting of only one syllable were removed from the data to keep the comparison fair. Parentheticals and swearwords together show a distribution significantly different from that of all other types of interjections ($p < 0.005$, chi-square = 80.5, df = 8).

Table 13: The distribution of interjections, swearwords and parentheticals over the clause

Position	**Interjections without swearwords**		**Swearwords and parentheticals**	
	#	**%**	**#**	**%**
LD-TOP	3	2.75	7	2.31
TOP	4	3.67	8	2.64
TOP-V1	2	1.83	15	4.95
V1-MI	2	1.83	59	19.47
MI	13	11.93	108	35.64
MI-V2	0	0.00	6	1.98
V2-EX	2	1.83	9	2.97
EX	2	1.83	7	2.31
#	81	74.31	84	27.72

[12] For more information on the parentheticals, see Schelfhout, Coppen & Oostdijk (2003).
[13] Interruptions at a transparent position (36 instances) were removed, as were instances occurring at the position RD, which was used in only 4 of 1210 instances. The position LD was used only once in this restricted set; this instance was removed for reasons of parsimony.

Totals	109	100	303	100

The difference between the distribution of parentheticals and swearwords is at the border of insignificance (p = 0.05, chi-square = 15.3, df = 8).

These differences suggest that the function of an interruption can cause a significant difference in the preference for a certain position. More specifically, parentheticals and swearwords show a stronger preference for positions preceding and within the middle field, V1-MI and MI, and occur less frequently at a clause boundary (#) in comparison to all other types of interjection. But does a different function explain all the differences? When we classify the entire set of data by the length of the interruptions,[14] we obtain a significant result as well.

Table 15: Distribution over the clause of interruptions with a length of one, two or more syllables

Position	**One syllable**		**Two syllables**		**More syllables**	
	#	**%**	**#**	**%**	**#**	**%**
LD	6	0.79	1	0.37	0	0.00
LD-TOP	34	4.49	7	2.59	3	2.10
TOP	54	7.13	10	3.70	2	1.40
TOP-V1	8	1.06	7	2.59	10	6.99
V1-MI	15	1.98	33	12.22	28	19.58
MI	89	11.76	74	27.41	47	32.87
MI-V2	6	0.79	3	1.11	3	2.10
V2-EX	5	0.66	9	3.33	2	1.40
EX	14	1.85	8	2.96	1	0.70
#	526	69.48	118	43.70	47	32.87
Totals	757	100	270	100	143	100

Now there is a clear subdivision between interruptions of one syllable, of two syllables and of three and more syllables long (p < 0.005, chi-square = 212, df = 8). The categories which contain polysyllabic interruptions contain a mixture of interjections (excluding swearwords), swearwords and parentheticals; it looks like the two-syllable and more-syllable categories are not identifiable with mainly parentheticals or mainly interjections, but really contain both kinds of structures.[15] Therefore, it appears that both the length of an interruption and its function influence the distribution.

[14] We used five categories for the length of an interruption. It can be 1 through 4 syllables long, or it is five or more syllables long.

[15] Of course, the monosyllabic category does not contain parentheticals.

7 Conclusion

Interjections can occur in all structural positions in the clause, but they cannot interrupt major constituents. Although we cannot be certain without more information on the relative use of topological fields, interjections seem to show preferences for certain positions. The position between clauses seems to be the most favoured one, followed by the position within the middle field. The explanation for these preferences is probably a combination of grammatical, functional and prosodic factors. The use of interjections is clearly different in spoken and written language. In written material, interjections are more frequent in text types which reflect spoken language use. In spoken material, we also see a difference between the use of interjections in private and public text types. The difference can be seen both in the types of interjection and in the positions they occupy. Those positions also seem to be influenced by the length of the interjection: monosyllabic interjections differ significantly from polysyllabic interjections, especially in written material. A closer look reveals that polysyllabic interjections in written material are mainly swearwords, which raises the question whether the length or the function of the interjection causes the difference in distribution. A comparison between two functionally different types of interruption, interjections and parentheticals, suggests that both prosodic and functional factors influences their distribution.

This study confirms and amplifies standard assumptions in the literature not previously tested with authentic material. It also raises issues for future research, amongst which the exact role of the function versus the prosody of interjections in deciding the distribution over the clause, their function in the discourse and the interaction of these factors. Also the need for a corpus annotated according to the topological descriptive model is underlined once more, since this is the only way to gain insight into the preferential distribution of interjections. In view of the varying frequency of interjections over different text types and in different positions, more data would be helpful for fine-tuning the analyses presented here, especially to determine whether there is a difference in use between swearwords and other types of interjection. More data could also be beneficial to determine whether there are differences between the various written text types, for which we have too few instances at the moment. Finally, the overall distribution shows that more than half the instances occur in a single position, between clauses. Since there are clauses of many different types, which can be coordinated or subordinated to each other, the variation within this group is large. It would be worthwhile to split up this group, for instance into positions between coordinated and subordinated clauses, or into positions preceding clauses which do or do not carry thematic roles, and see if this provides more information.

Acknowledgements

We would like to thank Antal van den Bosch and Hans van Halteren for their help in tagging the written corpus and Bill Fletcher for proofreading this paper. We are also thankful to two anonymous reviewers and the editors of Neerlandistiek.nl for their comments and helpful suggestions.

References

Brummel, G. (1978). Enkele opmerkingen over interjecties. In Berkel, A. (ed.), *Proeven van Neerlandistiek aangeboden aan Prof. Dr. Albert A. Sassen*, 155-175. Groningen: Nederlands instituut.

Haegeman, L. (1984). Interjections and Phrase Structure. In *Linguistics* 22: 1, 41-49.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn, eds. (1997). *Algemene Nederlandse Spraakkunst*. Groningen / Deurne: Martinus Nijhoff / Wolters Plantyn.

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

Romijn, K. (1998). Eh: substitutie- en aarzelingsinterjectie. In *TABU* 28: 2, 72-87.

Schelfhout, C., P.-A. Coppen & N. Oostdijk (2003). Positions of parentheticals and interjections: A corpus-based approach. In Cornips, L. & P. Fikkert (eds.), *Linguistics in the Netherlands 2003*, 155-166. Amsterdam: John Benjamins Publishing Company.

——— (2004). Finite comment clauses in Dutch: a corpus-based approach. In *Journal of Germanic Linguistics* 16: 4, 331-349.

Van den Toorn, M.C. (1968). De interjectie als woordsoort. In Hoogteijling, J. (ed.), *Taalkunde in artikelen*, 115-119. Groningen: Wolters-Noordhoff. Also printed in *De Nieuwe Taalgids* 53 (1960), 260-264.

De Vriendt, S. (1992). 'kom', 'kijk', 'zeg' als interjectie. In *Studia Neerlandica et Germanica*, 513-520.