The 10-million-words Spoken Dutch Corpus and its possible use in experimental phonetics

Louis C.W. Pols

Institute of Phonetic Sciences / ACLC, University of Amsterdam
Amsterdam, The Netherlands
(Louis.Pols@hum.uva.nl)

1. Introduction

In the Netherlands and in Flanders a 10-million-words corpus of contemporary standard Dutch is now being collected. This very exciting 5-year project started in June 1998 and will result in 7 releases of audio files plus various transcriptions and annotations, distributed for the time being on CD-ROM. The Dutch Language Union holds all rights (http://www.taalunie.org/). The speech material collected (about 1,000 hours) varies from read aloud texts to interviews and spontaneous telephone conversations. The entire corpus will be orthographically transcribed, lemmatized and annotated with part-of-speech tags. A selection of one million words will be phonetically transcribed and syntactically annotated. For this set the automatic time alignment will be manually checked at the word level. A subset of 250,000 words will receive a prosodic annotation. To facilitate the various transcription processes, the interactive signal processing tool PRAAT is being used (http://www.fon.hum.uva.nl/praat/). For more details about the project itself see http://lands.let.kun.nl/cgn/.

This large corpus of spoken Dutch addresses the needs of language and speech technologists, but also of linguists from various backgrounds. Exploitation software will be developed to enable users to access efficiently and with relative ease the sound files as well as the various other types of data files.

This corpus promises to be a gold mine for phonetics research as well. It allows users to study regional variants of standard Dutch. For the first time we will be able to study truly conversational speech, with all its hesitations, restarts, so-called ungrammaticalities and mispronunciations, turn-taking behavior, etc. We will get a better idea about vowel and consonant reduction, phoneme and syllable deletion, within- and between-word assimilation, co-articulation, realization of prominence and phrasing, speaking rate variation and rhythmic effects, effects of frequency-of-occurrence, and so on.

2. Corpus design

The “Corpus Gesproken Nederlands” (Spoken Dutch Corpus) will result in a large corpus of contemporary standard Dutch as it is spoken nowadays in the Netherlands and in the Flemish part of Belgium (Flanders). The goal is to collect 10-million words, which is expected to be around 1,000 hours of speech, from a variety of adult male and female speakers in a variety of conditions. These conditions are carefully selected in such a way that they appropriately cover both monologues and dialogues (or multilogues), both read and spontaneous speech, and both face-to-face conversations as well as telephone conversations. Also topic-wise there will be a great variety, from interviews, to lectures, to debates, to commentaries and news. Ten percent of the data are selected for more advanced annotations (phonetic transcription and syntactic annotation) and a quarter of that will also be prosodically annotated. Table 1 provides some information about the overall corpus design, for more details, see Oostdijk (2000). All recordings are digitized at 16 kHz with 16-bit resolution, apart from the telephone recordings that have an 8-kHz sampling frequency and 8-bit resolution. Information about the speakers and the recording conditions are collected in a meta database.

Corpus components		Total corpus	Phonetic transcr.	Syntactic annotat.	Prosodic annotat.
1	conversations (face-to-face)	3,000	150	550	100
2	interviews	460	50	50	20
3	telephone conversations	3,000	300	100	50
4	business transactions	175	15	15	10
5	interviews and discussions	750	75	75	10
6	discuss., debates, meetings	375	35	35	10
7	lectures	350	35	35	0
8	description of pictures	40	10	10	0
9	spontaneous commentary	250	25	25	10
10	news rep., current affairs progr.	250	25	25	10
11	news bulletins	250	25	25	10
12	commentary	200	25	25	10
13	lectures, speeches	275	30	30	10
14	read aloud text	1,000	200	0	0
	Total	**10,375**	**1,000**	**1,000**	**250**

Table 1: Number of words x 1000 in the various corpus components (1-7 concern dialogues and multilogues, whereas 8-14 concern monologues). Columns represent the whole corpus, as well as various selections for which more advanced annotations are envisaged.

