CGN, an annotated corpus of spoken Dutch

Ineke Schuurman and Machteld Schouppe
Centrum voor Computerlinguïstiek
K.U.Leuven (Belgium)
ineke.schuurman,machteld.schouppe@ccl.kuleuven.ac.be
Heleen Hoekstra and Ton van der Wouden
UiL-OTS
Utrecht University (The Netherlands)
Heleen.Hoekstra,Ton.vanderWouden@let.uu.nl

Abstract

Although there are two variants of Dutch, the northern variant being the one used in the Netherlands and the southern variant in Flanders (Belgium), one corpus of spoken Dutch is under construction, the Spoken Dutch Corpus (CGN). In this paper first the principles of this corpus will be discussed, thereafter a few small case studies will show what the merits of such a corpus are.

1 Credits

The authors would like to thank Bram Renmans and Michael Moortgat. 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.

2 Introduction

Dutch is the official language of approximately 21 million speakers: 15 million in the Netherlands and 6 million in Flanders, the northern part of Belgium. Since 1982 an intergovernmental institution, Nederlandse Taalunie (NTU) (lit. Dutch Language Union), is responsible for the language policy in both the Netherlands and Flanders. It supports, amongst other things, projects leading to dictionaries, grammars, and other language resources, and it advises the Dutch and Flemish government on language policy issues (a.o. within the settings of the European Union). In recent years, the NTU has also become interested in the creation of an electronic infrastructure for language in order to strengthen the position of Dutch in the international information society, in which language and speech technology (LST) has become increasingly important. By its very nature, the development of language and speech technology for a language has an important national (or even nationalistic) component, but in the case of Dutch it was the coordinating NTU that decided that the creation of a series of basic, publicly available, language resources of good quality, the so-called BLARK (Basis LAnguage Resources Kit) for Dutch should be stimulated (Cucchiarini and D'Halleweyn, 2002), to be of help in creating LST applications.

Sponsored by the NTU, quite some research has been done with respect to the creation of such a BLARK for Dutch (Bouma and Schuurman, 1998). This has resulted in a list of priorities, formulated by the LST-platform (Daelemans and Strik, 2002). One of the things that was found to be lacking (Bouma and Schuurman, 1998) was a resource for research into spoken language. Dutch descriptive linguistics has mainly focused on written language, while there is as yet hardly any systematic knowledge of the much more evasive spoken form of the language. So far for Dutch only written text corpora are available. But in 1998, work at the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) has started (Oostdijk et al., 2002).

CGN appears to be unique in that it concerns

two completely equivalent subcorpora (both with respect to the design of the corpus and the annotation schemes used), which together do constitute one large corpus. This way many interesting subjects of research with respect to the relation of both variants can be dealt with, whereas at the same time a large corpus of the standard language as such is available as well. In case Flanders and the Netherlands both would have gone their own way in creating a corpus, they would have had to spend more money in order to get a corpus of sufficient size, whereas it would have been more difficult to compare both variants of Dutch.

3 CGN

The CGN project (1998-2003) aims at developing a corpus of approximately 1,000 hours of speech from adult speakers of standard Dutch, that is circa 10 million words: 2/3 of it will be northern standard Dutch and 1/3 southern standard Dutch (cf. the respective numbers of inhabitants). The corpus is to serve as a major resource for Dutch, for use in a number of widely different fields of interest, including linguistics, language and speech technology, and education. Its design must anticipate the various research interests arising from these fields and provide for them, while the different transcriptions and annotations should be as sophisticated as possible given the present state of the art. Moreover, its construction conforms to national and international standards where available, or else follows recommendations and guidelines or adopts best practice as it has emerged from other projects.

All data in the corpus will be orthographically transcribed, lemmatized and annotated with part-of-speech (POS) information. For part of the corpus, additional transcriptions and annotations will be available. Among these is the syntactic annotation of 1 million words. 'Only' 1 million words because this layer of annotation is much more time-consuming than, for example, POS-annotation (cf. below).

3.1 What will be annotated

In speech corpora, orthographic transcription determines to a large extent what will be annotated, for things that are not transcribed cannot be annotated. And the annotation at the level of POS is also of importance for the syntactic annotation.

