Building a Corpus of Spoken Dutch

Nelleke Oostdijk
University of Nijmegen

Abstract

In this paper the Spoken Dutch Corpus Project is presented, a joint Flemish-Dutch undertaking aimed at the compilation and annotation of a 10-million-word corpus of spoken Dutch.* Upon completion, the corpus will constitute a valuable resource for research in the fields of computational linguistics and language and speech technology. The paper first gives an overview of the project. It then goes on to describe the data that are available in the first release of the first part of the corpus that came out March 1st, 2000.

1 Introduction

In June 1998 the Spoken Dutch Corpus project was started, a five-year project aimed at the compilation and annotation of a 10-million-word corpus of contemporary standard Dutch as spoken in the Netherlands and Flanders. The project is funded jointly by the Flemish and Dutch governments and Science Foundations with a budget of some 4.6 MEuro. The entire corpus will be orthographically transcribed, lemmatized and annotated with part-of-speech information. For a selection of one million words, further, more detailed annotations are envisaged, including an auditorily verified broad phonetic transcription and a syntactic annotation. A selection of 250,000 words will receive a prosodic annotation. To enable effective access to the speech recordings, the transcriptions will be enriched with pointers into the speech files. The automatic time alignment will be manually checked on the word level for that part of the corpus for which a verified phonetic transcription is available.

In section 2 of this paper I describe the project in more detail. In section 3 I focus on the data that are available in the first release of the first part of the corpus that came out on March 1st 2000. The paper concludes by discussing the position of the Spoken Dutch Corpus in the international context.

* This publication was supported by the Netherlands Organization for Scientific Research (NWO) under grant number 014-17-510.

2 The Spoken Dutch Corpus Project

2.1 Background and motivation

Standard Dutch is the official language in the Netherlands (some 15 million people speak northern standard Dutch) and in Flanders (the northern part of Belgium, 5.6 million people speak southern standard Dutch).[1] While variants of the same language, there are considerable differences between northern standard Dutch and southern standard Dutch. These differences occur with regard to syntax, morphology, lexis and phonetics/phonology (*cf.* Donaldson, 1983; Van de Velde *et al.*, 1998).

As one of the smaller languages in Europe, Dutch is under serious threat of gradually disappearing as it is losing ground to English. The availability of the necessary resources has placed the English language and speech technology in the leading position it holds today and has thus further strengthened the position of English for business communication. The fact that to date for Dutch few relevant language resources are available forms a serious complication for the advancement of Dutch language and speech technology (*cf.* Bouma and Schuurman, 1998). The present project seeks to ameliorate this situation.

Apart from the interests held by language and speech technologists, the corpus is intended to serve several other research interests. The corpus addresses the needs of linguists from various backgrounds. So far for Dutch the only more or less substantial data collections derive from written sources. As a consequence, studies of Dutch linguistics in the past have focused on the written language, leaving the spoken language rather poorly documented. Another field in which the corpus will be of significant use is that of education. The insights that can be gained into every-

[1] In addition, Dutch is the official first language in Surinam and the Dutch Antilles. However, since it concerns very small populations (some 360,000 and 240,000 speakers respectively) who use Dutch predominantly in formal settings, these have not been included.

day language use are indispensable for developing Dutch language courses and course materials.

2.2 Project organization

The Spoken Dutch Corpus project is directed by a board whose members include representatives of the two governments, the Dutch Language Union, Dutch and Flemish research foundations and one of the Dutch national research schools (LOT).[2] Chairman of the board is Professor W. Levelt of the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands.

Appointed by the board there is a steering committee consisting of experts from various linguistics (sub)disciplines and expert language and speech technologists, that is responsible for the project's progress and finances.

Project activities are coordinated from two sites: Ghent for Flanders and Nijmegen for the Netherlands. Each site is managed by a project leader. The project leaders in collaboration with three specialist working groups (one for corpus design and compilation, one for signal processing and one for corpus annotation) are responsible for the design and implementation of the various project activities.

2.3 Project outline and timetable

The project aims to compile a 10-million-word corpus that will constitute a plausible sample of contemporary standard Dutch as spoken in Flanders and the Netherlands. One third of the data will be collected in Flanders, two thirds will originate from the Netherlands. The entire corpus will be transcribed orthographically, lemmatized and tagged with part-of-speech information. Users will be able to access the speech recordings through pointers in the transcriptions. For a selection of one million words it is envisaged that an auditorily verified, broad phonetic transcription will be available, while for this part of the corpus the automatic time alignment will be manually checked on the level of the word. For most of the recordings which are not checked by hand the pointers are expected to be accurate within less than 100 ms. Also for one million words a syntactic annotation will be available and 250,000 words will receive a prosodic annotation.

