Harvesting Dutch Trees: Syntactic Properties of Spoken Dutch

Ton van der Wouden, Ineke Schuurman, Machteld Schouppe, Heleen Hoekstra

CCL, K.U. Leuven and UiL-OTS, Utrecht University

Abstract

In this paper, we report on quantitative research into certain word order phenomena in Dutch. In our research, we use the Spoken Dutch Corpus (CGN), a major new resource for research into contemporary spoken Dutch. After briefly introducing the primary data, the annotations added, and some of the tools to explore the primary data and the annotations, we illustrate how the Corpus may be utilized to answer certain linguistic questions concerning the Dutch language.

1 Introduction

Theoretically, the word order in main clauses in languages such as Dutch is relatively free. In practice, however, not all word orders that can happen will also actually occur, or at least not very often. This paper seeks to investigate in a quantitative way some of the peculiarities of Dutch word order. But we start by introducing the corpus and some of the tools to explore it.

2 The CGN

The aim of the Spoken Dutch Corpus project (abbreviated as CGN, from the Dutch name *Corpus Gesproken Nederlands*) is to build an annotated corpus of about one thousand hours of continuous speech, which amounts to 10 million words. It is a collaborative effort of several Dutch and Flemish universities (Oostdijk 2000a, 2000b). The project started in June 1998, and runs for five years.[1]

The corpus is intended as a major resource both for linguistic research and for language and speech technology. To serve this dual purpose, it contains materials recorded in a variety of communicative settings: spontaneous face-to-face and telephone dialogues, interviews, debates, lectures, news broadcasts and book passages read aloud. Two-thirds of the material is collected in the Netherlands, one third in the Dutch speaking part of Belgium. Upon completion, the corpus will be the largest and most diverse database of spoken Dutch collected so far.[2]

The project adds various levels of annotation to the primary speech data. The complete corpus is orthographically transcribed (Goedertier et al. 2000); all words

[1] This publication was supported by the project "Spoken Dutch Corpus" (CGN-project) which is funded by the Netherlands Organisation for Scientific Research (NWO) and the Flemish Government. Thanks are also due to Lisa Cheng, Norbert Corver, Crit Cremers, Helen de Hoop, Frank Jansen, Bob Kirsner, Michael Moortgat, Bram Renmans, Maaike Schoorlemmer, Rint Sybesma and Arie Verhagen for discussion; the usual disclaimers apply.

[2] The CGN is built under the auspices of the Nederlandse Taalunie (lit. Dutch Language Union), which is an intergovernmental organization of the Netherlands and Flanders; distribution of the corpus is done by the Evaluations and Language resources Distribution Agency ELDA (Paris).

receive a (contextually disambiguated) part-of-speech (POS) tag (Van Eynde et al. 2000). In addition, broad phonetic transcription (Demuynck et al. 2002) and syntactic annotation (Hoekstra et al. 2001, van der Wouden et al. 2002, Schuurman et al. 2003) are provided for a representative selection of 10 percent of the data. A selection of 250,000 words receives a prosodic annotation (Buhmann et al. 2002).

In order to yield a maximally consistent result in the time allotted, many of the annotation tasks are carried out (semi-)automatically with the help of tools that are developed for the purpose or taken from elsewhere. Transcriptions and annotations try to adhere to international standards (setting such standards if necessary (Salverda et al. 2001)), rather than following the most recent theories: the goal of the corpus is to serve as many users from as many backgrounds possible (Hoekstra et al. 2001).

3 Tools to Explore the CGN

3.1 COREX

Building a corpus such as CGN with many automated procedures, checks and warrants to guarantee optimal data quality is one thing, users of the CGN of course will want to utilize and explore the data as well. As the data are presented in a number of formats, one can use one's own tools and programming languages to extract the information that he/she is looking for. For example, to get information on collocations one can use standard packages such as the Ngram Statistics Package, WordSmith and MonoConc.[3] To derive word frequency lists one can write one's own script in one's favorite programming language, and one is of course free to relate information from the various annotation levels at one's choice.

