Orthographic Transcription of the Spoken Dutch Corpus

Wim Goedertier*, Simo Goddijn†, Jean-Pierre Martens*

* Electronics and Information Systems (ELIS),
University Gent,
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
{odul, martens}@elis.rug.ac.be

†Speech Processing Expertise Centre (SPEX),
Department of Language and Speech,
Catholic University of Nijmegen,
Erasmusplein 1, NL-6525 HT Nijmegen, the Netherlands
s.goddijn@let.kun.nl

Abstract

This paper focuses on the specification of the orthographic transcription task in the Spoken Dutch Corpus, the problems encountered in making that specification and the evaluation experiments that were carried out to assess the transcription efficiency and the inter-transcriber consistency. It is stated that the role of the orthographic transcriptions in the Spoken Dutch Corpus is twofold: on the one hand, the transcriptions are important for future database users, on the other hand they are indispensable to the development of the corpus itself. The main objectives of the transcription task are the following: (1) obtain a verbatim transcription that can be made with a minimum level of interpretation of the utterances; (2) obtain an alignment of the transcription to the speech signal on the level of relatively short chunks; (3) obtain a transcription that is useful to researchers working in several research areas and (4) adhere to international standards for existing large speech corpora. In designing the transcription protocol and transcription procedure it was attempted to establish the best compromise between consistency, accuracy and usability of the output and efficiency of the transcription task. For example, the transcription procedure always consists of a first transcription cycle and a verification cycle. Some efficiency and consistency statistics derived from pilot experiments with several students transcribing the same material are presented at the end of the paper. In these experiments the transcribers were also asked to record the amount of time they spent on the different audio files, and to report difficulties they encountered in performing their task.

1. Introduction

The Spoken Dutch Corpus, abbreviated as CGN[1] (from the Dutch name *Corpus gesproken Nederlands*) is intended to be an annotated speech corpus of about one thousand hours of continuous speech (10 million words). The project, which started in June 1998, will run for five years, and is a co-operation between several Dutch and Flemish universities [Oostdijk, 2000].

The corpus is to be a major resource for several research areas such as linguistics, phonetics, and language and speech technology. Therefore, it will contain material recorded in a variety of communicative settings: spontaneous face-to-face and telephone dialogues, interviews, discussions, debates, lectures, broadcast news and read book passages. As such, the corpus will be the largest and most diverse database of Dutch speech ever collected.

All the material will be orthographically transcribed and every word will receive a part-of-speech (POS) tag. In order to maximize the accessibility of the speech data, every word of the transcription will get a pointer to its position in the audio file. This indexing operation will be accomplished automatically for the entire corpus. For a selection of 10 percent of the data, a manual verification of the word pointers, a broad phonetic transcription, as well as syntactic annotations are envisaged. Prosodic annotation is envisaged for 250.000 words

The role of the orthographic transcriptions in the CGN is twofold. In the first place, the orthography is the most valuable piece of information for the future users, because it is the simplest symbolic representation of the speech file, and because together with the POS information, it is the only annotation that will be made available for all the speech material.

In the second place, the orthographic transcription is indispensable to the realisation of the corpus itself. It is the basis for every other layer of annotation that is added to the speech samples. Connection with lexical databases is enabled by this transcription and, for example, grammatical tagging and automatic word alignment fully depend on it.

For these reasons, it is of great importance that the specification of the orthographic transcriptions is well considered and that the quality of the transcriptions is high. All of the decisions made in the specification phase of the orthographic transcription should be interpreted in this light.

2. Objectives

The first objective has been to pursue a verbatim transcription requiring a minimum level of interpretation of the utterances (i.e. no correction of grammatical errors, no completion of truncated words, etc.).

