Syntactic Analysis in the Spoken Dutch Corpus (CGN)

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

*Utrecht University, Uil-OTS, Trans 10, 3512 JK Utrecht
{heleen.hoekstra, vdwouden}@let.uu.nl
†University of Leuven, Center for Computational Linguistics, Maria-Theresiastraat 21, 3000 Leuven, Belgium
{bram.renmans,ineke}@ccl.kuleuven.ac.be

Abstract

The paper describes the syntactic annotation of the Spoken Dutch Corpus ("Corpus Gesproken Nederlands" or CGN), the Dutch-Flemish project (1998-2003) aiming at the collection, description and annotation of ten million words of spoken Dutch. In the first part, the background of the parsing strategy is discussed, as well as some details concerning the actual implementation of the parsing process. The second part discusses some examples of practical applications of the result of the parsing process.

1. Introduction

Although Dutch is among the world's best studied languages, very little is known about spoken Dutch, as most linguistic studies deal with written variants of the language. The Spoken Dutch Corpus ("Corpus Gesproken Nederlands" or CGN) is meant to change this. It is a Dutch-Flemish project (1998-2003) aiming at the collection, description and annotation of ten million words of spoken Dutch, two thirds from the Netherlands, one third from the Dutch speaking part of Belgium (Oostdijk, 2000; Oostdijk et al., 2002).[1]

After enriching the speech data with an orthographic transcription, a first layer of linguistic annotation concerns the assignment of base forms and morphosyntactic tags to all words of the corpus (Van Eynde et al., 2000).

A second layer deals with the syntactic analysis. This is carried out for a subcorpus of one million words only, as this type of annotation turns out to be much more time-consuming than e.g. Part-of-Speech-tagging or lemmatisation. The reasons for this difference are threefold. At the beginning of the project,

- there was no such thing as a tool which can automatically parse spoken Dutch sentences with an acceptable degree of quality;
- there was no syntactically annotated corpus of spoken Dutch which could be used as a learning corpus for a statistics based general purpose automatic parser in order to develop such a thing;
- there was no formalised grammar of spoken Dutch: formal analyses of many constructions found in the spoken variants of the language only were simply lacking.

Therefore a considerable amount of time was spent writing a syntactic annotation manual (Moortgat et al., 2002), developing and testing tag sets (Hoekstra et al., 2001), establishing a manually annotated corpus for bootstrapping purposes, etc.

The output of the process of syntactic annotation is a set of dependency trees which are aimed to be as theory neutral as possible (Skut et al., 1997), sticking rather closely to traditional Dutch syntactic analysis as exemplified by the large ANS grammar (Haeseryn and others, 1997). In our view, this is the best way to serve as many potential users as possible: these dependency trees are input to other modules producing data structures useful to users from various theoretical backgrounds and with various practical aims (Hoekstra et al., 2001; Moortgat and Moot, 2001).

This paper consists of two parts. In the first part, some details of the annotation philosophy and the tag sets used are given. In the second part, we discuss some examples of ways in which this annotated corpus can enrich (and has already enriched) our knowledge of and insight into some of the pecularities of spoken Dutch.

2. The annotation process

2.1. Background

Input for the syntactic annotation is a POS-tagged orthographic transcription of the primary sound files. The material is segmented in annotation units. A real life example of such a unit is given in (1).[2]

(1) *ik zal u gaan uitleggen hoe we dat*
I will you go explain how we that
zo'n beetje hebben aangepakt dat probleem.
such-a bit have tackled that problem.
'I will explain to you how we more or less tackled it, that problem'

An example of a POS-tagged unit is shown in Table 1. The leftmost column has the complete sentence in a one word per line manner, the middle column contains the POS-information (main category in caps, features within brackets), the last column has the lexical lemmas.

[1]The research reported on here 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.

