The Spoken Dutch Corpus Project

Nelleke Oostdijk, University of Nijmegen

Introduction

The Spoken Dutch Corpus project is a five-year project that started in June 1998. It is funded jointly by the Flemish and Dutch governments and Science Foundations with a budget of some 4.6 MEuro. The project aims at the compilation and annotation of a 10-million-word corpus of contemporary standard Dutch as spoken in the Netherlands and Flanders. 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 a broad phonetic transcription and a syntactic 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.

Background, 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-based 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 everyday actual language use are indispensable for developing Dutch language courses and course materials.

Project organization

The Spoken Dutch Corpus project is directed by a board whose members include representatives of the two governments, the Dutch Language Union[2], 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 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.

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 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 eight six-months 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.

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). 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 particularly interesting speaker characteristics such as gender, age, geographical region, and socio-economic class are used as sampling criteria, otherwise they are merely recorded as part of the meta-data. The overall design of the corpus is given in Table 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 colleges, ask 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. Information about the recording conditions, the equipment that was used, etc. is recorded as part of the meta-data.

Orthographic transcription

For all recordings a verbatim transcript will be available.[3] To facilitate the transcription process, use is made of the interactive signal processing tool PRAAT.[4] During the transcription process, transcribers segment the audio files in relatively short chunks 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.

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). The tagset distinguishes ten major word classes, while with each of these word classes additional morpho-syntactic features are recorded. 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. Apart from the POS tag, for each word also the associated lemma is given.

Phonetic transcription

For the broad phonetic transcription of the data, use is made of the SAMPA set. In order to speed up the transcription process and also to maximize consistency, transcribers are provided with an automatically generated transcript which they are asked to verify and/or correct.

Syntactic annotation

An annotation scheme is currently being developed. The syntactic analyses will contain information about the constituent structure of major linguistic units as well as the syntactic relations that hold between constituents. Syntactic annotation will be carried out semi-automatically, using the ANNOTATE software. [5]

Dissemination of the results

During the project, prospective users are kept informed about its progress by means of a newsletter and a website.[6] Intermediate results of the project will be 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 discussed and evaluated. Upon completion of the project, the corpus including the recordings will be distributed on CD-ROM through ELRA.

For more information, 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

Notes

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

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

3. The orthographic transcription conforms to a large extent to 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. See Goedertier and Goddijn (2000). At present, the protocol is in Dutch. An English motivation will be available shortly.
4. For more information on PRAAT see http://www.fon.hum.uva.nl/praat/
5. More information on ANNOTATE can be found at http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html
6. http://lands.let.kun.nl/cgn/

References

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

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. See also http://lands.let.kun.nl/cgn/release1/info/paginas/gg_2000.htm

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

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'.

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>discussions, 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>