The Spoken Dutch Corpus. Overview and first Evaluation

Nelleke Oostdijk
Dept. of Language and Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
n.oostdijk@let.kun.nl

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 overall description of the project, its aims, structure and organization. It then goes on to discuss the considerations – both methodological and practical – that have played a role in the design of the corpus as well as in its compilation and annotation. The paper concludes with an account of the data that are available in the first release of the first part of the corpus that came out on 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.

The present paper aims to introduce the Spoken Dutch Corpus Project to researchers in European industry and acedemia. It also solicits comments, criticisms, and suggestions on various aspects of the work that is being done, from corpus design and compilation to corpus annotation, but also with regard to the dissemination and evaluation of the results. The paper is structured as follows: In Section 2, I describe the project in more detail. In Section 3 the design of the corpus is discussed, while Sections 4 and 5 describe various aspects of the compilation and annotation of the corpus. In Section 6 an account is given of 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 with a brief evaluation of our experiences in the project so far.

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[2] 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, 1998a,b). 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 everyday 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

[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.

[2] Examples are (the spoken part of) the British National Corpus (BNC; Burnard, 1995; Aston and Burnard, 1998), the Cambridge version of the Wall Street Journal text corpus (Fransen *et al.*, 1994), and the Switchboard Corpus (http:/www.cis.upenn.edu/~ldc/readme/switchbrd.readme.html).

governments, the Dutch Language Union[3], Dutch and Flemish research foundations and one of the Dutch national research schools (LOT). 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 directed by a project manager. The project managers 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.

A user group has been set up whose principal role is to monitor and critically assess the design and implementation of procedures and protocols and to evaluate (intermediate) results.

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. While it may seem that the start-up phase of the project has been rather lengthy, it should be pointed out that for such a project – aimed at the compilation of a corpus of a lesser documented and researched language such as Dutch – a great deal of time must necessarily be spent on these preparatory activities.

Over the remaining four years the corpus will be compiled, transcribed and annotated incrementally in eight six-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. 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.[4] Intermediate results of the project are made available at regular (roughly) six-month 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 are discussed and evaluated. Upon completion of the project, the corpus including the recordings will probably be distributed through ELRA.

3. 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. In this respect, the Spoken Dutch Corpus is unique. Unlike other corpora, the Spoken Dutch Corpus is not being compiled for a specific purpose or in the interest of a (single) well-defined user group. 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[5], 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).

[3] The Dutch Language Union is an intergovernmental organization, based on the 1980 Dutch Language Union Treaty between the Netherlands and Belgium concerning their language policy. In the case of the Spoken Dutch Corpus it is the Dutch Language Union which holds all rights.

[4] http://lands.let.kun.nl/cgn/

[5] This constitutes a major difference between the Spoken Dutch Corpus and for example the British National Corpus: for the latter the recordings are not publicly available.

<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>

Table 1: Overall design of the corpus

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. The overall design of the corpus is given in Table 1.

In all, 14 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 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.

Component:	total number of words in the corpus	type of annotation and amount of data (no. 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	10,000	10,000	0
9. spontaneous commentary	250,000	25,000	25,000	10,000
10. newsreports, current affairs progr.	250,000	25,000	25,000	10,000
11. nieuwsbulletins	250,000	25,000	25,000	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	1,000,000	200,000	0	0
Total	**10,375,000**	**1,000,000**	**1,000,000**	**250,000**

Table 2. Selections for which more advanced annotations are envisaged

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.

4. Corpus compilation

4.1. 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 transcripts) 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, etc.). For yet other components, such as the lectures, research assistants working for the project contact the schools (or institutions, or such like), 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.

4.2. Speaker-related meta-data

All speakers in the corpus are 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.[6] Thus we avoid descriptions that would make it possible to identify the speaker without much effort. Instead we classify speakers according to their age class, socio-economic class, etc. Such classifications are also useful for research purposes, more specifically where research focuses on groups of speakers rather than on individuals. The number of classes distinguished is generally small (between 2 and 5). Where considered useful, subclasses are introduced. 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.

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 occur in a prepared monologue, while in another he or she is one of the interlocutors in a highly interactive spontaneous conversation.