[2]For expository purposes, we have picked a short 14-word unit. Real life annotation units are anywhere between one and more than 150 words.

```
 <au id=1 t=0.000 sp=N00052>
ik          VNW(pers,pron,nomin,vol,1,ev)          ik
zal         WW(pv,tgw,ev)                          zullen
u           VNW(pers,pron,nomin,vol,2b,getal)      u
gaan        WW(inf,vrij,zonder)                    gaan
uitleggen   WW(inf,vrij,zonder)                    uitleggen
hoe         BW()                                   hoe
we          VNW(pers,pron,nomin,red,1,mv)          we
dat         VNW(aanw,pron,stan,vol,3o,ev)          dat
zo'n        VNW(aanw,det,stan,prenom,zonder,agr)   zo'n
beetje      N(soort,ev,basis,onz,stan)             beetje
hebben      WW(pv,tgw,mv)                          hebben
aangepakt   WW(vd,vrij,zonder)                     aanpakken
dat         VNW(aanw,det,stan,prenom,zonder,evon)  dat
probleem    N(soort,ev,basis,onz,stan)             probleem
.           LET()                                  .
```

Table 1: POS-tagged input

The first line of the input fragment in Table 1 shows a unique reference to a CGN soundfile. The second line states that the element *ik* ('I') is a pronoun, to wit, a personal pronoun, nominative case, non-reduced form, first person singular, the lexicon entry form of which is *ik*, etc.[3]

The actual POS-tagging is done in a way comparable to the syntactic annotation, viz., semi-automatically: the output of an ensemble of automatic taggers, using some 300 different morphosyntactic tags and with an accuracy around 95%, is checked and corrected by hand (Van Eynde et al., 2000; Van Eynde, 2001).

Table 2 gives an example of the kind of syntactic analysis that is generated for sentence (1) within the CGN project. It illustrates some prominent features of the CGN annotation:

- The annotation is a dependency structure and not a constituent structure or functional structure. The resulting object is therefore not a classical tree structure, but a graph. In such a graph, branches may cross and daughters may have more than one mother.
- The dependency relations are independent of surface word order and constituency. For example, the verb *uitleggen* 'explain' selects for a direct object (taking the form of an embedded question in this case) marked OBJ1, and an indirect object *u* 'you', marked OBJ2. In the surface string, however, *u* is between the main clause finite verb *zal* 'will' (marked HD) and the auxiliary verb *gaan* 'go', the head of a verbal complement (VC), leading to crossing dependencies in the annotation graph.
- The question word *hoe* 'how' has two mother nodes: it is both the head of the subordinate question WHSUB, and modifier of the participial group PPART, which itself is embedded in that subordinate question. This double function is expressed by the two dependency labels WHD and MOD that connect *hoe* with the mother nodes WHSUB and PPART.[4]
- "Right dislocation" and comparable phenomena are considered not to be part of syntax proper. The discourse relation between the main clause and the "moved" constituent *dat probleem* "that problem" is expressed by grouping these consituents under the label DU (for Discourse Unit) where they are assigned the dependency roles of NUCL and SAT, nucleus and satellite, respectively. If, in a later phase of the annotation process, anaphoric relations are going to be marked as well, a link may be made between the cataphoric pronominal element *dat* in the nucleus component, and the satellite full noun phrase *dat probleem*.
- Interpunction, such as the full stop (which is inserted in the orthographic transcription) on the last line of Table 1, is left out of consideration within the syntactic annotation process.[5]

2.2. Implementation

For the actual annotation process, which is carried out in Leuven (for the Belgian part of the corpus) and Utrecht (for the Dutch part), the syntactic annotation tool Annotate, developed by DFKI Saarbruecken, is used (Plaehn, 1998; Brants, 1999). Developed for the annotation of written German newspaper text, it is now employed for the analysis of (the orthographic transcription of) spoken Dutch text. Therefore a completely different tagset had to be developed, which in its current form consists of 316 morphtags (i.e. the tags used by the POS tagging), 72 wordtags (in fact simplified POS-tags) and some 100 tags for syntactic nodes and edges (Moortgat et al., 2002; Hoekstra et al., 2001).

[3] Apart from question marks, full stops and ellipsis marks (...), no interpunction is added in the orthographic transcription. Note that interpunction does not play a role in the syntactic annotation.

[4] In principle, it is even possible for an element or constituent to fulfill more than two roles.

[5] We do not want to claim that interpunction is irrelevant from a linguistic point of view (Nunberg, 1990), but it is definitely not part of the spoken language.