The first year of the project has been devoted to corpus design, the development of various protocols and annotation schemes, and the selection and adaptation of tools and supporting resources. During this year also a 50,000-word pilot corpus was compiled which was used for testing purposes.

Over the remaining four years the corpus will be compiled, transcribed and annotated incrementally in seven six-to-eight-month periods. At the end of each period part of the material will be released. Thus the data will be available to users from an early stage onward, while the project may benefit from the feedback given by these users.

2.4 Project activities

2.4.1 Corpus design

The design of the corpus was guided by a number of considerations. First of all, there is the fact that the corpus must serve many and rather diverse interests. Different user groups have different requirements when it comes to the quality and quantity of the data, the number and type of speakers, and so on. Second, the total budget available for the entire project is fixed at 4.6 MEuro, i.e. this should cover all costs involved in recording and collecting data, transcribing and annotating these data, etc. And finally, the issue of copyright complicates matters. Since the corpus will be distributed including the speech files, the consent of all speakers is required as well of any other parties that have any rights to the recorded material.

The design of the corpus takes into account the various dimensions underlying the variation that can be observed in language use. In the overall design of the corpus the principal parameter is taken to be the socio-situational setting in which language is used. This leads us to distinguish a number of components, each of which can be characterized in terms of its situational characteristics such as communicative goal, medium, number of speakers participating, and the relationship between speaker(s) and hearer(s).

---

[2] The Dutch Language Union was founded in 1980 and is the result of a treaty between Flanders and the Netherlands concerning their language policy. In the case of the Spoken Dutch Corpus it is the Dutch Language Union which holds all rights.

Table 1. Overall design of the corpus

<table>
<tr><td rowspan="7">dialogue / multilogue 8,110,000</td><td colspan="2" rowspan="4">private 6,635,000</td><td rowspan="4">unscripted 6,635,000</td><td rowspan="2">direct 3,460,000</td><td>conversations (face-to-face) 3,000,000</td></tr>
<tr><td>interviews 460,000</td></tr>
<tr><td rowspan="2">distanced 3,175,000</td><td>telephone conversations 3,000,000</td></tr>
<tr><td>business transactions 175,000</td></tr>
<tr><td rowspan="3">public 1,475,000</td><td>broadcast 750,000</td><td colspan="2">more or less scripted 750,000</td><td>interviews and discussions 750,000</td></tr>
<tr><td rowspan="2">non-broadcast 725,000</td><td colspan="2" rowspan="2">unscripted 725,000</td><td>discuss., debates, meetings 375,000</td></tr>
<tr><td>lectures 350,000</td></tr>
<tr><td rowspan="7">monologue 1,890,000</td><td colspan="2">private 40,000</td><td colspan="2">more or less scripted 40,000</td><td>descriptions of pictures 40,000</td></tr>
<tr><td rowspan="6">public 1,850,000</td><td rowspan="4">broadcast 950,000</td><td colspan="2">unscripted 250,000</td><td>spontaneous commentary 250,000</td></tr>
<tr><td colspan="2" rowspan="3">scripted 700,000</td><td>newsreports, current affairs programmes 250,000</td></tr>
<tr><td>news 250,000</td></tr>
<tr><td>commentary 200,000</td></tr>
<tr><td rowspan="2">non-broadcast 900,000</td><td colspan="2" rowspan="2">scripted 900,000</td><td>lectures, speeches 275,000</td></tr>
<tr><td>read aloud text 625,000 (+ 375,000)</td></tr>
</table>

The specification of each of the components is given in terms of sample sizes, total number of speakers, range of topics, etc. Where this is considered to be of particular interest, speaker characteristics such as gender, age, geographical region, and socio-economic class are used as (demographic) sampling criteria; otherwise they are merely recorded as part of the meta-data.[3] The overall design of the corpus is given in Table 1.