Not every linguist, however, is a skilled programmer, and linking the information from the various annotation levels can be quite difficult. To fulfill at least a subset of the prospective users's possible wishes, CGN comes with a specially tailored exploration tool called COREX (for CORpus EXploration), developed by the technical staff of the Max Planck Institute for Psycholinguistics. The Corex program allows one to listen to, to view and to analyse the corpus. It supports the following features (Kilpatrick and Hellwig 2002):

- easy navigation to sub-parts of the corpus, based either on predefined groupings (sex and age of the speaker, the region where (s)he grew up, the text type) or on user-defined groupings (for search purposes or as search results),
- display of synchronized audio and annotation data,

[3] Van der Wouden (2001, 2002) demonstrates the usefulness of the CGN for investigating collocational effects in function words. The Ngram Statistics Package was developed by Ted Pedersen and Satanjeev Banerjee of the University of Minnesota, Duluth. It is a library of routines written in Perl allowing the researcher to efficiently extract bigrams from a corpus and apply various statistical metrics on these bigrams, among other things. The program is free software under the terms of the GNU General Public License; the code can be found at the internet address `http://www.d.umn.edu/~tpederse/code.html`. The WordSmith tools were developed by Mike Scott, available via `http://www.oup.co.uk`. MonoConc is by Steve Barlow, `http://www.ruf.rice.edu/~barlow/mono.html`.

- display, search and statistical analysis of annotation data,
- display and search of metadata descriptions (i.e. information about the kind of data contained in the corpus, such as information about the speakers).

To give just one example, COREX allows one to investigate the geographical and sociolinguistic spread of the prefix *kei* (lit. 'boulder, stone'), which used to combine only with the adjective *hard* 'hard, fast, loud', but can nowadays be found with other adjectives too, as in *keigoed* 'very good', *keileuk* 'very nice' and *keiveel* 'very much/many'; moreover, COREX makes it possible to listen to the various instances directly (cf. (Oostdijk and Broeder 2003) for details and more examples).

3.2 Other Tools

Searching syntactically annotated corpora is a non-trivial task. To be able to fully explore such corpora, the researcher needs to be able to cast his or her queries in terms of abstract syntactic structures. The COREX tool does not allow for such queries yet. Annotate, the tool used to annotate the corpus sentences syntactically, is a development tool rather than an exploration tool, hardly allowing for any interesting queries.[4] Richard Moot has built a special purpose tool called Portray to visualize the CGN syntactic trees, but that does not have query possibilities either.[5]

However, both for corpus explorations and consistency checks, the need for such a tool grew more and more. And rather than re-inventing the wheel, we have chosen to adopt TIGERSearch, a well-established specialized search engine for syntactically annotated corpora, developed at the Institut für Maschinelle Sprachverarbeitung at Stuttgart University (Lezius et al. 2002).[6] TIGERSearch has been developed in the context of the TIGER Project, whose aim is to construct a German newspaper corpus of ca. 55,000 syntactically annotated sentences. Apart from the difference in language (German vs. Dutch) and type of language (newspaper vs. spoken language), the Tiger Project and the CGN have a lot in common. A common trait that is particularly important in this respect is the strategy for syntactic annotation: both projects have borrowed ideas and tools from the Saarbrücken NEGRA project, both projects use the Annotate tool in the semi-automatic annotation process. This made adoption of the TIGERSearch tool a rather trivial matter.

4 Exploring the CGN

In this section, we present some first results of exploration of the CGN. We restrict ourselves to syntactic aspects of Dutch.

[4] See Plaehn (1998) and `http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html`.

[5] `http://www.let.uu.nl/~Richard.Moot/personal/cgn/portray.html`.

[6] Cf. also `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/`. Version 2 was announced for May 2003, but for this paper, we used version 1.01. Large parts of TIGERSearch's functionality will be integrated in the next version of the COREX tool.