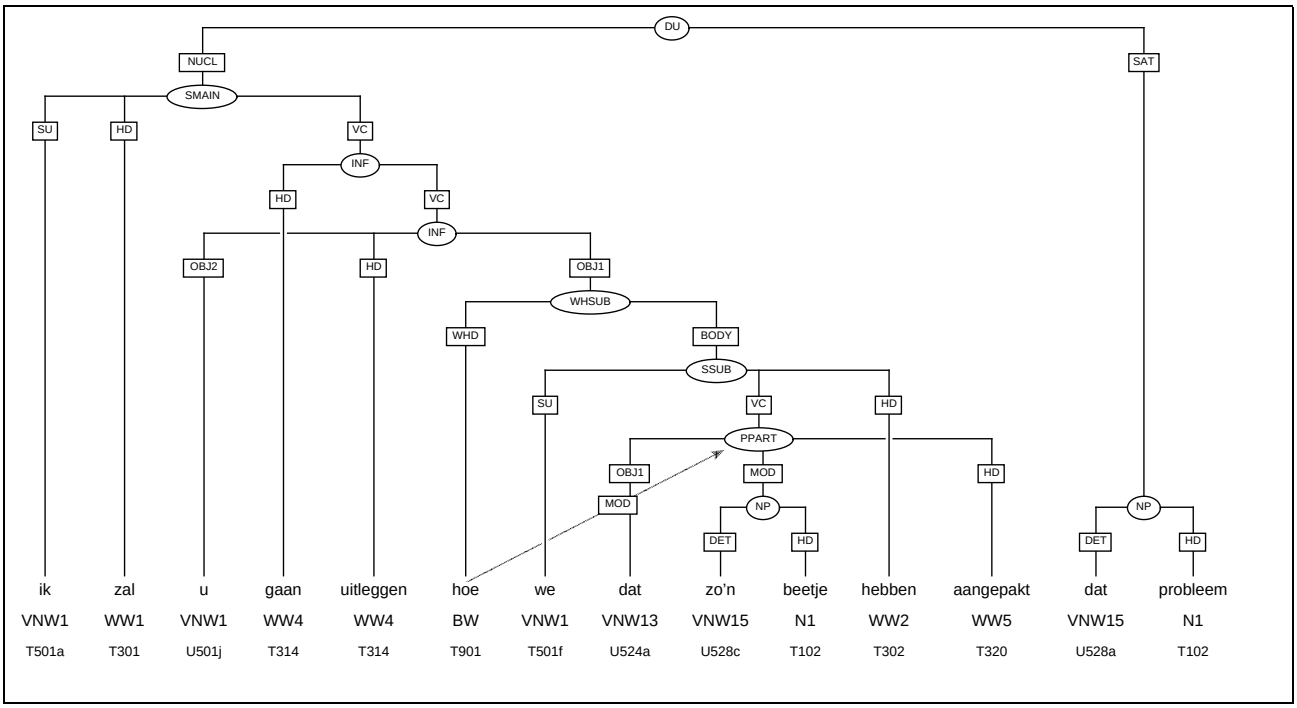

Table 2: Sample analysis

The tags are the same for the Dutch and the Flemish part of the corpus. This makes the resulting annotated corpus of great importance not only to people interested in 'general' or 'standard' Dutch, but also to those wishing to study differences between the variants of Dutch spoken in the Netherlands and Belgium.

The Annotate tools are designed to work together with parsers supporting the manual annotation and running in the background via a defined interface. In this phase of the project, we work with Thorsten Brants's Cascaded Markov Models (CMMs) approach, which supports learning on the basis of an existing annotated corpus (a tree bank) (Brants, 1999). The CMM approach implements a bootstrapping strategy: starting off with a small corpus, using the hypotheses of the parser to gain speed and quality in manually annotating the next part, add this part to the corpus and let the program refine its hypotheses, and so forth. In practice, however, the statistics based parser turned out to be less useful for spontaneous spoken Dutch than was hoped for: a considerable part of the actual parsing is done by hand, and the output is checked and re-checked both with automatic tools and by hand again. In later phases of the project, the CMM approach will be used in combination with other parsers, so that we can integrate the rich information of the lexicon with the statistical approach.

3. Applications

Many areas of the Dutch language are virtually unexplored. For example, grammatical analysis usually deals with written variants of the standard language – the aforementioned large ANS grammar (Haeseryn and others, 1997) is essentially still a prescriptive grammar of the written (van der Wouden, 1998). The Spoken Dutch Corpus is meant to help fill part of this gap by supplying a large body of text for research in this largely unexplored area. In this section, we present some first results.

