

On certain syntactic properties of spoken Dutch

Heleen Hoekstra*, Michael Moortgat*, Bram Renmans+,
Ineke Schuurman+, Ton van der Wouden*

(*UiL-OTS Utrecht, +CCL Leuven)*

CLIN-dag, 30 november 2001

*Nog vele problemen met woorden, zinsdelen en zinnen
wachten op een behandeling waarbij de computer een rol kan spelen
(Brandt Corstius 1970:163)*

Abstract

Certain Dutch construction types are much more frequent in spoken than in written language. Topic drop sentences such as (1) and constructions as in (2) (which we might call ‘mirror sentences’) are relatively common in spoken language but rare in written prose (with the exception, perhaps, of the most informal of text types such as Internet chat¹).

- (1) doen we
do we
‘(we agree that) we will do that’
- (2) je moet heel snel moet je weer wisselen
you must very soon must you again change
‘you have to change again very soon’

In this paper we present some preliminary quantitative and qualitative results with respect to typical spoken language phenomena in various subcorpora of the Spoken Dutch Corpus (CGN).

Concept-structuur:

- inleiding: weinig bekend van gesproken Nederlands. CGN
- wat typische spreektaalfenomenen
 - spiegels
 - topicdrop
- wat kwantitatieve eigenschappen

*Thanks to Laura Korte and Vincent Vandeghinste for computational support, and to Norbert Corver and Maaike Schoorlemmer for discussion.

¹Cf. Grondelaers *et al.* (2000).

1 Introduction: Formal properties of spoken Dutch

Dutch is one of the languages in the world that are studied most thoroughly. Still, many areas of the language are virtually unexplored. For example, grammatical analysis usually deals with written variants of the standard language – the large ANS grammar (Haeseryn *et al.* 1997) is essentially about the written language used by a well-educated elite.² This does not mean that there is no tradition of interest into properties of spoken Dutch: starting with the studies by the Groningen high school teacher Wobbe de Vries around 1910 (de Vries 1910; de Vries 1911; de Vries 1914) through the dissertations of Bertha Uijlings (Uijlings 1956) and Frank Jansen (Jansen 1981) to the huge recent book by Jelle de Vries (de Vries 2001), plus a considerable number of smaller papers, there is a fair amount of data and analyses. However, these studies dealing with oral Dutch can be divided into two groups: either they are in-depth studies of one or two spoken language constructions, or anecdotal overviews of a great number of peculiar constructions, with shallow analyses at best. There is nothing like a comprehensive grammar of spoken vernacular Dutch.

To mention another *terra incognita*, there hardly exist any quantitative data with respect to Dutch, apart from word frequency counts (van Berckel *et al.* 1965; Uit den Boogaart 1975) and some quantitative (corpus-based) explorations into the productivity of bound morphemes (Baayen 1989). As far as we know, however, hardly anything is known about, e.g., the distribution of various clause types and sentence types in spoken Dutch (or written Dutch, for that matter). There haven't been any studies comparable to de Haan & Oostdijk (1994) or Dick & Elman (2001). We assume that the main reason for this lack of quantitative data concerning Dutch lies in the fact that although there is no lack of corpora of various types of text (e.g. the Eindhoven Corpus and the various corpora of the INL), none of these is syntactically annotated.

Given that there has been little systematic research into properties of spoken Dutch, and hardly any tradition of quantitative linguistics with Dutch as an object language, it can't surprise us that hardly anything quantitative is known about spoken Dutch – the only exception being the word frequency lists in de Jong (1979). In any case, there is nothing that comes even close to a book such as Miller & Weinert (1998), a corpus based study into syntactic peculiarities of spontaneous spoken English.