Orthographic conventions

In the orthographic transcription of CGN, words are spelled the way they occur in the official spelling guide for Dutch (Renkema, 1997) and in case of missing words or obvious mistakes, the way they occur in the Van Dale dictionary (Geerts and Den Boon, 1999) . Sometimes some code will be added:[1]

- Foreign words that are not (yet) part of the Dutch language (i.e. do not occur in either de Woordenlijst Nederlandse Taal (Renkema, 1997) or Van Dale (Geerts and Den Boon, 1999)) will get a code `*v`.
- Words of which the transcriber is not sure will get an `*x`, a word (or a series of words) that are unintelligible will be represented as '*xxx*' or '*ggg*' (the latter in case of giggles etc).
- In case of mispronunciations (be it on purpose or not) a `*u` is added.
- When a word is interrupted, it is marked with an `*a`.

The only punctuation marks used are the full stop (.), the question mark (*?*) and the omission mark (...). A comma, for example, is not used because it turned out to be too problematic to assign it in a consistent way. So-called silent pauses will result in either a full stop or an omission mark, i.e. they will never occur within the sentence.

POS conventions

Most of the words with codes will get a special treatment at the level of POS-tagging.

- Words with `*v` will get a special tag `SPEC(vreemd)`. The tagset is not tailored to suit foreign language, and the proper POS will not even always be known.
- Words with `*x`, or xxx/ggg will get a tag `SPEC(onverst)`.

[1] There are a few more codes, but these are not relevant for the Syntactic Analysis.

- Mostly words with `*u` will be analyzed the way the 'correct' word would have been analysed. Cf.

*probleren**u	*proberen*	(try)
*om-uh-dat**u	*omdat*	(because)

  Sometimes, when it is completely unclear which word was meant, the transcription will get `SPEC(onverst)`.

- Words with `*a` will get a tag `SPEC(afgebr)`.

Syntactic conventions

Words that received the code `*v` (foreign words) will be treated like Dutch words, except when they appear in series. In that case they will be treated as a MWU (multi word unit). Words with `*a` will be neglected, unless the annotator knows for sure which word was to be realized (usually when a very small part of the word is missing).

Disfluencies are dealt with in various ways:

- fillers: whether or not a word appears with an filler like '*uh*' in it (as in "*TV-uh-scherm*" (tv screen)) doesn't matter for SA, as it has the same POS tag as the word without such a filler (in both cases `N(soort,ev,stan)`). A filler as a separate element will be neglected at the level of syntax, i.e. it is not part of the graph assigned to the sentence. Note that this does not mean that the element is deleted.

- speech repairs: only the corrections will be taken into account when constructing the graphs.

- repetitions: only the last occurence will be taken into account. When complete constituents are repeated they will all be constructed up to that level, but only the last one will be part of the graph assigned to the sentence as a whole.

- fresh starts: only the correction will be taken into account.

- silent pauses: see 'orthographic conventions'.

It is, however, not the case that sentences are normalized. Words that do not fit in will not be neglected, even if this leads to 'ungrammatical' sentences. And, unlike for example the Switchboard corpus (Meteer et al., 1995), conjunctions are not left out in order to start a new sentence. Note that this way we may end up with sentences of more than 150 words, and with several subjects and/or finite verbs.

Sometimes even short sentences will end up with two subjects and two finite verbs, for example in the so-called 'spiegelzinnen' (lit. mirror sentences).

ik ben eigenlijk ben ik docente Frans
(lit. I am in fact am I teacher French)

In (Huesken, 2001) ample evidence is given for not considering such sentences as involving a fresh start.

3.2 How will it be annotated

At the time of the first reflections on the syntactic annotation of the CGN,

- there was no 'full' grammar of spoken Dutch available, at least not in a formalised way,

- most grammars describe the nothern standard variant (even for written Dutch), cf. the ANS and also (De Vries, 2001),

- there was no syntactically annotated corpus of Dutch (written nor spoken) available to train a statistics based parser on, and

- there was no adequate (automatical) parser for Dutch available, not even for written Dutch.[2]

Therefore, an annotation scheme and manual had to be developed, which turned out to be a very time-consuming task, especially because many constructions which are common in spoken language will not show up in grammars dealing mainly with written language.

[2] At least not adequate for our purposes: the parser we were looking for had to be theory neutral and to give access to categorial as well as functional information. The Amazon parser (Coppen, 2002) for example doesn't provide functional information.