3.1. Application 1: textual differences between The Netherlands and Belgium

Inspired by (Biber, 1988), among others, we start with trying to find differences between the Dutch and the Flemish subparts. As we do not have any ideas regarding 'typical' or 'standard' values of what we count, we are more concerned with possible differences between interesting subparts of the corpus (cf. (van der Wouden et al., 2002)). However, we expect a direct correlation between the complexity of a text and its formality, that is to say: we expect the more formal texts to have higher average word length, sentence length, degree of sentence embedding, etc., than the less formal texts.

A few results with regard to such standard text properties, based on the 1 October 2001 version of the corpus, are given in Table 3.

The data in Table 3 all point in the same direction, viz., that the Dutch spoken in Belgium is somewhat more complex in terms of average word length, average sentence length and degree of sentence embedding. However, before we jump to the conclusion that on the whole, Belgian spoken Dutch is more formal than Dutch spoken Dutch, we have to check whether the two subcorpora are completely comparable. It turns out that this is not exactly the case: for the time being, the Flemish part contains more material from the more formal text types (in an intuitive sense) than the Dutch part. In due time, this skewedness in the composition of the corpus will of course be corrected.

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
SMAIN	6529	3963	1,65
words/SMAIN	14,2	14,5	0,98
embedded tensed clauses	3523	2391	1,47
embedded tensed/SMAIN	0,54	0,60	0,89

Table 3: Textual differences between The Netherlands and Flanders

3.2. Application 2: textual differences between text types

The corpus metadata allow for other selections from the corpus as well. We therefore proceed and investigate differences between four text types within the corpus, both from the Netherlands and Belgium:

- interviews (with teachers of Dutch);
- parliamentary speeches (recordings of the Dutch "Tweede kamer" and the Flemish "Vlaamse raad");
- radio (various types of broadcasts);
- spontaneous conversation (recorded especially for the CGN).

As a starting hypothesis, we expect parliamentary speeches to be the most formal, spontaneous conversations the least formal, and the other two text types somewhere in between. Some data are given in Table 4 (van der Wouden et al., 2002).

If we assume that more formal texts show greater complexity, the data in (4) may be taken as support for our initial hypothesis. The average sententence length (taken as number of words divided by the number of main clauses) 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, which points in the same direction. Surprisingly, however, the average word length is highest in the radio subcorpus, with parliamentary speeches in second position only.

Perhaps our initial hypothesis should therefore be adjusted just a little bit, in the sense that a subdivision be made between parliamentary speeches and radio recordings on the more formal side of the scale, and interviews and spontaneous conversions on the less formal side.

According to (Biber, 1988, 241), English discourse particles (*well, now, anyway, anyhow, anyways* are mentioned in particular) 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, 2002) argued that the situation in Dutch may be considerably more complex than that, in the sense that not all particles are equal in this respect. For example, of the almost synonymous focus particles *slechts* 'only' and *alleen* 'only', *slechts* is felt to be the more formal word by native speakers, and it is found relatively more often in more formal text types (van der Wouden, 2002).[6]

There is also considerable variation between speakers: e.g. (Miller and Weinert, 1998, 7) observe a major split in their speakers between those that heavily use the discourse marker *like* and those that do not – all within the same text type. We have the impression that Dutch *of zo* (literally 'or so') closely parallels English *like* (cf. also (Fleischman, 1999)) in some of its usages:

(2) *we waren met uh achttien man of zo*
we were with uh eighteen men or so
we were with like eighteen men

Table 5 offers some counts of three particle-like lexical items: the aforementioned *of zo*, the multifunctional modal particle *wel*, "a typical Dutch noise with no particular meaning" (Foolen, 1986), and the *toch*, which has contrastive and modal uses.[7]

(3) *ja dat doe 'k wél.*
yes that do I PART
'yes I do do that' (emphatic use of *wel*)

(4) *dat vind 'k altijd wel leuk.*
that find I always PART nice
'I a always sort of like that' (mitigating use of *wel*)

(5) *en toch moet Borst doorgaan.*
and PART must Borst continue
'and yet Borst has to continue'

(6) *dan is 't toch helemaal niet goed?*
then is it PART completely not good
'then it is terribly wrong, isn't it?'