4.1 Who's On First

Theoretically, the word order in main clauses in languages such as Dutch is relatively free (Haeseryn et al. 1997): details aside, much is possible, at least in principle, as long as the finite verb occupies the second position in the sentence. The following variants differ pragmatically, but they are all well-formed syntactically:[7]

(1)	*Jouw lot*	*kan*	*mij*	*vandaag*	*niets*	*schelen.*
	Your fate (SU)	can (FIN)	me (IO)	today	nothing (DO)	care (INF)

'I don't care about your fate today'

(2) *Niets kan jouw lot mij vandaag schelen.*

(3) *Mij kan jouw lot vandaag niets schelen.*

(4) *Vandaag kan jouw lot mij niets schelen.*

(5) *Schelen kan jouw lot mij vandaag niets.*

Restricting ourselves to the first position in the sentence, we see that it may be filled by the subject (1), the direct object (2), the indirect object (3), a sentence modifier (4), and an infinitival verbal complement (5), and this doesn't exhaust the possibilities (cf. below).

In practice, however, not everything that can happen will also actually occur, or at least not very often. In the unmarked case, according to the grammar books, sentences have a subject and that subject is in first position, as in (1). Other word orders are seen as marked ('inversion').[8]

In this section, we investigate the corpus in order to see whether this standard presentation is reasonable. The TIGER query in (6) looks for all main clauses with a finite verb and a subject:[9]

(6) `#n1:[cat="SMAIN"] > SU#n2 & #n1 > HD#n3`

This yields 16,505 for the Netherlands part of the Corpus (Release 6, as of Fall 2002), and 17,981 for the Belgian part.[10]

[7] This part of the paper is inspired by an e-mail question from Gisbert Fanselow.

[8] Dutch main clauses may, under certain circumstances, come without a subject. Two cases can be distinguished, viz. impersonal passives on the one hand, e.g. *in het stadion wordt gevoetbald* 'in the stadium is (being) soccer-played' 'they are playing soccer in the stadium' (Bennis 1986), and subject drop as a special instance of 'topic drop' (Jansen 1981, Ch. 5), e.g. *ben even bier halen* 'am for-a-moment beer fetch' 'I am out to get some beer'. Cf. also note 11.

[9] The query says: look for a node n1 which is SMAIN (main clause) directly dominating a node n2 which is SU (subject) where that node n1 also directly dominates a node n3 which is HD (verbal head).

[10] Note that all numbers in this paper should be handled with extreme care, as we are still in the process of building up the corpus; the composition of the various subcorpora is not necessarily completely comparable yet. The stronger trends below, however, may be expected to be robust.

A little extension of the query in (6) renders all and only main clauses with the subject[11] in first position.[12]

(7)
```
#n1:[cat="SMAIN"] > SU#n2 & #n1 > HD#n3 &
#n2 . #n3
```

In both the northern and the southern parts of the corpus, over 50% of the main clauses containing a subject turn out to have a subject in first position, which corroborates the traditional idea that main clauses with a subject in first position are the unmarked case.[13]

Along the same lines, we may search for main clauses in which the constituent in first position has another syntactic function than subject. Below, we give examples from the corpus involving a sentential modifier (8), a direct object (9), a dummy subject (10) (Bennis 1986) and a locative argument (11) in first position:[14]

(8) ***misschien*** *gaan we dan wel weer zo door*
perhaps go we PART PART PART PART through
'we might continue that way'

(9) ***dat*** *weten de vrouwen die hier zitten heel goed*
that know the women that here sit very well
'the women here are very well aware of that'

(10) ***er*** *zijn plannen gemaakt*
there are plans made
'plans have been made'

(11) ***daar*** *zetten we 'm nu ook wel vaak*
there put we him now PART PART often
'these days, we often tend to put him over there as well'

Below we give an overview of number of occurrences in first position of the most important main clause functions distinguished in the syntactic annotation.