The resulting Syntactic Annotation is as theory neutral as possible (in order to be broadly usable), sticking rather closely to the ANS (1997), the widely accepted reference grammar for Dutch. The annotation scheme for CGN has developed into a *de facto* standard for syntactic annotation of Dutch, and it is now also used by the Alpino Treebank project (Bouma et al., 2001).

The annotation provides two types of information: categorial information at the level of syntactic constituency, and dependency information to capture the semantic connections between constituents.

The CGN tagset tries to strike a balance between informativeness and practical usability. It uses 25 phrasal category labels and 34 dependency labels. Conciseness is obtained by giving the labels a context-sensitive interpretation. The `MOD` label, for example, denotes adverbial modification in verbal domains, but also adnominal adjuncts in noun phrases. Levels of granularity that are bound to lead to inter-annotator discrepancies (such as the twenty kinds of adverbial phrases distinguished in the ANS grammar) are avoided.

The rich POS tagset (with 316 labels (Van Eynde, 2001) is reduced to some 50 distinctions relevant for the dependency annotation. The reason for doing so is that otherwise, especially in the beginning of the project, it would have been more difficult to train the system (sparse data). The full tags, however, are available as well (via their unique code).

The NEGRA annotation format (Skut et al., 1997) uses data structures expressive enough to naturally encode dependency relations, also where they are at odds with syntactic constituent structure. Formally, the annotation structures are directed acyclic graphs (DAGs) instead of dependency trees (Hoekstra et al., 2001). The vertices are decorated with a syntactic category label: a POS label for the leaves, a phrasal label for the internal nodes. The edges carry dependency labels. They capture the grammatical function of the immediate constituents of a phrase, distinguishing head, complements and adjuncts.

The structures are as flat as possible, i.e. a new hierarchical level will only be introduced when this is induced by a new head, and there are no non-branching nodes.

Special provisions have been made for the annotation of typical spoken language phenomena. The category label `DU` (discourse unit) for example, allows for an articulation in terms of dependency notions such as nucleus versus satellite, tags or discourse links. An overview of the tagset can be found in (Hoekstra et al., 2001), the full annotation manual is to be found in (Moortgat et al., 2002).

The annotation makes full use of the expressivity of DAGs as compared to trees. Discontinuous dependencies result in crossing branches that would be problematic in a conventional syntactic constituent structure format. Allowing items to simultaneously carry multiple dependency roles (like making use of 'secondary edges') results in a simple annotation schema for phenomena that would require 'movement' or similar devices in tree-based theoretical frameworks.

Finally, annotation graphs with disconnected components are useful to provide partial analyses for interrupted phrases, interpolations and the like. The syntactic annotation procedure, which like the POS tagging is performed semi-automatically, uses the interactive annotation environment developed within the German NEGRA project (http://www.coli.uni-sb.de/sfb378/negra-corpus/negra-corpus.html). A simple visualisation tool (Portray) for the annotation graphs is freely available from the Utrecht CGN site (http://cgn.let.uu.nl). In a later phase of the project, the CGN exploitation software (COREX tools) will provide more advanced display and search facilities for the syntactic annotation.

4 Variants of Dutch

Dutch as it is used in Flanders is not completely identical to the language as it is used in the Netherlands, especially not when spoken language is concerned. 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.

Within CGN people are asked to speak 'standard' Dutch. In the Netherlands everybody[3] will interpret this in more or less the same way. But

[3]Only native speakers of Dutch will be involved.

not so in Flanders. Due to historical reasons (Goossens, 2000), (Wils, 2001), there are two tendencies as far as the use of a standard language is concerned. Some people aim at a standard language in Flanders that comes close to (or is even the same as) the standard language used in the Netherlands. This is more or less the language that is used in the newspapers or by the newsreaders of the public broadcasting companies (Hendrickx, 2000).

On the other hand there is a strong tendency towards the use of a daily speech variant that is non-dialectical and can be used everywhere in Flanders, but contains more regional aspects than the first variant (Goossens, 2000). This variant is known as 'Verkavelingsvlaams' (lit. "Allotment Flemish"). This variant is also often used in light entertainment programmes on TV, esp. in soap operas.

Dictionaries like Van Dale and reference grammars like the ANS (Haeseryn et al., 1997) mainly describe the northern variant of Dutch. They do contain several words and constructions only used in the southern variant (esp. the first one), but these will be marked as such (whereas words and constructions that are only used in the northern variant are not specifically marked).