We observe that according to Table 5, the least formal genres have the highest scores for the modal particle *wel*, which is what we expect. In the case of *toch*, however, we hardly find any difference between the various subcorpora. And the picture of *of zo* is particularly noteworthy: parliamentary speech ranks lowest, as expected, 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.

[6] Numbers on differences between *slechts* and *alleen* are not given in Table 5 because CGN in its current form is still too small for that.

[7] In the table, Kword is short for 1000 words.

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
embedded tensed clauses	1921	809	379	1493
embedded tensed/SMAIN	0,61	1,1	0,57	0,37

Tabel 4: Textual differences between four subcorpora of CGN

More quantitative properties of four subcorpora of CGN (N=121468)								
	interview		parliament		radio		spontaneous	
words	45502		13850		10144		51972	
	N	N/Kword	N	N/Kword	N	N/Kword	N	N/Kword
wel PART	435	9,6	56	4,0	47	4,6	521	10
toch PART	167	3,7	49	3,5	34	3,4	200	3,9
of zo 'like'	30	0,66	0	0	18	1,8	80	1,5

Table 5: Discourse markers in CGN

If anything, these preliminary counts seem to justify the following conclusions:

- not all of Biber's results with respect to English carry over to Dutch;
- not all parameters order bodies of text in the same way on the same scale – which probably means that text variation is a multidimensional phenomenon (cf. Biber);
- CGN, even in its current uncompleted form, is a useful tool for quantative text variation studies.

It goes without saying that statistics could be used to assess the validity of the findings presented, but we leave that for another occasion (cf. e.g. (Grondelaers and Speelman, 2001)).

3.3. Application 3: spoken language phenomena

Spoken Dutch knows a number of constructions that are considered to be unwellformed for written language, which explains why they are seldomly treated in the literature (but cf. (de Vries, 1911; Jansen, 1981; de Vries, 2001), etc.). For example, constituents can be left out for reasons having to do with discourse or performance:

(7) *[Dat] Is goed!*
[That] is good
'agreed' ('topic-drop': sentence without a subject)

(8) *[Dat] doen we!*
[That] do we
'agreed, we'll do that' ('topic-drop': no direct object)

On the other hand, one also meets sentences with more than one instance of subject, verbal head, or other constituents:[8]

(9) *ik ben eigenlijk ben ik docente Frans*
I am actually am I teacher French
'I am a French teacher, actually'

It is not the task of the CGN syntactic annotation group to judge what is 'proper Dutch' and what is not: all utterances in the corpus are considered to be grammatical, in principle (but cf. below), and should therefore get a syntactic analysis. Non-standard sentence types, as exemplified in (7–9), receive a 'standard' annotation. And if necessary, this annotation will have two verbal heads, or two subjects, or whatever – or lack such constituents altogether.

Note that these spoken language construction should be distinguished carefully from errors of speech or performance: if speakers obviously repair or correct (part of) their utterance, the corrected part is left out of the syntactic annotation graph.

This approach of annotating all utterances makes it easy to collect examples of these types of constructions for further research. For example: to find sentences of type (9) one simply has to look for clauses (annotation graphs) containing more than one subject node, verbal head node, etcera.

4. Concluding remarks

In this paper, we have given a brief overview of the Spoken Dutch Corpus CGN. We have discussed details both of the philosophy behind the process of syntactic annotation and of its actual implementation.

After that, we have shown how the CGN can be used to learn things about variation in Dutch we did not know before. Much more can be looked at, of course: (Biber, 1988) counts no less than 67 variables. Some of these searches are still quite tedious, but that will improve with the further development of the CGN exploration tools.

[8](Huesken, 2001) speaks of 'mirror constructions'; one also finds terms such as 'repetition' and 'anacolouthon'.

In the last part, we demonstrated how CGN can be used to investigate constructions found in spoken discourse only.

From all this it will be clear, we hope, that the Spoken Dutch Corpus can be a valuable tool for research both into the properties of spoken Dutch in general and into register variation within the language, among many other things.

5. References

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge [etc.].