Before discussing the numbers presented there, however, we have to discuss a complication. Consider a real life sentence such as (12):

[11] If we were more precise (cf. also note 8), we would write "one of the subjects", as spoken Dutch also allows for sentences with more than one subject (Jansen 1981, Ch. 7), e.g. *ik ben eigenlijk ben ik docente Frans* 'I am actually am I teacher (of) French'. This appears to be a real construction (with a special rhetorical function) rather than a performance error (Huesken 2001, van der Wouden et al. 2002). All in all, the Netherlands part of the Corpus contains 274 main clauses with more than one subject (immediately dominated by the the same main clause node), and the Belgian part 245. This latter number also covers a few cases of (dialectal) clitic doubling, as in *'k ga 'k ik 'n keer gaan* 'I-SU go-FIN I-SU I-SU a time go-INF' 'I'm going to leave now'.

[12] The last part of the query says that node n2 should immediately precede node n3.

[13] Jansen (1981) and Zwart (1993) discuss additional arguments that the first position in the sentence is the unmarked one for the subject, e.g. the fact that the only unstressed pronominal clitics that may occupy this position are subject clitics.

[14] Untranslatable modal particles etc. are glossed PART.

(12) *daar heeft men een een woord voor bedacht*
there has one a a word for be-thought
'one has invented a a word for that'

The repetition of the article *een* is, just like other performance errors, left out of consideration in the syntactic annotation, so that is not a problem. *Daar* 'there', however, the word in the first position of the sentence, is part of a constituent *daarvoor* that is a daughter of the main clause node; the rest of the prepositional phrase, *voor*, is close to the verb. In the syntactic annotation of CGN, this type of structures is analyzed by means of crossing branches (see figure 1).[15]

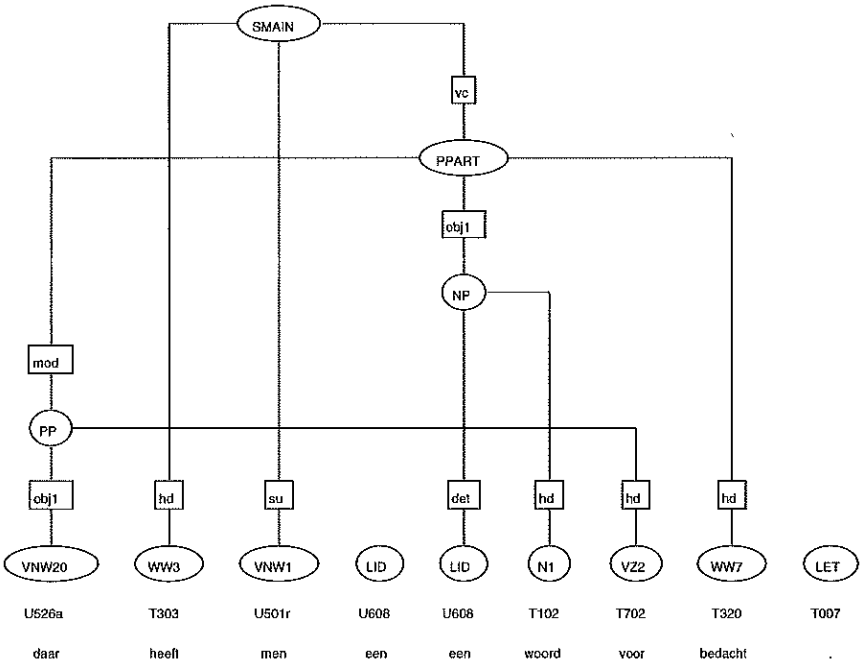

Figure 1: Analysis of example sentence (12)

It goes without saying that this is only one way of analyzing this type of phenomenon; at least since van Riemsdijk (1978) there is also a tradition of deriving this kind of word order via (cyclic) movement. However, the CGN has chosen not to use movement, traces, etc. (cf. (Moortgat et al. 2002, Hoekstra et al. 2001) for discussion). Given this choice, the problem is how to characterize the element in the first position of the sentence. As the TIGERSearch tool considers this word, part of a sentential modifier, to be a sentential modifier, we do so as well for the

[15] Example analyses are presented in Portray format.