This situation, however, is going to change with the *Corpus Gesproken Nederlands* (CGN), the Spoken Dutch Corpus. Among the goals of the corpus project are the following (Oostdijk 2000a; Oostdijk 2000b):

- to collect 10 million words of spoken Dutch, both from the Netherlands and Flanders, the Dutch speaking part of Belgium (2/3 – 1/3)
- to transcribe everything orthographically
- to supply lemmatisation and morphosyntactic annotation for everything (Van Eynde *et al.* 2000)
- to syntactically annotate 1 Million words, both from the Netherlands and Flanders (2/3 – 1/3) (Hoekstra *et al.* 2000)
- etcetera

In this paper, we will first briefly discuss some constructions which are typical for spoken language. After that, we will offer some preliminary quantitative data from the part of the CGN that has already been annotated syntactically.

²Not to mention its prescriptive aspect.

2 Some spoken language phenomena

2.1 Topic drop sentences

It is not uncommon in colloquial spoken Dutch to leave out the first constituent of a sentence. A few examples are given in (1).

- (1) a. — *Is goed!*
[dat/het] ‘that/it’ (SU) is well
‘OK!’
- b. — *Maken ook onze kostuums zelf en zo dus*
[We] ‘we’ (SU) make also our costumes self and so PART
‘we also make our costumes ourselves and things like that’
- c. — *Lijkt me wel een leuke gast in principe.*
[dat] ‘that’ (SU) seems me PART a nice guest in principle
‘he looks like a nice guy to me’
- d. — *Doen we!*
[dat] ‘that’ (DO) do we
‘let’s do that’
- e. — *Heb ik toch verteld van die onafgewerkte*u lach toch?*
[dat] ‘that’ (DO) have I PART told of that unfinished? laugh PART
‘I told you about that unfinished laugh, didn’t I?’
- f. — *Zeg je?*
[wat] ‘what’ (DO) say you?
‘What do you say?’
- g. — *Ben ik helemaal niet mee akkoord maar enfin*
[daar] ‘there’ am I totally not with agreement but OK (preposition object)
‘I completely disagree but OK’
- h. *Nee — heb ’k niet zo’n zin in.*
no [daar] ‘there’ have I not such-a sentence in (preposition object)
‘No I don’t like that too much’

It is clear that the syntactic function of the first element is not crucial: we find (at least) subjects, objects and prepositional objects. Still, the distribution of such “phonetically empty categories” – to borrow some GB terminology (de Haan & Tuijnman 1988) – “is heavily constrained. One of the requirements on such categories is, that they have to be recoverable.”³

A more or less standard analysis in GB-like terms goes along the following lines:⁴ (matrix clause) topicalisation is actually operator movement. The topicalized element is linked to a null operator in [Spec,CP],

³Cf. also Haeseryn *et al.* (1997:1114): “In het algemeen kunnen onvolledige zinnen gebruikt worden als de betekenis ervan voldoende duidelijk blijkt uit de situatie.” Zie ook Huang (1984). Stoett (1923:§219) geeft Middelnederlandse voorbeelden van het weglaten van het voornaamwoordelijk bijwoord, bijvoorbeeld *want zekerlic ic moestet zoeken ende om waken vele nachte* en verwijst naar de Vries (1910:79), die in een discussie van topicalisatie-zinnen (*die arme mensen die zag ik daar, die mensen die geef ik niets, die arme visschers die der schip is vergaan*) opmerkt dat met name *daar* als deel van het voornaamwoordelijk bijwoord wel weggelaten wordt, kortom, de bananen-zinnen in de terminologie van van der Horst & van der Horst (1999): *dat mes kan ik niet mee snijden, kinderen moet je niet zoo ruw tegen wezen, arbeiden heeft hij een hekel aan, de gunstige werking die het . . . uitoefent zijn H.H. doctoren het beslist over eens*; veel zeldzamer is volgens hem *ik heb de(n) notaris gesproken; is een gekke kerel*.

⁴After Neeleman (1994:31), who refers to Chomsky (1977); Koster (1978); Weerman (1992).

which in turn binds a trace (2a). Optionally, the null operator can be spelled out (2b) or the pre-clausal category linked to the null operator may be absent if the operator can be interpreted via the context – the so-called topic-drop sentences (2c).