The second objective has been to obtain a transcription that is aligned to the speech signal on the level of relatively small chunks. These chunks already enable

[1] The corpus produced by the CGN project will be the property of The Dutch Language Union. The Dutch Language Union is an intergovernmental organization, based on the 1980 Dutch Language Union Treaty between the Netherlands and Belgium. The aim of this treaty is the integration of the Netherlands and the Flemish community in Belgium in the field of Dutch language and literature in the widest sense.

focused access of the corpus: one can locate any desired word in a short stretch of signal covering just a few other words. Furthermore, the chunk level alignment offers a good starting point for the automatic alignment of the speech with its orthographic transcription at the word level. Finally, it is quite natural for the transcribers to process the long speech files (up to 20 minutes long) by selecting a short stretch of signal, by transcribing it and by moving on to the next chunk.

Since the CGN is intended to be a spoken language resource for several research areas, the third objective has been to construct an orthographic transcription that is beneficial to speech and language technologists as well as to linguists, lexicologists, phoneticians etc. For this reason, researchers working in these fields have been involved in the specification phase of the orthographic transcription. Decisions have been made only after extensive discussions between representatives of the different research areas had taken place. This does not mean however, that every decision has been made unanimously.

The fourth objective has been to adhere to current international standards for large spoken language corpora. The EAGLES [Gibbon et al, 1997] and CHILDES [Mac Whinney, 1999] documents on orthographic transcription have served as references during the specification phase, as well as the documentation supplied with several large speech corpora (including Switchboard from the Linguistic Data Consortium [LDC, 1994]).

Given these objectives, and keeping in mind the available budget, a set of criteria was defined that eventually led to the "Protocol for Orthographic Transcription" (after this: Protocol). Apart from that, a procedure was defined for the practical realisation of the transcriptions. This procedure is described in section 5.

3. Design Criteria

The orthographic transcription protocol had to be designed in such a way that an optimal consistency, accuracy and usability of the transcriptions can be expected.

Consistency is important from a logical point of view, i.e. identical situations should be transcribed identically, but it also facilitates searching the database. A linguist who is interested in the use of the interjective "hé" *(hey)* is helped by the fact that the word is spelled as <hé> consistently and not as <hee> or <hey>. Also, consistency facilitates the automatic processing that is necessary to realise further annotations, like grammatical tagging.

The quality of a spoken language resource depends largely on the accuracy of the transcriptions. Therefore, the transcription protocol should clearly specify what phenomena to transcribe and how to do it. It should leave little or no room for interpretation. Future users (or the external evaluator) of the corpus should have at their disposal a set of well-defined rules against which they can evaluate the transcriptions. This is also a necessary precondition for the automatic generation of phonetic transcriptions that is foreseen for all of the data.

Usability is the third criterion that must be satisfied. For example, a speech technologist who wishes to train a model for a certain word should be able to rely on the fact that words that are heavily regionally accented have received a special code, so that they can easily be left out of the training material to prevent the models from being contaminated. Again, the realisation of the corpus will benefit from this as well: knowing the properties of the transcriptions helps in automatically processing the data.

On the one hand, it is important to construct a set of rules that help in obtaining transcriptions that meet the above design criteria. On the other hand, one has to keep in mind that the work, i.e. making the transcriptions, is to be done by human beings (in our case by students), who are by definition fallible. A number of measures were taken in order to minimise the chance of errors. They are described below.

In order to attain consistency, all words in the transcription are checked on line against the orthographic entries of a lexicon. If the check fails, the transcriber is required to mark the word with one of the special codes mentioned in the Protocol. Furthermore, it was decided, as suggested in the EAGLES handbook as well as in the LDC manual for Switchboard, to make use of the conventional spelling as much as possible: it is thought easier for the transcribers and therefore beneficial to the consistency of the transcriptions.

For the same reason of consistency however, it was decided to deviate from the conventional spelling in a small number of cases, for example in the use of punctuation. The punctuation in the CGN is restricted to sentence endings, which means that no commas, colons, semicolons etc. are used. Not only is it very difficult to have these punctuation marks transcribed consistently, they also require a certain level of interpretation by the transcribers (which does not agree with the first objective outlined in section 2).