Table 1: Preposed constituents in Dutch main clauses

	Netherlands			Belgium		
sent. part	total	fronted	percentage	total	fronted	percentage
SUP	562	372	66%	835	538	64%
SU	16505	8925	54%	17981	10615	59%
MOD	17525	2579	15%	17306	2173	13%
OBJ1	4640	665	14%	4648	401	9%
PC	604	84	14%	805	42	5%
LD	1218	166	14%	1106	114	10%
VC	4399	460	10%	5551	345	6%
POBJ1	48	2	4%	60	5	8%
PREDC	3765	96	3%	4011	60	1%
OBJ2	150	2	1%	152	4	3%
PREDM	397	4	1%	294	6	2%
SVP	990	0	0%	1060	3	0%
SE	87	0	0%	156	0	0%

purpose of this paper. This implies that what is counted as a preposed constituent in the table below is not necessarily a complete constituent.

Table 1 gives an overview of the results for the most important main clause functions distinguished in the syntactic annotation.[16]

The general conclusion to be drawn from this table seems to be, that subjects and dummy subjects are indeed the clause parts with the strongest preference for occupying the first position in Dutch main clauses. Sentence modifiers, direct object, inherent locative objects and material from the verbal complement are sometimes found in this position, and other clause parts are extremely rare there. In this respect, no interesting differences seem to exist between the northern and southern variants of the language.[17]

In subsequent research one might want to refine these statistics a little bit. For example, according to Jansen (1981), direct or indirect objects occur in first position more easily if they are pronominal, animate objects occur more often in first position than inanimate ones, etc. It is moreover probable that not all sentential

[16] SUP = dummy subject, SU = subject, MOD = modifier, OBJ1 = direct object, PC = prepositional complement; LD = locative object, VC = verbal complement, POBJ1 = dummy (direct) object, PREDC = predicate complement, OBJ2 = indirect object, PREDM = secondary complement, SVP = verbal particle, SE = reflexive object; cf. Moortgat et al. (2002) and Hoekstra et al. (2001) for details.

[17] Dutch is far from unique in having a preference for subjects in first position (Bakker 1994), and many principles have been proposed for this cross-linguistic tendency, e.g. in terms of topic and comment structure (Li 1976), of properties of the human parser (Hawkins 1994, Gibson 1998), etc.

modifiers will be equal in this respect either.[18]

Observations such as the ones sketched above are not entirely new (Jansen 1981), although they were only seldom based on such a large corpus of data.

4.2 Long Distance Preposing

So far, we have only looked at (material from) constituents in the first position of the clause that were immediate daughters of the sentence node. In principle, however, the 'original position' may be deeper as well. Consider the corpus sentence in (13).

(13) *bebossing* *heb* *ik* *al* *vermeld*
afforestation have I already mentioned
'I already mentioned afforestation'

The CGN syntactic analysis of this sentence is as in figure 2. That is to say, *bebossing* 'afforestation' is analyzed as dependent on the past participle *vermeld* 'mentioned' rather than of the finite (auxiliary) verb *heb* 'have'.[19]

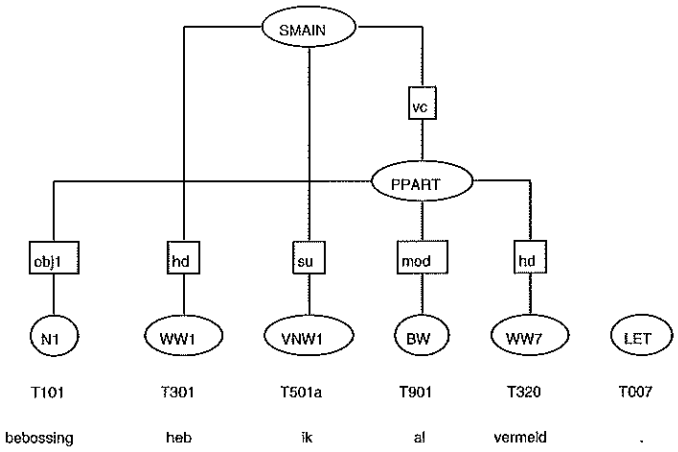

Figure 2: Analysis of example sentence (13)