- (2) a. [Zulke boeken]_i 0_i heeft Jan nooit t_i gelezen
 such books has John never read
 b. [Zulke boeken]_i die_i heeft Jan nooit t_i gelezen
 such books those has John never read
 c. A: Wat vindt Jan van zulke boeken?
 What thinks John of those books?
 B: 0_i heeft hij nooit t_i gelezen
 has he never read

The following example shows that it's not just the topic that can be dropped: a (light) verb may be left out as well:

- (3) — *niet leuk.* — *is echt niet leuk om te zien.*
 [dat is] 'that is' not nice. [dat] is really not nice for to see
 'that's not nice. that's really not nice to look at'

One might wonder how GB syntacticians would deal with the first sentence of this example. However, it is not our present goal to reject the classical analysis of this type of sentences and propose an alternative. Rather, we wanted to show that the CGN offers ways to quickly find new, real data that may help to shed more light on many language phenomena.

2.2 Mirror sentences

Another spoken language phenomenon – less known and less frequent – involves something which looks like a gross violation of all we think we know about syntax: main clauses containing more than one inflected verb, two subjects, etcetera.⁵

- (4) Ik zie Piet zie ik

Again, this construction (or phenomenon) has not gone unnoticed in the literature. Verdam (1923:111-113) sees it as a symptom of the degeneration of the Dutch spoken language

- (5) “een symptoom van de verwildering van de Nederlandse spreektaal, een gevolg van onze tot bandeloosheid overhellende vrijheidszucht en oorzaak dat de spreektaal verder dan nodig was van de schrijftaal is afgedwaald”

(Verdam (1923:111-113) as quoted in Sassen (1967:31))

According to Jansen (1981:p.224), this type of construction is an amalgam of two constructions, the result of some sort of sentence entanglement cum deletion:

- (6) a. Ik zie Piet zie ik
 b. Ik zie \cancel{P} \cancel{z} \cancel{e} \cancel{z} Piet zie ik

⁵Mooie voorbeelden (passen ook op het scherm, ook in Portray) 283, 123 283, 183 283, 233 283, 289 321, 237 347, 86 430, 13

We will refrain from giving the details of Jansen's contraction transformation⁶ The transformation is a very special one, because it works across sentence boundaries.

Another approach to the phenomenon can be found in Nicole Huesken's Master's Thesis (Huesken 2001). She talks about "mirror constructions", as the sentences involved often show a kind of mirror structure, with some constituent (hardly ever the inflected verb) as the central point.

According to her the mirror centre functions as a discourse link (p.51): usually the sentence part after topic plus inflected verb is the focus. The speaker, however, doesn't want the inserted link to be interpreted as focus, and therefore continues the sentence as if the mirror part is the topic.

Whatever one's analysis (one could also think of a kind of reduplication or something analogous to the spelling out of traces of movement (van der Wouden & Hoekstra 2001)), the Corpus Gesproken Nederlands makes it fairly easy to quickly find examples of the phenomenon by looking for main clauses with more than one verbal head. A few examples are given in (7) – note that various constituents can function as the mirror centre.

- (7) a. *Maar in wezen is een obligatie is gewoon een onderhandse lening* (V-SU-V)
 but in fact is a bond is simply a private loan
 'Actually, a bond is simply a private loan'
- b. *En dat is een erg interessant blad is dat* (SU-V-PREDC-V-S)
 and that is a very interesting journal is that
 'That is a very interesting journal'
- c. *We hebben sinds een paar jaar hebben we een nieuwe spelling* (S-V-MOD-V-S)
 We have since a few years have we a new spelling
 'We have a new spelling since a few years'

If one looks for main clauses with more than one verbal head, one also find sentences such as the ones in (8), in which the verb in the second part of the construction is not a 'mirror' image of the first verb, but where it is 'replaced' by another verb. In some cases, as in (8b), the function of the mirror centre may be different with respect to the two verbs.