It is attempted to increase accuracy by having every transcription checked in an extra transcription cycle. From experiments in het start-up phaseof the CGN-project we know that a number of errors can be detected at low cost during this second cycle. Not only the transcription itself is checked in this extra cycle, also the positions of the time markers, the match of the words against the lexicon and the attribution of utterances to the correct speakers in dialogues and multilogues (i.e. conversations containing more than two speakers) are checked.

The number of rules in the Protocol is kept as small as possible. This makes it easier for the transcribers to memorize them and apply them correctly. The Protocol also comprises a lot of examples illustrating the rules. By testing the Protocol and by recording the transcribers' feedback, ambiguities in the rules were discovered and the Protocol was improved. It took three iterations to arrive at the Protocol that is actually used.

An internal system of bug reporting from further annotation layers back to the orthographic transcription and the use of post-processing scripts to check the syntax of the transcriptions should give an extra guarantee for consistency, accuracy, and in concordance with the Protocol.

4. Pilot Transcriptions

4.1. Objectives

During the start-up phase of the CGN-project, closely monitored pilot transcriptions were carried out in Flanders

as well as in the Netherlands, using a test version of the Protocol (that still contained the use of commas). The purpose of these transcriptions was twofold. Firstly, they enabled an evaluation of the Protocol of that time in terms of clarity, ambiguity, completeness and usability. Secondly, they made it possible to get an estimate of the inter-transcriber consistencies and of the time it takes to transcribe a certain amount of speech, both in relation to the speech style and communicative setting of the speech material to be transcribed.

4.2. Speech material and task

Fragment	Description
A	read speech (text available, monologue)
B	prepared lecture (text available, monologue)
C	live radio coverage of a soccer game (monologue)
D	debate in the Lower House (simple multilogue)
E	radio debate with reactions of callers (difficult multilogue)
F	informal meeting with discussion (very difficult multilogue)
G	multilogue during lunch/reception (very difficult multilogue)

Table 1: Overview of the test fragments

A number of speech fragments representing different speech styles (from read speech to spontaneous comments) and communicative settings (from monologues to multilogues with 4 or more speakers) were selected for transcription, in Flanders as well as in the Netherlands. Table 1 gives on overview of the selected fragments: fragments C and G come from the Flemish pilot and the remaining fragments come from the Dutch pilot. The Dutch fragments were transcribed by 5 Dutch students and the Flemish fragments were transcribed by 6 Flemish students.

For the fragments C to G, a transcription had to be produced from scratch, whereas a text was already available for the fragments A and B. For these fragments, the text had to be checked and it had to be made conform the Protocol. In the discussion of the results, a distinction is made between the two different tasks. Insertion of time markers was part of the task for all of the fragments.

4.3. Results

Table 2 gives an overview of the results of the pilot. The first row listst the average time needed for a transcriber to produce a transcription, expressed in "times real-time" (for example, the figure 8.7 indicates that is takes 8.7 minutes to transcribe 1 minute of speech). Clearly, the speech fragments A and B are not to be compared with the remaining speech fragments, because of the difference in transcription task. A comparison between the transcription time needed for fragment A and fragment B and a comparison between the transcription time needed for fragments C to G indicates that the factor increases with the difficulty of the transcription task.

To obtain an idea of the inter-trancriber consistency the transcriptions of different transcribers were aligned with one-another using a Viterbi-alignment procedure with a cost function measuring differences between words. Every transcription was aligned with the remaining transcriptions of the same fragment, and the average percentages of discrepancies between aligned transcriptions were measured: deletions, insertions and substitutions of ordinary words and punctuation marks respectively. The results obtained for the different fragments are listed in rows 5 and 6 of table 2.

The first thing to be observed is that the use of punctuation marks is not very consistent: for the fragments C to G the percentage of sentence ending differences is (far) over 30%. Furthermore, it can be seen that the number of word differences is very low for the fragments A and B, for which a prompt text was available during the transcription. It is considerably higher for the fragments C, D, E, F and G, fragments containing more or less spontaneous speech. The very high percentage for sample F are due to the fact that the transcriptions were aligned speaker by speaker and that different transcribers attributed the same speech to different speakers.