TigerSearch allows us to systematically search for embedded constituents that end up before the main clause finite verb, i.e. in the first position of the sentence. Some results are given in table 2.

[18] As a reviewer suggested, one may also want to investigate whether special intonation or stress patterns are associated with sentences in which the subject does not occupy the first position, but that would better wait until the various annotation levels have been integrated more tightly.

[19] Note that this is not the only possibility: in a HPSG-like analysis, which opts for 'flatter' structures, the past participle *vermeld* and the object *bebossing* would be sisters of the main verb *heb*.

Table 2: Preposed 'deep' constituents

	Netherlands			Belgium		
sent. part	total	fronted	percentage	total	fronted	percentage
deep OBJ1	3987	309	8%	4800	230	5%
deep VC	2147	57	3%	2912	58	2%
deep LD	882	33	4%	1002	22	2%
deep PREDC	999	8	1%	1246	18	1%
deep OBJ2	169	4	2%	263	4	2%
deep POBJ1	22	1	5%	38	4	11%
deep SU	3440	0	0	3951	0	0
deep SUP	177	0	0	244	0	0
deep PREDM	190	0	0	264	0	0
deep SVP	642	0	0	600	0	0

The conclusion to be drawn from this table is that all 'long distance movement' of 'normal constituents' (as opposed to special ones, such as WH-words) is rare in Dutch, albeit that some of these 'movements' are rarer than others.[20]

4.3 Dutch Verb Clusters

Dutch as it is used in Flanders is not completely identical with the language as it is used in the Netherlands, especially not when spoken language is concerned. One finds differences between the northern and southern variant in all areas of the language: vocabulary, pronunciation, morphology, syntax, and probably also pragmatics. We will refer to the standard Dutch language spoken in the Netherlands as the northern variant, and to the language spoken in Flanders as the southern variant.

One syntactical difference that has received a lot of attention in the theoretical literature is the verbal cluster. As in other Germanic V2-languages, nonfinite verbs may form groups of considerable length at the end of the sentence (Haeseryn et al. 1997, 946):

(14) *Ik had je die olifanten graag willen zien laten*
I had you those elephants please want-INF see-INF let-INF
dansen.
dance-INF
'I would have liked to have seen that you had made those elephants dance'

[20] Under a 'raising' analysis of auxiliary verbs, one would find 'long distance movement' of subjects galore, but CGN has chosen not to take this approach: in our analysis of *Jan schijnt ziek te zijn* 'John seems ill to be' 'John seems to be ill', *Jan* is taken to be the subject of the 'raising verb' *schijnen* 'seem' rather than that of the embedded verb *zijn* 'be' (or even of the adjective *ziek*).

In Belgium, but not in the Netherlands, non-verbal material may 'intrude' into this verbal cluster (the example is from (van der Horst and van der Horst 1999, 199, 292)):

(15) *Het had kunnen waar zijn.*
It had can-INF true be-INF
'It could have been true'

Van der Horst and Van der Horst (1999) show that such examples are far from rare in standard northern Dutch before the 20th century, e.g.

(16) *beloovende den volgenden morgen te vijf ure te zullen*
promising the next morning at five hour to will
aanwezig *zijn*
present be
'promising to be present the next morning at five'
(Jacob van Lennep, *Reisdagboek* 1823)

(17) *als ik jaar in jaar uit, een sommetje kan* ***terzij*** *leggen voor*
if I year in year out, a sum-DIM kan aside lay for
den ouden dag
the old day
'if I can put aside a little sum for when I am old'
(Multatuli, *Max Havelaar*, ed. G. Stuiveling, 1949 (1859))

Vanacker (1970) has claimed that this 'intrusion', which has a very 'Belgian flavor' to northern ears, is virtually absent in various Belgian regions, and altogether less frequent than much of the literature suggests.