- (8) a. *Want je krijgt dus wel 10 procent van 't bedrag mag je zelf houden*
 because you get PART PART 10 percent of the amount may you self keep
hè (S-V-DO-V-S)
 'You get/may keep 10 percent of the amount'
- b. *nou dan gaat er zeven centimeter over de volle breedte snijden ze d'r*
 PART PART goes there seven centimetre over the full width cut they there
zo af (V-SU/DO-V-S)
 PART off
 'A strip with a width of seven centimeters is cut off'

These sentences may be taken as an argument in favour of an analysis of a construction *apo koinou* as found in Overdiep (1937) and in Haeseryn *et al.* (1997:1259 ff.). According to this analysis, the mirror centre has a double function: it is the last element of the first sentence fragment, and the first element of the of the second sentence fragment.

Again, we are not here to defend one or the other analysis, we just want to show that the CGN can be used to quickly find relevant data of some construction.

⁶They can be found at Jansen (1981:p.225).

3 Some quantitative data on spoken Dutch

In this section, we will present the result of some elementary corpus counts, inspired by Biber (1988), among others. But before we do so, we will have to give some explanation of what it exactly is that we count.

- SMAIN: main clause with the main verb in second position

(9) *Jan slaat Marie*
John hits Mary
'John hits Mary'

(10) *een deel van de budgetten van de interne afdelingsbudgetten van de*
a part of the budgets of the internal department budgets of the
*geormerkte budgetten zijn natuurlijk in*a intern ook wel degelijk*
earmarked budgets are of course ??? internally PART PART PART
geormerkt.
earmarked
'part of the internal departmental budgets are earmarked of course'

- SSUB: subordinate clause: main verb in last position (but beware: prepositional phrases may show up after the verb)

(11) *...dat Jan Marie slaat*
...that John Mary hits
'that John hits Mary'

(12) *...zoals dat ook 't geval is met laten we zeggen de additionele middenlaag*
...as that also the case is with let we say the additional middle layer
'as is the case with, let us say, the additional middle layer too'

- WHREL: headless relative

(13) *wie honden slaat (is een slecht mens)*
who dogs beats (is a bad human)
'whoever beats dogs (is a bad human being)'

(14) *wat ik nu zie is in dit verdelingsvoorstel dat de budgetten toegewezen*
what I now see is in this division-proposal that the budgets allotted
worden aan de opleidingen binnen een onderwijsinstituut.
are to the schools within an educational institution
'what I see is that according to this proposal the budgets are endowed to schools within educational institutions'

- WHSUB: embedded WH-question

(15) *ik vroeg hoeveel je denkt dat hij weegt*
I asked how-much you think that he weighs
'I asked how much you think he weighs'

(16) *allerbelangrijkste is een goeie samenvatting waarin staat met wie je 't doet*
 most-important is a good resume wherein stands with who you it do
wie verantwoordelijk is voor wat hoe je het gaat doen wat je gaat doen
 who responsible is for what how you it go do what you go do
en waarom je denkt te menen jouw onderzoek goed te moeten keuren.
 and why you think to assume your research good to assess
 'most important is a good resume in which you state with whom you do it, who is responsible
 for what, how you will do it, what you will do and why you think your researched should be
 approved'

- SVAN: embedded sentences with the preposition *van* 'of' functioning as a kind of complementizer – something like one of the modern usages of *like* in English.

(17) *die vroeg aan mij van: is die dan getrouwd?*
 that asked to me VAN: is that PART married
 who asked me like: is he married?

(18) *dan heb ik zoiets van: laat maar...*
 then have I something VAN: let PART
 'then I am like: leave it'

(19) *dus dat even voor wat betreft van wat voor soort van activiteiten*
 so that PART for what concerned VAN what for kind of activities
worden nou vergoed?
 are PART reimbursed?
 'so far for the moment about the kind of activities that are reimbursed'

- SV1 sentences with the main verb in first position (yes/no questions, imperatives, topic drop sentences ...)

(20) *wordt u geholpen?*
 are you served?
 'are you being served?'

(21) *kijk maar uit*
 watch PART out
 'you'd better watch out'

(22) *doen we*
 do we
 'OK'

Now we are ready for the first table on your hand-out (23).