In all, 14 different components are distinguished. The total number of words varies from component to component. Since not for all components a full specification is available as yet, the total number of words per component remains at this point somewhat arbitrary. At this time, however, we assume that no adaptations will be necessary. Considerations that have played a role in determining the present sizes of the components are the following:

- there is a great demand for spontaneously spoken language data; this explains the overall bias towards unscripted language;
- interaction is considered to be a typical characteristic of spoken communication; therefore it is felt that dialogues and multilogues should be amply represented in the data;
- certain language varieties display a great deal more variation than others; in order to capture this variation, more heterogeneous components generally are represented in the corpus by a larger number of samples than the more homogeneous ones;
- the sample size differs from component to component; while it is impossible to know what the optimum sample size is, intuitive judgements are brought into play when it comes to deciding what constitutes an appropriate sample. Here the ‘natural’ length

[3] See also Oostdijk (2000).

of a spoken text also plays a role: an item in a radio news broadcast is per definition shorter than the spoken commentary in a television documentary;

- some types of data are easier to collect than others
- in order to meet the needs of particular user groups some components require a certain minimum amount of data; this is especially true for components that are used for the development of technological applications such as the telephone conversations and read aloud text.

Once the overall design of the corpus had been established, it remained to be decided which part(s) of the corpus should be included in the selection of one million words (or 250,000 words in the case of prosodic annotation) for which more advanced annotations are envisaged. Preferably, the selection should in some way reflect the composition of the full corpus. While it would have been straightforward to simply select 10 per cent of each component, there were two powerful arguments that were raised against this procedure. First, there is the given fact that some user groups require certain minimum amounts of data with specific higher level (or more advanced) annotations that exceed the 10 per cent norm. Second, not all types of data can be annotated with the same rate of success and/or at the same expense. Therefore, in the light of the quality standards that are to be upheld and the time and money available, certain types of data are given priority over others. The selections that were decided upon for each type of advanced annotation are displayed in Table 2.

2.4.2 Recording and collecting data; digitization

Ten million words of data amount to roughly 1,000 hours of speech. The recordings are obtained in a variety of ways. Where, as in the case of broadcast data, recordings (sometimes accompanied by rough transcriptions) can be obtained through other parties, contracts are negotiated that allow us to use the data. For components such as the direct face-to-face conversations, volunteers are recruited and asked to participate in the recording of conversations in their home environment, while a relatively small group of people is instructed to go out and record in a variety of settings (in shops, at work, in a restaurant, the theatre, etc.). For yet other components, such as the lectures, research assistants working for the project contact the schools (or institutions, or whatever), ask their permission and make the necessary arrangements for them to come and do the recording on site. On occasion there are collaborative actions where the Spoken Dutch Corpus project obtains data through other projects, as in the case of the private interviews that have been recorded within the project *The pronunciation of Standard Dutch. Varieties and variants in Flanders and the Netherlands* (Van de Velde *et al*., 1998).

All recordings are digitized. All non-telephone recordings have a sampling frequency of 16 kHz and a 16-bit resolution, while telephone recordings have a sampling frequency of 8 kHz and an 8-bit resolution. As the data are stored, no compression is applied. Information about the recording conditions, the equipment that was used, etc. is recorded as part of the meta-data.

Every speaker in the corpus is assigned a unique identification code. Information about the speakers is made available as part of the meta-data in such a fashion that it does not in any way endanger the speakers' anonymity.[4] Thus we avoid descriptions such as the following since these would make it possible to identify the speaker without much effort: a 56-year-old ex-college professor from Eindhoven who was born in Rosmalen, attended high-school in 's Hertogenbosch and graduated from Delft University, who is currently a member of the senate. Instead we classify speakers according to their age class, socio-economic class, etc.[5] Such classifications are also useful for research purposes, more specifically where research focuses on groups of speakers rather than on individuals.

[4] Of course, in the case of publicly well-known figures it is virtually impossible to keep their identity from being revealed.

[5] For example, three age classes are distinguished: young, i.e. 18-24 years of age, middle, i.e. 25-55 years of age and old, i.e. over 55 years of age. A further subclassification of the middle class distinguishes between people between 25 to 34 years of age, 35 to 44, and 45 to 55 years of age.