A corpus like CGN is a suitable tool to record the state of affairs at a certain moment in time, and also under which circumstances which variant is used (region, age, education, setting, etc.).[4] At this moment it is far from clear which variant will in the end become the southern variant of the standard language of the future (cf. also (Van Haver, 1989)).

Most variation between nothern and southern standard language shows up with respect to pronunciation, vocabulary, and morphology. Some examples of the latter are given below:

- vocabulary

Nl	Fl	
sinaasappel	*appelsien*	(orange)
stookolie	*mazout*	(fuel oil)
accu	*batterij*	(battery)
lopen	*(te voet) gaan*	(to walk)
rennen, hardlopen	*lopen*	(to run)

Note that the different use of *lopen* may sometimes lead to confusion!

Quite often words and constructions are valid in both variants, but the preferences are different:

- preferences wrt plurals

Nl	Fl	
appels	*appelen*	(apples)
leraren	*leraars*	(teachers)

- preferences wrt past tense

Nl	Fl	
zeiden	*zegden*	(said)

- preferences wrt gender

Nl	Fl	
het/de filter	*de filter*	(the filter)
het gilde	*de gilde*	(the guild)

Moreover: in Flemish there are still three genders: masculine, feminine and neuter, whereas in the Netherlands there are only two genders left: neuter and non-neuter. Therefore in Flanders one will often say when refering to a door "*Ze staat open*" (She is open) whereas in the Netherlands one will use "*Hij staat open*" (He is open).

- preferences wrt. particle verbs

Nl: *dat ze hem* **op** *wilde* **bellen**
Fl: *dat ze hem wilde* **opbellen**
(that she wanted to call him)

5 Further remarks

In order to make (the syntactic part of) CGN accessible for users with various backgrounds, and therefore various wishes with respect to the way the output is presented, the CGN output can be converted into other formats (categorial grammar, showing non-branching trees, showing traces, ...) as well (Moortgat and Moot, 2001). The CGN exploitation software should also allow for interaction with the other layers of annotation.

[4] It may well be the case that certain constructions that are said to be typically Flemish also turn up in the southern part of the Netherlands ("below the rivers"). This part of the country has several cultural and/or religious resemblances with Flanders. On the other hand, influences of education, newspapers etc will be more country-dependent. The way CGN is conceived allows us to look more closely to the language used by people living in this area, i.e. in Noord-Brabant and Limburg.

6 Some (preliminary) results

The CGN corpus is a very powerful means to perform - amongst other things - research concerning the variation between northern standard and southern standard Dutch, which is what we will do in this section.

However, it should be noted here that since the CGN project has not been completed yet, the figures are still somewhat tentative, since the northern part and the southern part of the corpus are in different stages of development.[5] When the project ends, the northern and southern part of the corpus should contain an equal amount of telephone conversations and other spontaneous speech on the one hand, and lectures, speeches and other more prepared speech on the other hand.

The research was carried out with the use of a search tool called TIGERSearch. Developed at the university of Stuttgart, TIGERSearch allows one to query a given corpus by making use of the TIGERSearch language (Lezius et al., 2002). TIGERSearch queries allow one to search for a given structure, specifying dominance and precedence relations, and properties of nodes. The specific (suspected) differences between northern standard and southern standard under consideration will be in the verbal domain.

6.1 Red versus green word order

In Dutch, the combination of a participle and finite verb in a subordinate clause can occur in two word orders: the red and the green order.

Red order: *Ik geloof niet dat hij* **is gekomen**
Green order: *Ik geloof niet dat hij* **gekomen is**

The red order has for a long time been considered to be the better variant, as the green one was considered to be a Germanism. More recently it is stated that both orders are correct, the red order being the common one in written text, the green one in spoken language, cf. the ANS (Haeseryn et al., 1997).

However, our research has shown that in the Netherlands the ratio between red and green order is almost equal as in the northern part of the corpus 292 occurrences of the red order and 286 occurrences of the green one were found. In the southern standard, however, there appears to be a clear preference for the green order. Of the 904 sentences with finite verb and participle, 560 had green order. The 346 remaining sentences had red order.
Thus, in the northern variant the claim that the green order is the predominantly used one in spoken language is falsified[6].