(23)

Quantitative properties of two subcorpora of CGN (N=150194)			
	Netherlands	Flanders	N/B
words	92631	57563	1,61
bytes	476793	311265	1,53
bytes/word	5,2	5,4	0,95
words/SMAIN	14,2	14,5	0,98
SMAIN	6529	3963	1,65
SSUB	2476	1658	1,49
REL	582	506	1,15
WHREL	172	104	1,65
WHSUB	188	77	2,44
SVAN	105	46	2,35
SSUB+REL+WHREL+WHSUB+SVAN	3523	2391	1,47
(tensed) embedded/SMAIN	0,54	0,60	0,89
SV1	1590	791	2,01
SV1/SMAIN	0,24	0,20	1,22

On the basis of these numbers one might want to draw the conclusion that the Flemish variant is slightly more formal: both the average sentence (computed as the total number of words divided by the total number of main clauses⁷) and the average word are somewhat longer in the Flemish subcorpus than in the Dutch material. In addition, there is a correlation with a somewhat higher degree of sentence embedding. Moreover, SVAN, which to Dutch natives sounds very informal, is much more frequent in the data from the Netherlands.

We should, however, not be jumping to conclusions too hastily. Another explanation for the differences found might be that the Flemish and the Dutch subcorpus are not completely comparable (yet): on both sides of the border, we are still working hard, but not always exactly on the same types of text.⁸

In order to get a better idea of the kind of differences there might exist (and are not an artifact of the current state of the corpus), we will now take a look at the properties of those subcorpora that have a reasonable size already.⁹ We selected four of these subcorpora:

- interview (with Dutch teachers)
- parliament (recordings of the Dutch “Tweede kamer” and the Flemish “Vlaamse raad”)
- radio (various types of broadcasts)
- spontaneous conversation (recorded especially for the CGN)

⁷Of course, we could compute the average sentence length in a different way, e.g. total number of words divided by the total number of main or embedded clauses. This yields 9,2 for the Dutch material vs. 9,1 for Flanders. What we cannot do, in any case, is divide the number of words by the number of full stops, as speakers don’t use or pronounce full stops. Miller & Weinert (1998:30-31) even claim that spoken language lacks sentences in the sense of entities that start with a capital and end with a full stop. Following Halliday, they assume that “the language system must be analyzed as having clauses combining into clause complexes” (p. 130). In the orthographic transcription of the CGN, the corpus data have been divided in text units (chunks) with a length of anywhere between 1 and over 130 words.

⁸Given that the complementizer *van* may function quite differently for many Flemish speakers – it may have a sort of modal meaning (Van Craenenbroeck 2000) – they may want to avoid the construction elsewhere in order not to cause misunderstanding.

⁹A note on corpus size: anyone saying that our corpus is too small to say anything interesting about spontaneous spoken Dutch will be referred to Miller & Weinert (1998) whose main English subcorpus consists of 50,000 words of spontaneous conversation (cf. their discussion on pages 10–14). Two of the subcorpora we are looking at are small compared to even that standard, but we will have to do with them for the time being.

As a starting hypothesis we expect parliamentary speeches to be most formal, spontaneous conversations the least formal, and the other two text types somewhere in between. The first results are in the second table (24):

(24)

Quantitative properties of four subcorpora of CGN (N= 121468)				
	interview	parliament	radio	spontaneous
words	45502	13850	10144	51972
bytes	251294	81875	62856	285421
bytes/word	5,5	5,9	6,2	5,5
SMAIN	3130	748	666	4083
words/SMAIN	14,5	18,5	15,2	12,7
SSUB	1347	580	258	1063
REL	382	162	95	199
WHREL	83	28	17	60
WHSUB	68	34	6	106
SVAN	41	5	3	65
SSUB+REL+WHREL+WHSUB+SVAN	1921	809	379	1493
embedded/SMAIN	0,61	1,1	0,57	0,37
SV1	622	107	78	1210
SV1/SMAIN	0,20	0,14	0,11	0,30

These numbers may be taken as support for our initial hypothesis: the average sentence length (taken again as words/SMAIN) is highest in the parliamentary speeches and lowest in the spontaneous conversations. The level of sentence embedding is also dramatically higher in the parliamentary material than elsewhere. The amount of SVAN and of verb initial constructions (SV1), both informal constructions intuitively, is also highest in the two subcorpora expected to be most informal, viz., interviews and spontaneous conversations. Surprisingly, however, the average word length is highest in the radio subcorpus with parliamentary speeches in second position.