Tabel 2. Selections for which more advanced annotations are envisaged

Component:	total number of words in the corpus	amount of data and types of annotation (in number of words)		
		phon.transcr. + allignment	syntactic annotation	prosodic annotation
1. conversations (face-to-face)	3,000,000	150,000	550,000	100,000
2. interviews	460,000	50,000	50,000	20,000
3. telephone conversations	3,000,000	300,000	100,000	50,000
4. business transactions	175,000	15,000	15,000	10,000
5. interviews and discussions	750,000	75,000	75,000	10,000
6. discussions, debates, meetings	375,000	35,000	35,000	10,000
7. lectures	350,000	35,000	35,000	0
8. descriptions of pictures	40,000	5,000	5,000	0
9. spontaneous commentary	250,000	27,500	27,500	10,000
10. newsreports, current affairs programmes	250,000	25,000	25,000	10,000
11. nieuwsbulletins	250,000	27,500	27,500	10,000
12. commentary	200,000	25,000	25,000	10,000
13. lectures, speeches	275,000	30,000	30,000	10,000
14. read aloud text	625,000 *(+ 375,000)*	200,000	0	0
Total	**10,000,000** ***(10,375,000)***	**1,000,000**	**1,000,000**	**250,000**

Since each speaker is assigned a unique identification code, it is possible – in so far as multiple recordings involving the same speaker are available – to compare the speech of the same speaker in different recordings. Thus in one recording the speaker may be the one speaker in a monologue type of prepared speech, while in another he or she is one of the interlocutors in a highly interactive spontaneous conversation.

2.4.3 Orthographic transcription

Of all recordings a verbatim transcript is made. To a large extent the transcripts conform to the standard spelling conventions. A protocol has been developed which describes what to transcribe and how to deal with new words, dialect, mispronunciations, and so on.[6]

The procedure that is followed in order to arrive at an orthographic transcript depends on the type of data and also on whether already some (kind of) transcript is available. In the latter case it is usually worthwhile to use the available transcript and adapt it to meet the project's standards. Of course when no transcript is available or when the transcript is of very poor quality, a transcript is made strictly on the basis of the auditory signal. It is estimated that making a verbatim transcript of one hour of recorded speech requires between 8 and 38 hours: 8 hours for read aloud text where an initial transcript of reasonable quality is available and can be used to base the definitive transcript on; 38 hours for spontaneous conversations with no transcript to start from. Apart from the availability of an initial transcript, transcription experiments have demonstrated that also the number of speakers and the amount of interaction constitute major factors when it comes to the time needed to arrive at a transcript. Monologues generally are much easier to transcribe than dialogues or even multilogues, while highly interactive types of text are much more difficult to transcribe than texts with little or no interaction. The difficulty not only lies in the fact that the speech of a speaker is interrupted by that of another, the identification of the speakers (especially when more than two speakers are involved) appears in many cases problematic.

To facilitate the transcription process, use is made of the interactive signal processing tool PRAAT.[7]

[6] See also Goedertier *et al*. (2000). At present, the protocol (Goedertier and Goddijn, 2000) is in Dutch. An English motivation will be available shortly.

[7] For more information on PRAAT see http://www.fon.hum.uva.nl/praat/.

