Spontaneous Speech in the Spoken Dutch Corpus

Lou Boves, Nelleke Oostdijk
Dept. of Language and Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
{l.boves | 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 corpus of 1,000 hours 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. Although the corpus will contain a fair amount of read speech (mainly to train initial acoustic models for speech recognizers), the lion's share of the data will consist of spontaneous speech, ranging from lectures to unobtrusively recorded conversations. The corpus is unique in that all speech recordings will be made available together with several levels of high quality annotations, from verbatim orthographic transcriptions to syntactic analyses and prosodic labeling.

1. Introduction

In June 1998 the Spoken Dutch Corpus (in Dutch: *Corpus Gesproken Nederlands*, or CGN) project was started, a five-year project aimed at the compilation and annotation of a corpus of 1,000 hours of speech originating from adult speakers of standard Dutch 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 orthographically transcribed, lemmatized and annotated with part-of-speech information. For a selection of approx. 100 hours (estimated to comprise about one million words), more detailed annotations are envisaged, including a manually verified broad phonetic transcription and syntactic annotation. About 25 hours (250,000 words) will receive a prosodic annotation. The CGN is reminiscent of the British National Corpus [3], but contrary to the BNC the CGN will make all audio files available along with the annotation data. In that respect the CGN is also similar to the Spontaneous Speech Corpus of Japanese [2], but the sampling criteria used in the latter corpus were different. To facilitate access to the recordings, the transcriptions will be enriched with pointers into the speech files.

The present paper presents an account of our experiences with collecting and annotating spontaneous speech within the CGN project.

2. The Spoken Dutch Corpus Project

2.1. Background and motivation

As one of the smaller languages in Europe, Dutch is under threat of gradually disappearing as a language for business communication. The fact that to date few relevant language resources are available is an impediment for the advancement of Dutch language and speech technology. The CGN project will 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. 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 outline and timetable

The first year of the project has mainly 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. In the subsequent three years the main focus has been on recording, transcribing and annotating the data. Now that the project has entered its fifth and final year, it is evident that it will be difficult to reach the target of 1,000 hours of recorded speech with all the annotations that were originally envisaged. We will, however, come very close.

In the course of the project, also software has been developed that enables users to access the data efficiently and with relative ease. The software deals 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.

2.3. Dissemination of the results

During the project, prospective users are kept informed about its progress by means of a newsletter and a website.[1] Intermediate results of the project are made available at regular (roughly) six-month intervals. The pre-final release of corpus was on November 1st, 2002. 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 full corpus will be distributed through ELRA.

3. Corpus design

The design of the corpus was guided by a number of considerations. First, the corpus is intended as a plausible sample of contemporary standard Dutch as spoken in the Netherlands and Flanders that should support a wide range of basic and applied research interests. Second, the corpus should constitute a resource for Dutch that holds up to international standards. Moreover, because of the time, financial and legal constraints under which the pro-

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

ject must operate, but also for practical reasons, it is impossible to include all possible types of speech and compromises are inevitable. For example, the fact that the corpus will be distributed including the audio files requires that the consent of all speakers is obtained as well of any other parties that have any rights to the recorded material. This, obviously, is not always feasible.

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). In all, 14 components are distinguished, specified 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 sampling criteria; otherwise they are merely recorded as part of the meta-data.[2] Considerations that have played a role in determining the present sizes of the components include 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 judgments 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 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 minimum amount of data; this is especially true for components that are used for the development of technological applications.

3.1. The ‘core corpus’

Once the overall design of the corpus had been established, it remained to be decided which parts of the corpus should receive more advanced annotations. 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, two powerful arguments were raised against this procedure. First, 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 accuracy 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 also detailed in [1].

4. Corpus compilation

4.1. Recording and collecting data; digitization

At the start of the project we estimated that 1,000 hours of speech amounts to roughly ten million words. In the course of the project it appeared, however, that on average the rate of speech is much higher in spontaneous conversations than in the more formal types of speech on which the initial estimate was based.

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 face-to-face conversations, volunteers were recruited who recorded conversations in their home environment, while a smaller group of people was 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 contact schools, ask permission and make the necessary arrangements to do the recording on site. On occasion there are collaborative actions where the CGN project obtains data through other projects.

All non-telephone recordings have a sampling frequency of 16 kHz and a 16-bit linear resolution, while telephone recordings have a sampling frequency of 8 kHz in the 8-bit A-law format. 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. We classify speakers according to their gender, age group, geographic region, 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. 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.

5. Corpus annotation

5.1. Orthographic transcription

Of all recordings a verbatim transcript is made. Following the recommendations made in [2,3], the transcripts conform to the standard spelling conventions as much as possible. A protocol has been developed which describes what to transcribe and how to deal with new words, dialect, mispronunciations, and so on.

[2] The overall design of the corpus is given in [1].

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 that transcript and adapt it to meet the project's standards. When no transcript is available or when the transcript is of very poor quality, a transcript is made just on the basis of the auditory signal.

Practice has shown that it takes between 8 and 38 hours to produce a verbatim transcript of one hour of recorded speech: 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. 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 a separate 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.

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 [4] and would be compatible with the authoritative Dutch reference grammar [5]. The CGN tagset consists of 316 tags. It distinguishes ten major word classes, while with each of these word classes additional morpho-syntactic features are recorded. 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. 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. <u>*leidde*</u> … <u>*af*</u>, where the verb is *afleiden*), prepositions (e.g. <u>*van*</u> … <u>*uit*</u> 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 many research aims a reliable narrow phonetic transcription of the full CGN would be a major asset. However, providing such transcriptions would require resources far beyond the budget. Moreover, there is some doubt whether a `reliable narrow phonetic transcription' can be made. Many believe that the degree of detail that one would require from a narrow phonetic transcription strongly depends on the aims and requirements of a specific research project. For example, an investigation focusing on regional differences in the degree of diphthongisation of long vowels might require another type of detail than a study into the degree of devoicing of fricatives. Thus, a coarse, yet reliable, which can be augmented later on by adding the details that are required by a specific project is to be preferred.