Perhaps we should therefore adjust our initial hypothesis a little bit, in the sense that we make a subdivision between parliamentary speeches and radio recordings on the more formal side of the scale, and interviews and spontaneous conversions on the less formal side. This finds support in the numbers of the following table (25), where we compare the numbers of nouns and verbs. Traditionally (ubi?), nominal style is seen as more formal than verbal style.¹⁰

(25)

More quantitative properties of four subcorpora of CGN (N= 121468)				
	interview	parliament	radio	spontaneous
words	45502	13850	10144	51972
nouns	5359	2114	1950	5081
verbs	7393	2346	1580	8705
nouns/verbs	0,72	0,90	1,20	0,58

According to this metric, the parliament and radio subcorpora are the most formal again.

¹⁰In the POS-tagging, which is used here, nominalized infinitives, as in *het roken van sigaren* 'the smoking of cigars' are labeled N. If we take these into account (by counting verbal elements functioning as head of a noun phrase as nominal rather than verbal) the picture might change dramatically.

According to Biber (1988:241), English discourse particles (he mentions *well, now, anyway, anyhow, anyways*) are “rare outside the conversational genres”. Comparable things have been said about Dutch modal particles, which supposedly occur more in informal than in formal genres. van der Wouden (2001) showed that reality may be somewhat more complicated than that, in the sense that not all particles are equal in this respect. Miller & Weinert (1998:7) observe a major split in their speakers between those that heavily use the discourse marker *like* and those that don’t (cf. also Fleischman (1999) on *like*). We have the impression that Dutch final *of zo* closely parallels certain usages of English *like*. Table (26) offers some counts of particle-like things.

(26)

	interview	kamer	radio	spontaan
words	45502	13850	10144	51972
# <i>alleen</i> ‘only’	30	12	9	47
# <i>alleen</i> /Kword	0,66	0,87	0,88	0,90
# <i>toch</i> ‘yet’	167	49	34	200
# <i>toch</i> /Kword	3,7	3,5	3,4	3,9
# <i>of zo</i> ‘like’	30	0	18	80
# <i>of zo</i> /Kword	0,66	0,0	1,8	1,5
# <i>omdat</i> ‘because, since’	48	18	8	48
# <i>omdat</i> /Kword	1,1	1,3	1,9	1,1
# <i>want</i> ‘because, since’	103	29	19	204
# <i>want</i> /Kword	2,3	2,1	1,9	3,9
# <i>wel</i> PART	435	56	47	521
# <i>wel</i> /Kword	9,6	4,0	4,6	10

We observe that the least formal genres score highest with the modal particle *wel*, which is what we expect. In the case of the contrastive *toch*, however, we hardly find any difference. And with *of zo* the picture is really strange: parliamentary speech ranks lowest, which is what we expect, but the subcorpus scoring highest is radio, which is completely unexpected on the basis of earlier calculations where we found it to be rather formal.

The last counts we want to present here involve personal pronouns. The table in (27) gives the numbers. To quote Biber (1988:225), “first person pronouns have been treated as markers of ego-involvement in a text. They indicate an interpersonal focus and a generally involved style. [...] Numerous studies have used first person pronouns for comparison of spoken and written registers.” We therefore expect that the most formal subcorpora to have the least first person pronouns. Surprisingly, this expectation is borne out for the radio corpus, but not for the parliamentary speeches.¹¹

¹¹One might think that almost all first person pronouns in this subcorpus would be plural (*modestiae* or *maiestatis* or whatever) but that is not the case: over 300 cases are singular.