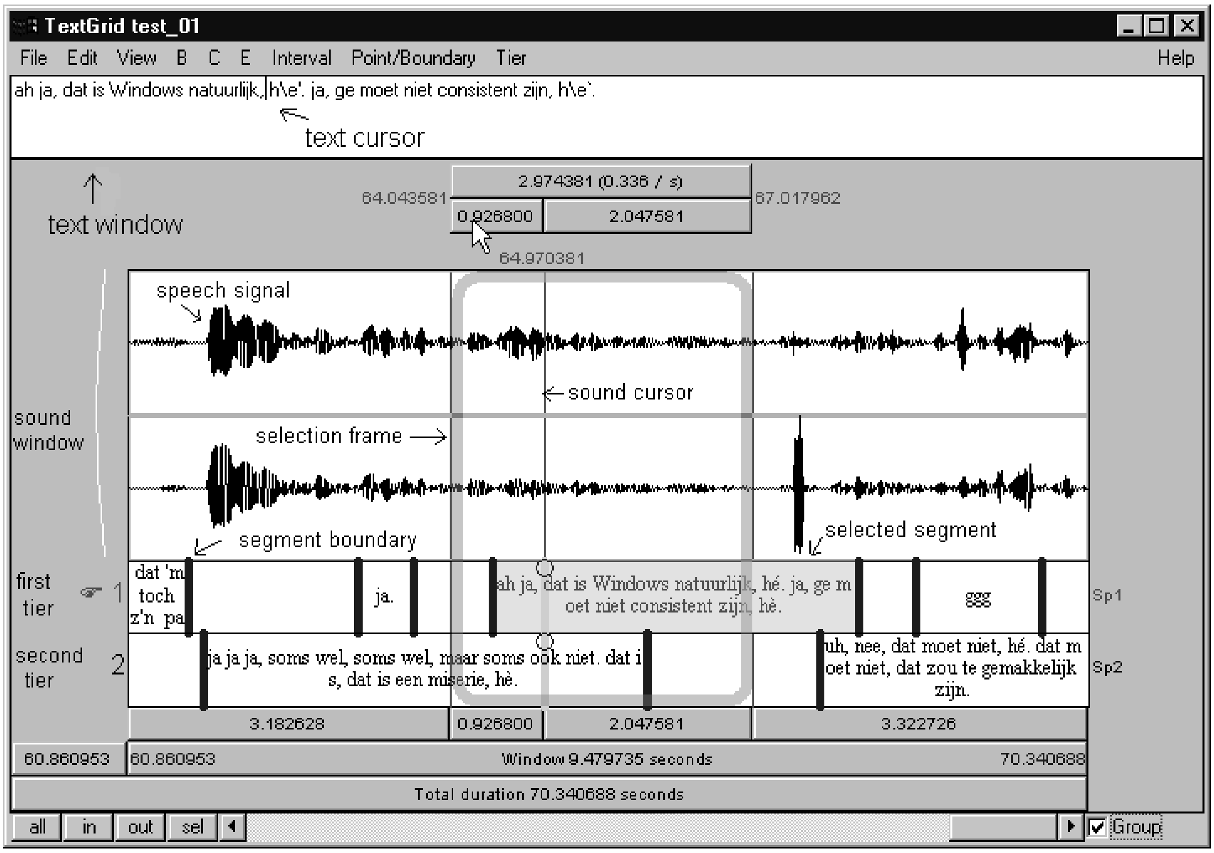

Figure 1. Screendump of the PRAAT software

In PRAAT it is possible to listen to and visualize the speech signal and at the same time create and view an orthographic transcription. Each speaker is assigned his or her own tier. For unknown speakers an additional tier is used. While the speech of unknown speakers is transcribed, no attempt is made to distinguish between multiple unknown speakers.

During the transcription process, transcribers segment the audio files in relatively short chunks (of approximately 2 to 3 seconds each) by inserting time markers in unfilled pauses between words. At a later stage these markers are used as anchor points for the automatic alignment of the transcript and the speech file.

2.4.4 Lemmatization and part-of-speech (POS) tagging

After an evaluation of taggers and tagsets available for Dutch, it was decided to define a tagset for Dutch that would conform to the EAGLES guidelines and would be compatible with the authoritative Dutch reference grammar, viz. the ANS (Haeseryn *et al.*, 1997).[8] The tagset distinguishes ten major word classes, while with each of these word classes additional morpho-syntactic features are recorded.[9] In all, the tagset consists of some 300 tags. For the tagging process a tagger has been developed which assigns the one most likely tag for a word in a given context. All output is manually checked and – where necessary – corrected. It is estimated that on average this takes about 10 hours for one hour of speech (approx. 10,000 words).

Apart from the POS tag, for each word also the associated lemma is given. In the first phase a lemmatizer is used to automatically associate with each token the appropriate lemma. The result is manually checked and corrected. At this stage the

[8] EAGLES stands for Expert Advisory Group for Language Engineering Standards. See also http://www.ilc.pi.cnr.it/EAGLES96/home.html.
[9] For a more detailed description see Van Eynde *et al*. (2000) and also Van Eynde *et al*. (2000).

constituent parts of split verbs (e.g. leidde … af), prepositions (e.g. van … uit) and such like items are lemmatized as if they occurred independently. At a later stage, a more advanced lemmatization is undertaken in which the constituent parts are considered jointly and a lemma is associated with the combination as a whole.

2.4.5 Phonetic transcription

For the broad phonetic transcription of the data, use is made of the SAMPA set.[10] In order to speed up the transcription process and also to maximize consistency, transcribers are to be provided with an automatically generated transcript which they are asked to verify and/or correct.[11] Before the exact procedure is decided upon, however, in a number of experiments it is attempted to establish whether phenomena such as cross-word assimilation should already be incorporated in the transcript that is presented to the transcribers, or whether these are best left out. It is estimated that it requires about 38 hours to yield a verified broad phonetic transcript for one hour of speech.

The part of the corpus for which a verified broad phonetic transcript is available (one million words) will be aligned automatically with the speech signal and checked manually on the word level.