Thorsten Brants. 1999. Cascaded Markov Models. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics EACL-99*, Bergen, Norway, 1999.

Wobbe de Vries. 1911. Dysmelie. Opmerkingen over syntaxis (vervolg). Verhandeling behoorende bij het programma van het gymnasium der gemeente Groningen voor het jaar 1911–1912.

Jelle de Vries. 2001. *Onze Nederlandse Spreektaal*. SDU Uitgevers, Den Haag.

Suzanne Fleischman. 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.

Ad Foolen. 1986. 'Typical Dutch noises with no particular meaning': modale partikels als leerprobleem in het onderwijs Nederlands als vreemde taal. In *Verslag van het negende Colloquium van docenten in de Neerlandistiek aan buitenlandse universiteiten*, pages 39–57. IVN, Den Haag.

Stefan Grondelaers and Dirk Speelman. 2001. Werpt het CGN een ander licht op de stratificatie van het Belgische Nederlands? voordracht Over Spraak Gesproken. Gebruikersworkshop van het Corpus Gesproken Nederlands, 21-12-2001, Antwerpen.

Walter Haeseryn et al., editors. 1997. *Algemene Nederlandse Spraakkunst*. Martinus Nijhoff and Wolters Plantijn, Groningen and Deurne. 2e, geheel herz. dr.

Heleen Hoekstra, Michael Moortgat, Ineke Schuurman, and Ton van der Wouden. 2001. Syntactic Annotation for the Spoken Dutch Corpus Project (CGN). In W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel, editors, *Computational Linguistics in the Netherlands 2000*, pages 73–87. Rodopi, Amsterdam/New York.

Nicole Huesken. 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.

Frank Jansen. 1981. *Syntaktische konstrukties in gesproken taal*. Ph.D. thesis, Leiden.

Jim Miller and Regina Weinert. 1998. *Spontaneous spoken speech. Syntax and Discourse*. Clarendon, Oxford.

Michael Moortgat and Richard Moot. 2001. CGN to Grail. extracting a type-logical lexicon from the CGN annotation. In W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel, editors, *Computational Linguistics in the Netherlands 2000*, pages 126–143. Rodopi, Amsterdam/New York.

Michael Moortgat, Ineke Schuurman, and Ton van der Wouden. 2002. Syntactische annotatie. Internal working document CGN, Utrecht, version January 2001.

Geoffrey Nunberg. 1990. *The linguistics of punctuation*. Center for the Study of Language and Information, Menlo Park, Calif. [etc.]. (CSLI lecture notes 18).

Nelleke Oostdijk, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat, and Harald Baayen. 2002. Experiences from the Spoken Dutch Corpus Project. Proceedings LREC 2002.

Nelleke Oostdijk. 2000. The Spoken Dutch Corpus. Overview and first evaluation. In M. Gavralidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of the second International Conference on Language Resources and Evaluation*, pages 887–893. ELRA, Paris.

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

W. Skut, B. Krenn, and H. Uzkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Washington, D.C. available via arxiv.org/format/cmp-lg/9702004.

Ton van der Wouden, Ineke Schuurman, Bram Renmans, Heleen Hoekstra, and Michael Moortgat. 2002. Variation across spoken Dutch. lecture TIN-dag, Utrecht, 26 january, submitted to *Linguistics in the Netherlands* 2002.

Ton van der Wouden. 1998. Dat had niet zo gehoeven: Modaliteit en negatie in de nieuwe ANS. *Nederlandse Taalkunde*, 3(3):237–252.

Ton van der Wouden. 2002. Partikels: naar een partikelwoordenboek voor het Nederlands. *Nederlandse Taalkunde*, 7(1):20–43.

Frank Van Eynde, Jakub Zavrel, and Walter Daelemans. 2000. Lemmatisation and morphosyntactic annotation for the Spoken Dutch Corpus. In Paola Monachesi, editor, *Computational Linguistics in the Netherlands 1999. Selected Papers from the Tenth CLIN Meeting*, pages 53–62. Utrecht University, Utrecht Institute of Linguistics OTS, Utrecht.

Frank Van Eynde. 2001. Part of speech tagging en lemmatisering. Technical report, Centrum voor Computerlinguïstiek K.U. Leuven, http://lands.let.kun.nl/cgn/publicat.htm.