(27)

More quantitative properties of four subcorpora of CGN (N= 121468)				
	interview	kamer	radio	spontaan
words	45502	13850	10144	51972
personal pronouns	3704	846	392	4894
pers.pron/Kword	81	61	38	94
pronoun 1st person	1755	477	121	1892
1st person/Kword	39	34	12	36
pronoun 2nd person	765	113	47	1357
2nd person/Kword	17	8	5	26
pronoun 3rd person	1184	256	226	1645
3rd person/Kword	26	18	22	32

Again according to Biber (1988:225), “second person pronouns require a specific addressee indicate a high degree of involvement with that addressee”. The informal subcorpora have more of them than the more formal ones, and the same holds for the third person pronouns. We leave the rest of the interpretation of these results for further research.

4 Concluding remarks

We have tried to show how the CGN can be used to learn things about register variation in Dutch we didn't know before. Much more can be looked at, of course: Biber (1988) counts not less than 67 variables. Some off the searches are still quite tedious, but that will improve, we hope, with the further development with the CGN exploration tool COREX.

It goes without saying that statistics should be used to assess the validity of the findings presented, but we leave that for another occasion. We still hope, however, that this corpus can be a valuable tool for research both into the properties of spoken Dutch in general and into register variation within.

References

- BAAYEN, R. HARALD. 1989. *A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation*. Vrije Universiteit Amsterdam dissertation.
- BIBER, DOUGLAS. 1988. *Variation across speech and writing*. Cambridge [etc.]: Cambridge University Press.
- BRANDT CORSTIUS, HUGO. 1970. *Exercises in computational linguistics*. Amsterdam: Universiteit van Amsterdam dissertation.
- CHOMSKY, NOAM. 1977. On WH-movement. In *Formal Syntax*, ed. by P. Culicover, T. Wasow, & A. Akmajian. New York: Academic Press.
- DE HAAN, GER, & KEES TUIJNMAN. 1988. Missing subjects and objects in child grammar. In *Language development*, ed. by Peter Jordens & Josine Lalleman. Dordrecht: Foris.
- DE HAAN, PIETER, & NELLEKE OOSTDIJK. 1994. Clause patterns in modern british english: A corpus-based (quantitative) study. *ICAME Journal* 18, 41–79.