2.4.6 Syntactic annotation

An annotation scheme for the syntactic annotation of one million words is being developed.[12] The scheme should cater for the idiosyncracies of spoken language data, including hesitations and false starts (*cf.* example [1]), extensions of the clause (as in [2] and [3]) and asyndetic constructions such as exemplified in [4].

[1] als je tenminste nog uh als je uh in je bed ligt
[if you at least still ehm if you ehm in your bed lie]
that is, if you're still ehm if you ehm are in bed

[2] dat verbaast me, dat je dat nog weet
[that surprises me, that you that still know]
I'm surprised that you should remember that

[3] dan heb ik zoiets van: laat maar, weet je
[then have I something of: leave it, know you]
then I have this feeling of: ah well never mind, you know

[4] (welke kranten lees jij?) bij de lunch, de Volkskrant; 's avonds, de NRC
[(which newspapers read you?) at the lunch, De Volkskrant, at night, the NRC]
(which newspapers do you read?) over lunch, De Volkskrant, at night, the NRC

The syntactic analyses will contain functional information in the form of dependency labels as well as category information (provided in the form of node labels). Syntactic annotation will be carried out semi-automatically, using the ANNOTATE software.[13]

2.4.7 Prosodic annotation

It is envisaged that 250,000 words will receive a prosodic annotation. What form exactly this annotation will take is as yet unclear. A committee of experts has been formed who are expected to write a proposal which pairs a useful interpretation of this task with what is feasible in the light of the available budget. It is intended that the annotation will encompass in any case the identification of the most important phrase boundaries as well as the one or two most important words (sentence accents) of each phrase.

2.4.8 Exploitation software

In the course of the project, software will be developed that will enable users to access the data efficiently and with relative ease. The software should be able to deal with sound files as well as various other types of data files. Basic functionality includes efficient storage, search and retrieval of data as well as an appropriate representation for each type of annotation. The generation of frequency counts and concordances are built-in standard procedures.

2.5 Dissemination of the results

During the project prospective users are kept informed about its progress by means of a newsletter and a website.[14] Intermediate results of

[10] SAMPA is an ASCII encoding system for various languages, including Dutch, based on the International Phonetic Alphabet (IPA). Zie ook http://coral.lili.uni-bielefeld.de/Documents/sampa.html.
[11] See also Hoste *et al.* (2000).
[12] Moortgat and Schuurman (in preparation).

[13] More information on ANNOTATE can be found at http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html.
[14] The website of the Spoken Dutch Corpus is situated at http://lands.let.kun.nl/cgn/.

Table 3. Data available in the first release

Component:	**Orthographically transcribed** (number of words)		**Lemmatized and tagged for POS information** (no. of words)	
	Flemish data	**Dutch data**	**Flemish data**	**Dutch data**
1. conversations (face-to-face)	--	1,500	--	--
2. interviews	44,787	162,167	--	18,714
3. telephone conversations	--	--	--	--
4. business transactions	0	--	0	--
5. interviews and discussions	8,359	7,127	--	--
6. discussions, debates, meetings	24,904	217,376	--	2,800
7. lectures	--	--	--	--
8. descriptions of pictures	--	0	--	0
9. spontaneous commentary	--	4,250	--	3,970
10. newsreps., curr. aff. programs	2,068	--	--	--
11. nieuwsbulletins	4,701	1,932	--	1,485
12. commentary	1,986	5,263	--	4,331
13. lectures, speeches	27,475	34,017	--	--
14. read aloud text	76,376	--	57,354	--
Total	**190,656**	**423,632**	**57,354**	**31,300**

the project are made available at regular (roughly) six-months intervals. The first release of the first part of the corpus was on March 1st, 2000. The date for the second release is set for September 1st, 2000. On a regular basis workshops and seminars are organized at which progress reports are presented and results discussed and evaluated. Upon completion of the project, the corpus including the recordings will probably be distributed on CD-ROM through ELRA.

3 Data available in the first release

In the first release of the first part of the corpus a total of some 615,000 words are available. Table 3 summarizes the data. For all data, sound files are available as well as an orthographic transcript. Part of the data have been lemmatized and tagged with part-of-speech information. Pending a definitive decision on the extent and nature of the meta-data, the information included in this release has been restricted to a bare minimum and must be considered provisional. More information will be made available in future releases. The meta-data that are included in this release are of two kinds: they give information about the text sample or they provide information about the speaker(s). Each text sample is classified in terms of one of the 14 components distinguished in the design of the corpus. Further information concerns the length of the sample, the number of words in the orthographic transcript, and the number of speakers. Speaker information includes the speaker's sex, age class, geographic region, and level of education.