A combination of budgetary and scientific considerations has thus resulted in the decision to restrict the phonetic transcription to a broad phonemic level. The starting point for the transcriptions is a phonemic representation of the orthographic transcription that is generated fully automatically. The set of symbols used in the transcriptions is derived from the SAMPA set. This set does not contain diacritics, so that the transcription is truly limited to the broad phonemic level. The design of the internal data structures of the CGN are completely based on the concept of words as units delimited by blank spaces. This principle was carried over to the level of phonemic transcription. However, cross-word assimilations and degeminations abound in continuous speech. To retain the one-to-one correspondence between the orthographic words and the phonemic transcriptions, a special notation had to be developed for cross-word degemination.

Work is under way to develop automatic transcription procedures that maximize the `quality' of the automatic transcription. Automatic phonemic transcriptions will be provided for the full CGN. For approximately 100 hours of speech the automatic phonemic transcriptions will be checked and corrected by students trained for this task. The work is supervised by trained phoneticians.

The procedure for the manual verification is defined in a detailed protocol. Extensive discussions were needed to define a protocol that is at the same time sufficiently detailed as well as practical. Based on our experiences now that the protocol has been in use for more than two years we can say that it has proved to be adequate for the task at hand. Transcribers encounter few problems, and if problems do occur, supervisors find it easy to arbitrate. An evaluation of the procedure for phonetic transcription in the CGN project is given in [9].

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 verified manually on the word level.

5.4. Syntactic annotation

A scheme has been developed for the syntactic annotation of part of the corpus. The scheme caters for the idiosyncracies of spoken language data, including hesitations and false starts, extensions of the clause and asyndetic constructions.

The syntactic analyses contain functional information in the form of dependency labels as well as category information (provided in the form of node labels). Syntactic annotation is carried out semi-automatically, using the ANNOTATE software.[3]

5.5. Prosodic annotation

Potential users of the prosodic annotation expressed a preference for a perceptually based annotation that can serve as a starting point for further detailed prosodic labeling, e.g., ToBI [6]. Therefore, the annotation is limited to marking prominent syllables, locating important between-word and within-word breaks, and marking prosodically relevant lengthening of individual vowels and consonants not carrying prominence. Syllables are marked as either prominent or not (there is no distinction between different degrees of prominence), and a break can either be weak or strong. This annotation scheme constitutes a compromise between what is desirable information for a large number of users, and what can actually be provided with a sufficiently high degree of consistency at a limited cost.

The prosodic annotation is produced by students working at four different sites, under the direction of four different supervisors. Therefore, procedures had to be developed to maximize the degree of consistency between students and sites. Since prominence and break strengths are basically ordinal variables, it is important to reach a common understanding of these labels. Therefore, we developed a protocol providing examples and describing the general rules and procedures to follow during the annotation. Moreover, the examples in the protocol are supplemented with speech fragments and their prosodic annotation. These real examples are supplied in the form of a learning corpus for which the supervisors created a consensus annotation.

6. Quality control and consistency

To maintain consistency between the annotation levels and to obtain optimal quality control, we have developed a set of procedures for validation and bug-reporting. During the transcription/annotation process transcriptions/annotations of one transcriber/annotator are checked by another transcriber/annotator. Upon completion, the transcription/annotation is checked by means of a dedicated tool which checks for illegal characters or symbols or suspect sequences of characters/symbols. All words (tokens) and lemmas in the orthographic transcriptions are validated against the lexicon, as are all combinations of type-tag pairs. Quality checks are also made on the basis of the information in the frequency lists that are updated regularly. Low frequency items typically help to pinpoint potential errors, while alternative entries for one and the same item help to identify inconsistencies.

In so far as one type of annotation builds - directly or indirectly - on another type (as POS tagging on orthographic transcription, but also for part of the material syntactic annotation on POS tagging or phonetic transcription on orthographic transcription), this automatically involves a verification of the output of a previous annotation. Upon the detection of what is perceived to be an error, a bug report is filed with the team responsible for the annotation. Based on the feedback of groups working on other transcription or annotation layers, many errors and inconsistencies are detected and subsequently corrected. While some of the corrections can be done automatically, part of the work needs at least some human interaction. Thus the feedback from other transcription and annotation levels has proved to be at once a very important tool to maximize the quality of the transcriptions and annotations, and a task that has turned out to be much more time-consuming than we had anticipated.

Tools that we have found useful for quality control and consistency include a customized spelling checker that enforces the conventions adopted in the protocol for orthographic transcription, an XML parser for validating the format of the data files, a tag selection program that is used for the manual verification of the tagger output.

7. Acknowledgement

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

8. References

[1] Oostdijk, N. (2000) The spoken Dutch corpus. Overview and first evaluation. *Proc. Second International Conference on Language Resources and Evaluation*, vol. 2, p. 887-893.

[2] Maekawa, K., Koiso, H., Furui, S and Isahara, H. (2000) Spontaneous speech corpus of Japanese. *Proc. LREC2000*, Athens, Greece, vol.2, pp.947-952.

[3] Aston, G. and Burnard, L. (1998) *The BNC Handbook.* Edinburgh: Edinburgh University Press.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] Binnenpoorte, D, Goddijn, S. and Cucchiarini, C. (2003) How to improve human and machine transcriptions of speech. *Proc. Workshop Spontaneous Speech processing and Recognition*, Tokyo 2003.

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