6.2 Infinitive vs. te+infinitive

Another aspect in which the northern and the southern standard differ is the presence or absence of the particle *te* 'to' in front of an infinitive in the verbal cluster. The ANS (Haeseryn et al., 1997) contains a table with (auxiliary) verbs and the form that the accompanying verb takes.

Among the verbs that **obligatorily** take a 'te+infinitive' (an infinitive preceded by the verbal particle *te* (to)) mentioned in this table there are a number that in the southern variant have an **optional** *te*.

For instance

beginnen 'to start'
proberen 'to try'
vergeten 'to forget'

In the ANS such constructions are marked as substandard, regional ones. In the southern variant, however, several instances were found, also in prepared types of speech (news broadcasts, current affairs programmes). An example:

"*en we gaan eruit met beelden van de Etna die vrijdag weer vuur is beginnen spuwen*"
(and we will conclude our broadcast with pictures of the Etna, which has started to erupt again last Friday) (from: De zevende dag, VRT)

Especially the verb *beginnen* often comes with a bare infinitival complement. 16 out of 17 hits are with a bare infinitive instead of the expected '*te* + infinitive'. *Beginnen* also triggers IPP in Dutch (Infinitivus Pro Participio, a construction in which a (bare) infinitive appears instead of a participle

[5] In order to overcome this problem we verified our findings in those parts of the corpus that have not yet been syntactically analysed, using 'grep' and the like.

[6] These findings are in line with what we found in the other parts of the corpus

when it is selected by the temporal auxiliary *hebben* (to have) or *zijn* (to be).). It seems that, by analogy with *hebben* and *zijn*, using a bare infinitival complement has been adopted by other auxiliaries (such as *beginnen*) as well.

"*ik wil weer beginnen zwemmen*"
(I want to start swimming again)

Note that *proberen* (to try) and *vergeten* (to try) trigger the IPP effect as well. In the part of the corpus that has been syntactically analysed at the moment these two verbs prefer a '*te* + infinitive' as verbal complement. A look in the other parts of the corpus shows that also *proberen* has a tendency to show up with a bare infinitive in the southern standard.

It even turns out to be possible to have such construction when the trigger is a finite verb.
"*. . . , dat mensen met regels op mensen hun vingers beginnen kloppen*"
(. . . , that people start to tap on other peoples fingers with a ruler)
"*. . . , dat de mensen beginnen nadenken*"
(. . . , that people start to think)
"*Heb je dat ook dat je namen begint vergeten?*"
(Does it also happen to you that you start forgetting names?)

These latter constructions, however, have so far only been found in more spontaneous speech. The triggering verb is always a plural (whose form is identical to that of an infinitive)

6.3 Om+te+Infinitive vs. om+infinitive

Usually, when an infinitival complement starts with *om* (for), this *om* is to be followed by *te*. Constructions without *te* are considered to be ungrammatical (they are not even mentioned in the grammar books). But in the Flemish part of CGN quite some instances of such constructions are to be found:

"*mooi om zien, hé?*" (nice to look at, isn't it?)
"*dat is belangrijk om weten*" (it is important to know that)

So far these constructions were only found in the more spontaneous part of CGN (telephone conversations and the like). No occurences were found in the northern variant.

6.4 Which temporal auxiliary is to be used?

In Dutch there are two temporal auxiliaries for the perfect tense *hebben* (to have) and *zijn* (to be). Which one is to be used depends on the verb that comes with it:

"*Hij is gevallen*" (He has fallen)
"*Hij heeft gegeten*" (He has eaten)

When more verbs are involved, there are several possibilities. It turns out that in the nothern and the southern standard the choices will not always be the same. It seems that in the northern standard the verb that comes with the temporal auxiliary is decisive, whereas in the southern standard it will often be the main verb. In CGN, sentences like the following are found for the southern standard:

news broadcast VRT:
Fl: *"hoe het ongeluk* **is** *kunnen* **gebeuren***"*
(How the accident could have happened)
meeting Flemish parliament:
Fl: *"hij* **heeft** *komen* **zeggen** *dat . . . "*
(He came and said that . . .)

Although the first sentence is not impossible in the northern standard, the second one is.