- DE JONG, EVELINE D. (ed.). 1979. *Spreektaal. Woordfrequenties in gesproken Nederlands*. Utrecht: Bohn, Scheltema & Holkema.
- DE VRIES, JELLE. 2001. *Onze Nederlandse Spreektaal*. Den Haag: SDU Uitgevers.
- DE VRIES, WOBBE, 1910. Dymelie. Opmerkingen over syntaxis. Verhandeling behorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1910–1911.
- , 1911. Dymelie. Opmerkingen over syntaxis (vervolg). Verhandeling behorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1911–1912.
- , 1914. De typen der mededeeling. Opmerkingen over syntaxis. In: Verhandeling behorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1914–1915.
- DICK, FREDERIC, & JEFFREY L. ELMAN. 2001. The frequency of major sentence types over discourse levels: A corpus analysis. *CRL Newsletter* 13. <http://crl.ucsd.edu/newsletter>.
- FLEISCHMAN, SUZANNE. 1999. Pragmatic markers in comparative and historical perspective: theoretical implications of a case study. Paper delivered at the Fourteenth International Conference on Historical Linguistics, Vancouver, BC, August 1999.
- GRONDELAERS, STEFAN, KATRIEN DEYGERS, HILDE VAN AKEN, VICKY VAN DEN HEEDE, & DIRK SPEELMAN. 2000. Het CONDIV-corpus geschreven Nederlands. *Nederlandse Taalkunde* 5, 356–363.
- HAESERYN, WALTER, & OTHERS (eds.). 1997. *Algemene Nederlandse Spraakkunst*. Groningen and Deurne: Martinus Nijhoff and Wolters Plantijn. 2e, geheel herz. dr.
- HOEKSTRA, HELEEN, MICHAEL MOORTGAT, INEKE SCHUURMAN, & TON VAN DER WOUDE. 2000. Syntactic annotation for the spoken dutch corpus project (cgn). paper delivered at Computational Linguistics in the Netherlands, Tilburg, November 2000, submitted for the Proceedings.
- HUANG, J. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15, 531–74.
- HUESKEN, NICOLE. 2001. Mirrorsentences. Repetition of inflected verb and subject in Spoken Dutch. Master's thesis, Utrecht University, General Linguistics. www.let.uu.nl/~Nicole.Huesken/personal/scriptie/scriptie.pdf.
- JANSEN, FRANK. 1981. *Syntaktische constructies in gesproken taal*. Leiden dissertation.
- KOSTER, JAN. 1978. *Locality principles in syntax*. Dordrecht: Foris.
- MILLER, JIM, & REGINA WEINERT. 1998. *Spontaneous spoken speech. Syntax and Discourse*. Oxford: Clarendon.
- NEELEMAN, AD. 1994. *Complex Predicates*. Utrecht dissertation.
- OOSTDIJK, NELLEKE. 2000a. Building a corpus of spoken Dutch. In *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*, ed. by Paola Monachesi, 147–157. Utrecht: Utrecht University, Utrecht Institute of Linguistics OTS.
- , 2000b. The Spoken Dutch Corpus. Overview and first evaluation. Proceedings LREC 2000.
- OVERDIEP, G.S. 1937. *Stilistische grammatica van het moderne Nederlandsch*. Zwolle: Tjeenk Willink.

- SASSEN, ALBERT. 1967. Syntactische implicaties van de zgn. herhalingsconstructie (*dat is een gek geval is dat*). *Handelingen van het Vlaams Filologencongres XXVI*, 300–47.
- STOETT, F.A. 1923. *Middelnederlandsche spraakkunst. Syntaxis*. 's-Gravenhage: Martinus Nijhoff. Derde herziene druk, vijfde oplage, 1977.
- UIJLINGS, B.J. 1956. *Praat op heterdaad*. Assen: Van Gorcum [etc.]. ook als diss. Utrecht onder de titel “Syntactische verschijnselen bij onvoorbereid spreken”.
- UIT DEN BOOGAART, P.C. (ed.). 1975. *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.
- VAN BERCKEL, J.A.TH.M., & OTHERS. 1965. *Formal properties of newspaper Dutch*. Amsterdam: Mathematisch Centrum.
- VAN CRAENENBROECK, JEROEN. 2000. Complementerend van. Een voorbeeld van syntactische variatie in het Nederlands. *Nederlandse Taalkunde* 5, 133–163.
- VAN DER HORST, JOOP, & KEES VAN DER HORST. 1999. *Geschiedenis van het Nederlands in de twintigste eeuw*. Den Haag/Antwerpen: Sdu/Standaard.
- VAN DER WOUDE, TON. 2001. Partikels: naar een partikelwoordenboek voor het Nederlands. lezing TIN-dag, Utrecht, 3 februari 2001, te verschijnen in *Nederlandse Taalkunde*.
- , & HELEEN HOEKSTRA. 2001. Spreektaalconstructies. lezing TABU-dag, 22 juni 2001.
- VAN EYNDE, FRANK, JAKUB ZAVREL, & WALTER DAELEMANS. 2000. Lemmatisation and morphosyntactic annotation for the spoken dutch corpus. In *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*, ed. by Paola Monachesi, 53–62. Utrecht: Utrecht University, Utrecht Institute of Linguistics OTS.
- VERDAM, J. 1923. *Uit de geschiedenis van de Nederlandsche taal*. Zutphen: . 4e dr.
- WEERMAN, FRED. 1992. *The V2 Conspiracy: A synchronic and a diachronic analysis of verbal positions in Germanic languages*. Mouton de Gruyter. Diss. Utrecht 1989.