5. Corpus annotation

5.1. Orthographic transcription

Of all recordings a verbatim transcript is made. Following the recommendations made in den Os (1998: 170f), the transcripts to a large extent conform to the standard spelling conventions. A protocol has been developed which describes what to transcribe and how to

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

deal with new words, dialect, mispronunciations, and so on.[7]

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 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.[8] In PRAAT it is possible to listen to and visualize the speech signal and at the same time create and view an orthographic transcript. 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.[9]

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.

5.2. 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[10] and would be compatible with the authoritative Dutch reference grammar, viz. the ANS (Haeseryn *et al*., 1997). The tagset distinguishes ten major word classes, while with each of these word classes additional morpho-syntactic features are recorded.[11] In all, the tagset consists of some 300 tags. For the tagging process a tagger has been developed which assigns the 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 constituent parts of split verbs (e.g. *leidde* … *af*, where the verb is *afleiden*), prepositions (e.g. *van* … *uit* instead of *vanuit*) 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.

5.3. Phonetic transcription

For the broad phonetic transcription of the data, use is made of SAMPA.[12] 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. 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.

5.4. Syntactic annotation

An annotation scheme for the syntactic annotation of one million words is being developed.[13] 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
[2] dat verbaast me, dat je dat nog weet
[3] dan heb ik zoiets van: laat maar, weet je
[4] (welke kranten lees jij?) bij de lunch, de Volkskrant; 's avonds, de 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

[7] See Goedertier, W. and S. Goddijn (2000). At present, the protocol is in Dutch. An English motivation will be available shortly.
[8] For more information on PRAAT see http://www.fon.hum.uva.nl/praat/
[9] For more information, we refer to Goedertier *et al*. (2000).
[10] Cf. the *Recommendations for the Morphosyntactic Annotation of Corpora* of the Expert Advisory Group on Language Engineering Standards (EAGLES, 1996).
[11] For a more detailed description, see Van Eynde (2000) and Van Eynde *et al*. (2000).
[12] The acronym SAMPA stands for Speech Assessment Methods Phonetic Alphabet, which is a machine-readable phonetic alphabet which has been applied to a variety of languages, including Dutch. See also Gibbon *et al*. (eds.) (1998), Vol. IV Appendix B.
[13] Moortgat and Schuurman (in preparation).

labels). Syntactic annotation will be carried out semi-automatically, using the ANNOTATE software.[14]

5.5. Prosodic annotation

It is envisaged that 250,000 words will receive a prosodic annotation. At this time it is as yet unclear what form this will take. 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. The first concern of the committee is to make an inventory of users' needs and stipulate the requirements for this type of annotation. Next, available annotation schemes such as ToDI will be taken into consideration. The acronym ToDI stands for Transcription of Dutch Intonation. The scheme was developed by C. Gussenhoven, T. Rietveld, and J. Terken and resembles the ToBI (Tones and Break Indices; Silverman *et al*. 1992) scheme that was developed for American English, but has (in an adapted form) also been applied to other languages (see den Os, 1998: 162 ff).[15] Since the prosodic annotation of a substantial amount of corpus data for Dutch is a novel development, experiments are necessary to establish to what extent a given annotation scheme can successfully be applied.

6. Data available in the first release

The first release of the part of the corpus was on March 1st, 2000. In this release a total of some 615,000 words are available; approximately 425,000 words originate from The Netherlands and 190,000 from Flanders. For all data, sound files are available as well as an orthographic transcript. Part of the data (some 90,000 words altogether, i.e. 60,000 from Flanders and 30,000 from the Netherlands) 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 (cf. Table 1. 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. The definitive format of the annotation files has not yet been decided upon but will probably be SGML or XML-conformant, following the guidelines and recommendations of the Text Encoding Initiative (TEI; Sperberg-McQueen and Burnard, 1994) and the Corpus Encoding Standard (CES; Ide, 1996).

For the first release also different types of frequency list 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 list of POS tags and a lemma list. In the latter list for each lemma it is listed which parts of speech occurred as well the corresponding word forms.

The first release has been distributed on CD-ROM exclusively among partners in the Spoken Dutch Corpus project and members of the user group. The latter group plays a crucial role in the evaluation process. With a resource of this kind that is intended to serve so many and diverse needs, it is of the utmost importance to get feedback from a very early stage onward, so that procedures and protocols may be revised, adapted or refined if and where required.