Nl: *"hoe het ongeluk* **heeft/is kunnen** *gebeuren"*
Nl: *"hij* **is/*heeft komen** *zeggen dat . . . "*

7 Conclusion

The observations represented in section 6 are not to be found as such in the leading Dutch reference grammar, the ANS (Haeseryn et al., 1997), maybe because the ANS covers mainly the language as it is written and because it represents the northern standard, sometimes mentioning that the situation is different in the southern variant. Shortly there will be a syntacticaly annotated corpus describing spoken language in both the Netherlands and Flanders. Using this Spoken Dutch Corpus a new ref-

erence grammar could (and should) be made. No need to say that CGN could also be profitable for a whole series of other uses.

References

Gosse Bouma and Ineke Schuurman, 1998. *De positie van het Nederlands in Taal- en Spraaktechnologie.*

Gosse Bouma, Gertjan Van Noord, and Robert Malouf. 2001. Alpino: Wide Coverage Computational Analysis of Dutch. In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in the Netherlands CLIN 2000*, pages 45–59. Rodopi, Amsterdam - New York.

Peter-Arno Coppen. 2002. Het geheim van de oude dame. *Nederlandse Taalkunde*, 7(4):312–334.

Catia Cucchiarini and Elisabeth D'Halleweyn. 2002. How to HLT-Enable a Language: The Dutch-Flemish Experience. http://www.hltcentral.org/page-996.0.shtml.

Walter Daelemans and Helmer Strik, 2002. *Het Nederlands in Taal- en Spraaktechnologie: prioriteiten voor basisvoorzieningen.*

Jelle De Vries. 2001. *Onze Nederlandse spreektaal.* Sdu Uitgevers, Den Haag.

Guido Geerts and Ton Den Boon. 1999. *Van Dale. Groot Woordenboek der Nederlandse Taal.* Van Dale Lexicografie, Utrecht - Antwerpen. 3 vol.

Jan Goossens. 2000. De toekomst van het Nederlands in Vlaanderen. *Ons Erfdeel*, 43(1):3–13.

Walter Haeseryn, Kirsten Romijn, Guido Geerts, Jaap De Rooij, and Maarten Van den Toorn. 1997. *Algemene Nederlandse Spraakkunst.* Martinus Nijhoff uitgevers/Wolters Plantyn, Groningen/Deurne. 2 vol.

Ruud Hendrickx. 2000. VRT en het Nederlands in België. http://www.taaldatabanken.com/, link Taalbeleid.

Heleen Hoekstra, Michael Moortgat, Bram Renmans, Ineke Schuurman, and Ton Van der Wouden. 2001. Syntactic Annotation for the Spoken Dutch Corpus Project (CGN). In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in the Netherlands CLIN 2000*, pages 73–87. Rodopi, Amsterdam - New York.

Nicole Huesken. 2001. Mirrorsentences. Repetition of inflected verb and subject in Spoken Dutch. Master's thesis, Algemene Taalwetenschap Universiteit Utrecht.

Wolfgang Lezius, Hannes Biesinger, and Ciprian Gerstenberger, 2002. *TIGERSearch Manual.* IMS, University of Stuttgart.

Marie Meteer et al. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus. Revised by Ann Taylor, June 1995.

Michael Moortgat and Richard Moot. 2001. CGN to Grail: Extracting a Type-logical Lexicon from the CGN Annotation. pages 126–143. Rodopi, Amsterdam - New York.

Michael Moortgat, Ineke Schuurman, and Ton Van der Wouden, 2002. *CGN Syntactische Annotatie*, January.

Nelleke Oostdijk, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat, and Harald Baayen. 2002. Experiences from the Spoken Dutch Project. In Manuel Gonzáles Rodríquez and Carmen Paz Suarez Araujo, editors, *LREC 2002. Third International Conference on Language Resources and Evaluation*, volume I, pages 340–347. Las Palmas de Gran Canaria, Spain. Proceedings.

Jan Renkema. 1997. *Woordenlijst Nederlandse Taal.* Sdu Uitgevers and Standaard Uitgeverij, Den Haag and Antwerpen. Composed by Instituut voor Nederlandse Lexicografie (INL), with an introduction by Jan Renkema.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97).* Washington, DC, USA.

Frank Van Eynde, 2001. *Part of Speech Tagging and Lemmatisering.* Centrum voor Computerlinguïstiek, K.U.Leuven, June. Corpus Gesproken Nederlands.

Jozef Van Haver. 1989. *Noorderman & Zuiderman. Het taalverdriet van Vlaanderen.* Lannoo.

Lode Wils. 2001. *Waarom Vlaanderen Nederlands spreekt.* Davidsfonds, Leuven.