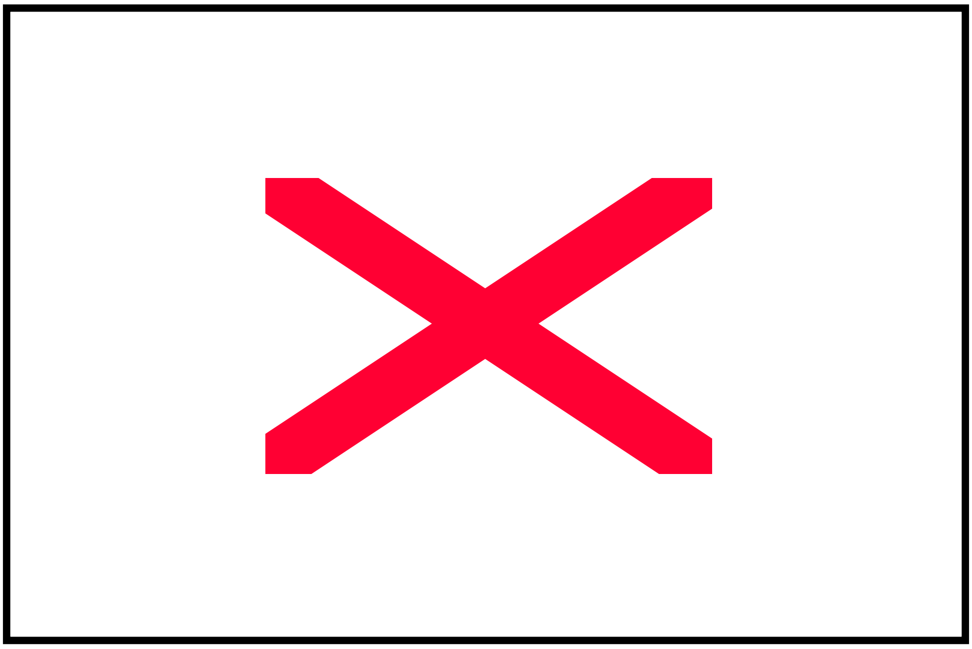

Figure 1: Screen dump of the transcription interface of the interactive speech signal processing package PRAAT. For more details, see text.

3. Orthographic transcription

The whole corpus will be orthographically transcribed by trained students (and checked in an extra transcription cycle) according to a protocol that has grown from experience and that aims at optimal consistency, accuracy and usability. Text interpretation is avoided as much as possible, grammatical errors are not corrected, truncated words are not completed, and interjectives are limited to a closed set with a fixed spelling to facilitate later consistent search. The transcripts follow the standard spelling convention as much as possible and are checked on line against the orthographic entries of a lexicon. As punctuation marks, only sentence endings are used, thus no commas, colons, semicolons, or question marks are used since they would require too much interpretation. The transcriptions are aligned to the speech signal at the level of so-called chunks of speech of up to a few seconds. For more details, see Goedertier et al. (2000).

Figure 1 is a screen dump of the interactive speech signal processing package PRAAT that is used extensively within this project, and more specifically for the orthographic transcription. For more details about PRAAT, see http://www.fon.hum.uva.nl/praat/ . This specific example shows two audio channels (stereo signal) and five tiers. It is no problem to represent simultaneous or overlapping speech. The first two tiers in this example have been selected for two specific speakers (N01011 and N01010) in this conversation, the third tier (speaker_ unknown) is meant for incidental speech that cannot be attributed to a known speaker, the fourth tier (background) can contain transcriptions of clearly audible and meaningful background noises, the final tier (comments) can contain general remarks that characterize the recording as a whole.

4. Part-of-speech tagging, lemmatization and syntactic annotation

All words in the corpus will also be automatically tagged with a parts-of-speech (POS) tag. A Dutch tag set is developed with 10 major word classes (noun, adjective, verb, pronoun, article, numeral, preposition, adverb, conjunction, and interjection), additional morpho-syntactic features per word class (such as *singular*, *dimunitive* and *neuter* for nouns) result in a total tag set of some 300 tags. A self-learning automatic tagger has been developed which assigns the most likely tag for a word in a given context. Also the associated base form (lemma) is provided, for verbs this is the infinitive, for most other forms this is the stem, i.e. a form without inflectional affixes. For more details, see Van Eynde et al. (2000). A semi-automatic annotation scheme for the syntactic annotation of 10% of the corpus is presently being developed. It will use the ineractive annotation software developed within the NEGRA project (http://www.coli.uni-sb.de/sfb378/negra-corpus/).