Various audio players can be used to listen to the recordings, while the orthographic transcripts can be viewed in any editor. The use of PRAAT, however, is recommended since it allows you to play the recordings and view the orthographic transcripts at the same time. The lemmatized and tagged data are available in a tab-delimited file in plain ASCII format and can be viewed in any editor. As an example, an excerpt of one of the text samples has been included here as Figure 2. The first column gives the tokens in the input, the second column contains the contextually appropriate POS tag for each item, and the third column lists the associated lemmas. Each input string is introduced by a marker of the form <au s=Nnnnnn> which indicates the beginning of an annotation unit and identifies the speaker (s) by means of the speaker identification code. The definitive format of the annotation files has not yet been decided upon but will probably be XML-conformant.

For the first release also a number of frequency lists have been compiled. Apart from the straightforward overall word frequency counts (available as alphabetical list and as rank order list), a word frequency list has been included in which the different components of the corpus are distinguished. Other types of frequency list that have been included here are a lemma list and a list of tags. Figure 3 gives an excerpt from the word frequency rank order list, listing the top-50 types encountered in the data. In Figure 4, an excerpt from the lemma frequency list is given. For each lemma it is listed which parts of speech occurred as well the corresponding word forms.

4. Conclusion

With the compilation of the Spoken Dutch Corpus we find ourselves in a position where we have a resource for Dutch that holds up to international standards. Both with respect to its design and with respect to the nature of the transcriptions and annotations the corpus conforms to international standards, guidelines and best practice. Upon completion the corpus will be comparable in size to, for example, the spoken component of the British National Corpus (Aston and Burnard, 1998; Burnard ed., 1995). The sophistication of the different transcriptions and annotations and the availability of the sound files place the corpus in the leading ranks of state-of-the-art corpus development. Thus the expectation that the Spoken Dutch Corpus will prove a valuable asset for language and speech technologists as well as linguists from various backgrounds is well-justified.

If you are interested in the results of the Spoken Dutch Corpus Project, or would like to receive the *Corpus Gesproken Nederlands Nieuwsbrief*, please contact the Spoken Dutch Corpus secretariat at the following address:

Bureau Corpus Gesproken Nederlands
NWO, Geesteswetenschappen
Ms. A. Dijkstra
P.O. Box 93120
2509 AC The Hague
The Netherlands
Email: dijkstra@nwo.nl

5 References

Aston, G. and L. Burnard. 1998. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Bouma, G. and I. Schuurman. 1998. *De positie van het Nederlands in Taal- en Spraaktechnologie*. Rapport in opdracht van de Nederlandse Taalunie.

Burnard, L. Ed. 1995. *Users Reference Guide for the British National Corpus*. Oxford: Oxford University Press.

Donaldson, B.C. 1983. *Dutch. A Linguistic History of Holland and Belgium*. Leiden: Martinus Nijhoff.

Goedertier, W. and S. Goddijn. 2000. *Protocol voor Orthografische Transcriptie*. CGN Internal publication. Available on http://lands.let.kun.nl/cgn/.

Goedertier, W., S. Goddijn and J.P. Martens. 2000. Orthographic Transcription of the Spoken Dutch Corpus. In M. Gravrilidou, G. Carayannis, S. Markantonatou, S. Piperiolis, G. Stainhaouer. Eds. *LREC 2000 Proceedings. Athems, Greece. 31 May-2 June 2000*. Vol. 2: 887-894.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij, and M.C. van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Groningen: Martinus Nijhoff.

Hoste, V., S. Gillis and W. Daelemans. 2000. A Machine Learning Approach to Phonemic Corpus Annotation. This volume.

Moortgat, M. and I. Schuurman (in preparation). *Syntactische annotatie.*

Oostdijk, N. 2000. Meta-Data in the Spoken Dutch Corpus. In *LREC-2000 Workshop Proceedings of the Workshop on Meta-Descriptions and Annotation Schemes for Multi-Modal/Multi-Media Language Resources*. 29-30 May 2000. 21-25. Athens, Greece.