In our experience so far, the collaboration in the project between the Flemish and Dutch partners has been a fruitful one, notwithstanding the many regional differences that exist. The design of the corpus is the result of extensive discussions and is expected to meet the collective needs in the way of a resource for spoken Dutch. While procedures and protocols are developed jointly, regional colouring has proven to be desirable and sometimes unavoidable. For example, in the protocol for orthographic transcription a number of phenomena are described that are characteric of either northern Dutch or southern Dutch. Moreover, practical circumstances sometimes lead us to use somewhat different procedures, although never to the extent that the results are in any way incompatible. Now that the first release is available, Dutch and Flemish data can be compared and evaluated, and the results can be used to the benefit of future releases.

7. Conclusion

Despite the fact that the project is somewhat behind schedule[16], I think it is fair to say that the Spoken Dutch Corpus project is well under way, especially now that main procedures and protocols (such as for orthographic transcription and POS tagging) have been established. On evaluation, looking at the development of the project up to now, we find that reaching a consensus over how an ambitious and complex project plan like this one is to be implemented is a very time-consuming process. At the same time, though, we are confident that the time and effort spent in the initial phase of the project will prove to be well-invested.

8. Obtaining further information

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:

---

[14] More information on ANNOTATE can be found at http: www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

[15] For more information on ToDI, see http://lands.let.kun.nl/todi.

[16] The original planning was to have a release of 1,250,000 words every six months (years 2 through 5).

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

9. Acknowledgements

This publication was supported by the Netherlands Organization for Scientific Research (NWO) under grant number 014-17-510. Thanks are due to Lou Boves and Louis Pols for their suggestions and comments on an earlier version of this paper.

10. 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, 1998a. Intergovernmental Language Policy for Dutch and the Language and Speech Technology Infrastructure. In Rubio, A., N. Gallardo, R. Castro and A. Tejada, Eds., *Proceedings First International Conference on Language Resources & Evaluation*, Granada, Spain 28-30 May 1998. 509-513.

Bouma, G. and I. Schuurman, 1998b. *De Positie van het Nederlands in Taal- en Spraaktechnologie.* Report for the Dutch Language Union.

Burnard, L. 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.

EAGLES, 1996. *Expert Advisory Group on Language Engineering Standards. Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG-TCWG-MAC/R. Version March 1996.

Fransen, J., D. Pye, T. Robinson, P. Woodland and S. Young, 1994. *WSJCAM0 Corpus and Recording Description.* http://morph.ldc/upenn/edu/catalog/docs/wsjcam0/wsjcam0.ps

Gibbon, D., R. Moore, and R. Winski, Eds., 1998. *Handbook of Standards and Resources for Spoken Language Systems. Vol. IV. Spoken Language Reference Materials.* Berlin, New York: Mouton de Gruyter.

Goedertier, W. and S. Goddijn, 2000. *Protocol voor Orthografische Transcriptie.* CGN Internal publication.

Goedertier, W., S. Goddijn, and J. Martens, 2000. Orthographic Transcription of the Spoken Dutch Corpus. *Proc. LREC2000*.

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

Ide, N., 1996. *Corpus Encoding Standard*. http://www.cs.vassar.edu/CES/

Moortgat, M. and I. Schuurman, in preparation. *Syntactische Annotatie.*

Os, E. den, 1998. SL Corpus Representation. In D. Gibbon, R. Moore and R. Winski (eds.), *Handbook of Standards and Resources for Spoken Language Systems.* Vol. IV *Spoken Language System and Corpus Design*. 146-174. Berlin, New York: Mouton de Gruyter.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg, 1992. ToBI: A Standard for Labeling English Prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing, ICSLP*, 12-16 October 1992 Banff, Canada. Vol. 2: 867-870.

Sperberg-McQueen, C.M. and L. Burnard, Eds., 1994, *Guidelines for Electronic Text Encoding and Interchange.* ACH-ACL-ALLC.

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 en Lemmatizering*. CGN Internal publication.

Van Eynde, F., J. Zavrel and W. Daelemans, 2000. Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus. *Proc. LREC2000*.