5. Phonetic transcription

From the orthographic transcription and the pronunciation lexicon an automatically generated phonetic transcript will be provided to the transcribers. They will have to verify this transcript and, if necessary, correct it. It is presently being tested whether it increases consistency to include certain between-word phenomena in the automatically generated transcript. It is expected that this phonetic transcription will require on average some 40 times real time. It should be kept in mind that no hand segmentation at the phoneme level will be provided. Instead, the one-million words that will be phonetically transcribed, will be aligned automatically with the speech signal (using speech recognition software) and checked manually at the word level. Figures about the transcription consistency or the word alignment accuracy are not yet available.

6. Prosodic transcription

Although the protocol for the prosodic transcription of a subset of 250,000 words from the corpus has not yet been fully established, it seems to be probable that, for practical reasons, this activity will be limited to indicating phrase boundaries and most prominent words. It will have to be tested whether it is not too much of an extra burden to the transcribers to ask them to indicate not so much the prominent words but actually the vowel that is most stressed in these prominent words, thus providing us with additional information about realized word stress (compared to lexical stress). Per phrase we may also ask for an indication of the local speaking rate (normal, versus extra fast or extra slow). Physical details about the actually realized intonation will not be scored, this is of course one of the many additions that could be made in subsequent studies.

7. Conclusions

At the time of writing of this short contribution (Oct. 2000), the second of seven releases has just been realized. The Spoken Dutch Corpus is thus a reality! The CD-ROMs can be ordered from the Dutch Language Union (http://www.taalunie.org/), who holds all rights, while future distribution may take place through ELRA. These first two releases don't contain yet phonetic, prosodic, or syntactic annotations, but mainly audio files with orthographic transcripts. Part of the data have been lemmatized and tagged with part-of-speech information. The meta data are yet incomplete but do provide information about the text samples and about the speakers. Also several frequency lists are included, like word frequency counts (also for the various components of the corpus separately), a list of POS tags and a lemma list. The annotation CD-ROM will be updated with every new release, whereas the package of other CD-ROMs will of course grow over the years, although new storage media may come instead. Via PRAAT, the recordings can be played and the orthographic transcripts can be viewed at the same time. Other exploitation software is presently being developed at the Max Planck Institute in Nijmegen according to CGN specifications. This should allow future users to access and search the data efficiently and with relative ease at various levels of representation.

It is too early yet to present results of studies that have used these data. However, given for instance our own earlier studies on various phonetic aspects of speech, and given our never ending complaints about insufficient data, we may expect that this large annotated database of spoken Dutch will contribute substantially to our future research. Also the protocols developed within the CGN-project will most probably appear to be extremely useful for other related projects, such as Rob van Son's post-doc project on "How efficient is speech? Studying the distribution of variation and information in speech" (Van Son et al., 2000), in which also a lot of speech data are collected and will have to be annotated and segmented at phoneme level. The same holds for the INTAS project proposal on "Spontaneous speech of typologically unrelated languages (Russian, Finnish and Dutch): Comparison of phonetic properties" (De Silva, 2000).

References

De Silva, V. (2000), “Spontaneous speech of typologically unrelated languages (Russian, Finnish and Dutch): Comparison of phonetic properties”, INTAS proposal.

Goedertier, W., Goddijn, S. & Martens, J.-P. (2000), “Orthographic Transcription of the Spoken Dutch Corpus”, Proceedings Second International Conference on Language Resources and Evaluation LREC-2000, Athens, Vol. 2, 909-914.

Oostdijk, N. (2000), “The Spoken Dutch Corpus. Overview and First Evaluation”, Proceedings Second International Conference on Language Resources and Evaluation LREC-2000, Athens, Vol. 2, 887-894.

Van Eynde, F., Zavrel, J. & Daelemans, W. (2000), “Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus”, Proceedings Second International Conference on Language Resources and Evaluation LREC-2000, Athens, Vol. 3, 1427-1433.

Van Son, R.J.J.H., Streefkerk, B. & Pols, L.C.W. (2000), “An acoustic profile of speech efficiency”, Proceedings International Conference on Spoken Language Processing ICSLP-2000, Beijing.