The corpus and TIGERSearch make it easy to corroborate at least part of Vanacker's claim: not more than some 8% of the Belgian verb clusters showed some kind of intrusion, whereas the phenomenon was completely absent in the Dutch part of the corpus. A few examples are given below:

(18) *ze zouden er kunnen futuristische stukken in opnemen*
they could there can-INF futuristic pieces in up-take-INF
'they might film futuristic scenes there'

(19) *dat ik eigenlijk toch dringend d'r wat zou moeten*
that I actually PART urgently there something should must
aan doen
on do
'that I really should do something about it immediately'

Checking the corpus for the other part of Vanacker's claim, viz., that the intrusion phenomenon is restricted to certain regions within Belgium, is better postponed until TIGERSearch is integrated in the COREX exploitation tool: only then it will be easy to relate this type of structural queries to the regional and sociolinguistic backgrounds of the speakers.

5 Concluding Remarks

In this paper, we have given an overview of some properties of the Spoken Dutch Corpus (CGN). After that, we have illustrated how this major new resource for research into contemporary spoken Dutch may be utilized to ask linguistic questions concerning the various variants of spoken Dutch.

It will be clear that only the surface of the possibilities has been scratched. One of the results of this paper is corroboration of the standard assumption that in the (quantitatively) unmarked case, subjects occupy the first position of main clauses. Further research may e.g. address the question whether this unmarked word order corresponds with unmarked intonation, but before this question can be asked, the various CGN annotation layers have to be integrated further in the COREX tool.

References

Bakker, D. (1994), *Formal and computational aspects of functional grammar and language typology*, PhD thesis, University of Amsterdam, Amsterdam.

Bennis, H. (1986), *Gaps and dummies*, PhD thesis, University of Tilburg, Tilburg. (published by I.C.G. Printing and Foris).

Buhmann, J., Caspers, J., van Heuven, V., Hoekstra, H., Martens, J. and Swerts, M. (2002), Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus, *in* M. G. Rodríguez and C. P. S. Araujo (eds), *Proceedings of the third International Conference on Language Resources and Evaluation*, ELRA, pp. 779–785.

Demuynck, K., Laureys, T. and Gillis, S. (2002), Automatic generation of phonetic transcriptions for large speech corpora, *Proceedings International Conference on Spoken Language Processing*, Vol. 1, Denver, pp. 333–336.

Gibson, E. (1998), Linguistic complexity: locality of syntactic dependencies, *Cognition* **68** (1), 1–76.

Goedertier, W., Goddijn, S. and Martens, J.-P. (2000), Orthographic transcription of the Spoken Dutch Corpus, *Proceedings LREC 2000*, Athens.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and van den Toorn, M. (eds) (1997), *Algemene Nederlandse Spraakkunst*, 2 edn, Nijhoff and Wolters Plantijn, Groningen and Deurne.

Hawkins, J. (1994), *A performance theory of order and constituency*, Cambridge University Press, Cambridge.

Hoekstra, H., Moortgat, M., Schuurman, I. and van der Wouden, T. (2001), Syntactic Annotation for the Spoken Dutch Corpus Project (CGN), *in* W. Daelemans, K. Sima'an, J. Veenstra and J. Zavrel (eds), *Computational Linguistics in the Netherlands CLIN 2000*, Rodopi, Amsterdam, pp. 73–87.

van der Horst, J. and van der Horst, K. (1999), *Geschiedenis van het Nederlands in de twintigste eeuw*, Sdu Uitgevers and Standaard Uitgeverij, Den Haag and Antwerpen.