Van de Velde, H., G. De Schutter, R. van Hout, P. Adank, W. Huinck and L. Op 't Eynde. 1998. *The Pronunciation of Standard Dutch in Flanders and the Netherlands.*

Van Eynde, F. 2000. *Part-of-Speech Tagging and Lemmatisering*. CGN Internal publication. Available on http://lands.let.kun.nl/cgn/.

Van Eynde, F., J. Zavrel and W. Daelemans. 2000. Bootstrapping Morphosyntactic Annotation for the Corpus of Spoken Dutch. This volume.

<au s=N00023>		
gadverdakkie*n	TSW()	gadverdakkie
't	VNW(pers,pron,stan,red,3,ev,onz)	het
is	WW(pv,tgw,ev)	zijn
een	LID(onbep,stan,agr)	een
beetje	N(soort,ev,dim,onz,stan)	beetje
grijs	ADJ(vrij,basis,zonder)	grijs
't	VNW(pers,pron,stan,red,3,ev,onz)	het
regent	WW(pv,tgw,met-t)	regenen
't	VNW(pers,pron,stan,red,3,ev,onz)	het
is	WW(pv,tgw,ev)	zijn
nat	ADJ(vrij,basis,zonder)	nat
.	LET()	.
<au s=N00023>		
de	LID(bep,stan,rest)	de
mensen	N(soort,mv,basis)	mens
die	VNW(betr,pron,stan,vol,persoon,getal)	die
de	LID(bep,stan,rest)	de
hond	N(soort,ev,basis,zijd,stan)	hond
hebben	WW(pv,tgw,mv)	hebben
moeten	WW(inf,vrij,zonder)	moeten
uitlaten	WW(inf,vrij,zonder)	uitlaten
die	VNW(aanw,pron,stan,vol,3,getal)	die
hebben	WW(pv,tgw,mv)	hebben
d'r	VNW(aanw,adv-pron,obl,red,3o,getal)	daar
alles	VNW(onbep,pron,stan,vol,3o,ev)	alles
al	BW()	al
van	VZ(fin)	van
mee	VZ(fin)	mee
kunnen	WW(inf,vrij,zonder)	kunnen
maken	WW(inf,vrij,zonder)	maken
.	LET()	.
<au s=N00023>		
het	VNW(pers,pron,stan,red,3,ev,onz)	het
regent	WW(pv,tgw,met-t)	regenen
nog	BW()	nog
behoorlijk	ADJ(vrij,basis,zonder)	behoorlijk
ook	BW()	ook
nog	BW()	nog
.	LET()	.

Figure 2. Excerpt from text sample fn000001 with POS tags and lemmas

24509	de	9064	die	4393	met	3241	we	2271	bij
18620	dat	8543	ja	4376	voor	3227	of	2238	uhm
18025	uh	6820	niet	4290	zijn	3219	daar	2145	was
15816	en	6501	ook	4003	wat	3106	er	2080	heel
13238	een	6219	dan	3895	dus	2859	nog	2073	naar
12902	ik	6129	maar	3733	als	2635	u	1983	nou
12451	van	5755	je	3617	wel	2549	zo	1848	nu
11759	het	5712	op	3543	om	2415	over	1838	moet
10081	in	5433	te	3539	ze	2409	hebben	1790	heeft
9637	is	5316	't	3452	aan	2352	heb	1772	toch

Figure 3. Excerpt from the word frequency rank order list (top 50 types)

8	rij		
	4 (0.500000)	N(soort,ev,basis,zijd,stan)	rij
	1 (0.125000)	N(soort,ev,dim,onz,stan)	rijtje
	3 (0.375000)	N(soort,mv,basis)	rijen
1	rijbewijs		
	1 (1.000000)	N(soort,ev,basis,onz,stan)	rijbewijs
19	rijden		
	5 (0.263158)	WW(inf,vrij,zonder)	rijden
	1 (0.052632)	WW(od,prenom,zonder)	rijdend
	1 (0.052632)	WW(pv,imp,ev)	rij
	1 (0.052632)	WW(pv,tgw,met-t)	rijdt
	2 (0.105263)	WW(pv,tgw,mv)	rijden
	3 (0.157895)	WW(pv,verl,ev)	reed
	4 (0.210526)	WW(pv,verl,mv)	reden
	2 (0.105263)	WW(vd,vrij,zonder)	gereden
1	rijdier		
	1 (1.000000)	N(soort,ev,basis,onz,stan)	rijdier

Figure 4. Excerpt from the lemma frequency list