Huesken, N. (2001), *Mirrorsentences. Repetition of inflected verb and sub-*

ject in Spoken Dutch, Master's thesis, General Linguistics, Utrecht University, Utrecht. `http://www.let.uu.nl/~Nicole.Huesken/personal/scriptie/scriptie.pdf`.

Jansen, F. (1981), *Syntaktische konstrukties in gesproken taal*, PhD thesis, University of Leiden, Leiden.

Kilpatrick, P. and Hellwig, B. (2002), Corpus Gesproken Nederlands (COREX), manual, version 5. CGN-CD.

Lezius, W., Biesinger, H. and Gerstenberger, C. (2002), *TIGERSearch Manual*, IMS, University of Stuttgart, Stuttgart.

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

Moortgat, M., Schuurman, I. and van der Wouden, T. (2002), Syntactische annotatie. Internal working document CGN, Utrecht, version January 2002.

Oostdijk, N. (2000a), Building a corpus of spoken Dutch, *in* P. Monachesi (ed.), *Computational Linguistics in the Netherlands 1999*, Utrecht Institute of Linguistics OTS, Utrecht University, Utrecht, pp. 147–157.

Oostdijk, N. (2000b), The Spoken Dutch Corpus. Overview and first evaluation, *in* M. Gavralidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer (eds), *Proceedings of the second International Conference on Language Resources and Evaluation*, Vol. 2, ELRA, Athens, pp. 887–893.

Oostdijk, N. and Broeder, D. (2003), The Spoken Dutch Corpus and its exploitation environment, *in* A. Abeillé, S. Hansen-Schirra and H. Uszkoreit (eds), *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, Budapest, pp. 93–100.

Plaehn, O. (1998), Annotate: Bedienungsanleitung. Document Projekt C3, Nebenläufige Grammatische Verarbeitung. Universität des Saarlandes, FR 8.7 Computerlinguistik.

van Riemsdijk, H. (1978), *A Case Study in Syntactic Markedness: The binding nature of prepositional phrases*, Foris Publications, Dordrecht.

Salverda, R., Hajic, J., Bird, S. and Höge, H. (2001), Mid term evaluation Spoken Dutch corpus. Conclusions and recommendations. Evaluation report October 2001, available via `http://lands.let.kun.nl/cgn/home.htm`.

Schuurman, I., Schouppe, M., Hoekstra, H. and van der Wouden, T. (2003), CGN, an annotated corpus of spoken Dutch, *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, Budapest, pp. 101–108.

Van Eynde, F., Zavrel, J. and Daelemans, W. (2000), Lemmatisation and morphosyntactic annotation for the Spoken Dutch Corpus, *in* P. Monachesi (ed.), *Computational Linguistics in the Netherlands 1999*, Utrecht Institute of Linguistics OTS, Utrecht University, Utrecht, pp. 53–62.

Vanacker, V. (1970), Een 'Zuidnederlandse' constructie in een paar Zuidnederlandse dialecten, *De Nieuwe Taalgids* **63**, 140–157.

van der Wouden, T. (2001), Collocational behaviour in non content words, *in* B. Daille and G. Williams (eds), *COLLOCATION: Computational Extraction, Analysis and Exploitation. Proceedings of a Workshop during the 39th*

Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter, Toulouse, pp. 16–23.

van der Wouden, T. (2002), Particle research meets corpus linguistics: on the collocational behavior of particles, To appear in *Belgian Journal of Linguistics* 16, special on particles ed. by Ton van der Wouden, Ad Foolen and Piet Van de Craen, 2002, 151–174.

van der Wouden, T., Hoekstra, H., Moortgat, M., Schuurman, I. and Renmans, B. (2002), Syntactische annotatie voor het Corpus Gesproken Nederlands (CGN), *Nederlandse Taalkunde* **7** (4), 335–352.

Zwart, J.-W. (1993), *Dutch syntax: A minimalist approach*, PhD thesis, University of Groningen, Groningen.