4.4. Conclusions

Test sample	A	B	C	D	E	F	G
Transcription time (times real-time)	8.7×	14.1×	23.4×	24.8×	33.4×	38.7×	47.0×
% of sentence ending differences	1.1 %	9.3 %	31.3 %	69.0 %	64.7 %	89.1 %	37.6 %
% of word differences	1.1 %	3.1 %	8.8 %	10.6 %	9.7 %	92.6 %	30.4 %

Table 2: Results of the pilot transcriptions

The closely monitored pilot transcriptions made clear that differences in the use of punctuation marks between transcribers are very large. It was concluded that obtaining a satisfactory consistency in the use of punctuation marks, especially for commas, is too ambitious a goal for the CGN. In consequence of this result, it was decided not to use commas in the CGN-transcriptions. The Protocol was changed accordingly.

The large number of speakers in fragments F and G created an extra problem for the transcribers: not only were they confronted with overlapping speech, they also were to recognise speakers by their voices, which proved to be a very difficult task (especially since the transcribers were not acquainted with the speakers). It was decided that the maximum number of speakers in spontaneaous speech fragments for the CGN should be four.

5. The Protocol

The "Protocol for Orthographic Transcription" comprises a set of rules defining what exactly should be transcribed and how speech fragments should be transcribed. The EAGLES guidelines have played a

decisive role in the development of these rules. The pilot experiments described in the previous section have had an influence on the improvement of the rules.

As mentioned before, it was decided to stay close to the ordinary spelling conventions. It is assumed that transcribers are acquainted with this spelling. Although some of the rules in the Protocol are included just to remind transcribers of the correct spelling or to give guidelines for the level of detail expected in the transcriptions, the main part of the Protocol contains rules for situations in which the transcriber is asked to deviate from the conventional spelling.

One of the more conspicuous deviations is that capitals are reserved for proper names, titles (of books, films, etc.), abbreviations and acronyms. No capitals are used at the beginnings of sentences. An interesting advantage of this approach is that it suffices to include words in the lexicon only once, in lower case. Another advantage is that removing or inserting a sentence ending (e.g. during the verification cycle) does not require any modification of the transcribed words.

A second deviation from conventional spelling that was already mentioned is the use of punctuation. Punctuation is restricted to full stops, question marks and continuations. Within-sentence punctuation and exclamation marks were considered too prone to interpretation to be included in the transcriptions. In the pilot stage of the CGN-project, we did use commas, but they showed to be responsible for a substantial portion of the disagreement between transcribers. It was therefore decided to leave them out.

Because spoken language differs substantially from written language, conventional spelling lacks the possibilities to adequately transcribe certain spoken language features that are nevertheless presumed important to future users. Examples are truncated words, mispronunciations, and words that may be misperceived. For these features, the following convention has been adopted: the word is supplemented with an asterisk (*) and a character specifying the feature in question. The words having an asterisk code are called marked words. (Table 3 lists the features that are marked.) They are important to the lexicon builder who has to decide whether and how to include a word in the lexicon. They are also important to the linguist who is looking for the frequency of occurrence of mispronunciations, foreign words, dialectal words, etc. in relation to the type of speech and the social background of the speaker. The marked words can also be of help during the training of the transcribers, because listing just these words found in the transcriptions produced by a transcriber can bring to light possible problems with the interpretation of particular rules.

The transcription of clearly audible non-linguistic sounds produced by the speaker (coughing, laughing, etc.) is accounted for, but for reasons of efficiency and consistency, no differentiation between these sounds is envisaged. Clearly audible background noises or background noises that clearly have an influence on the course of a conversation are accounted for by means of an extra background tier (see section 5.1 for an explanation of the notion "tier"). In addition, the transcribers are

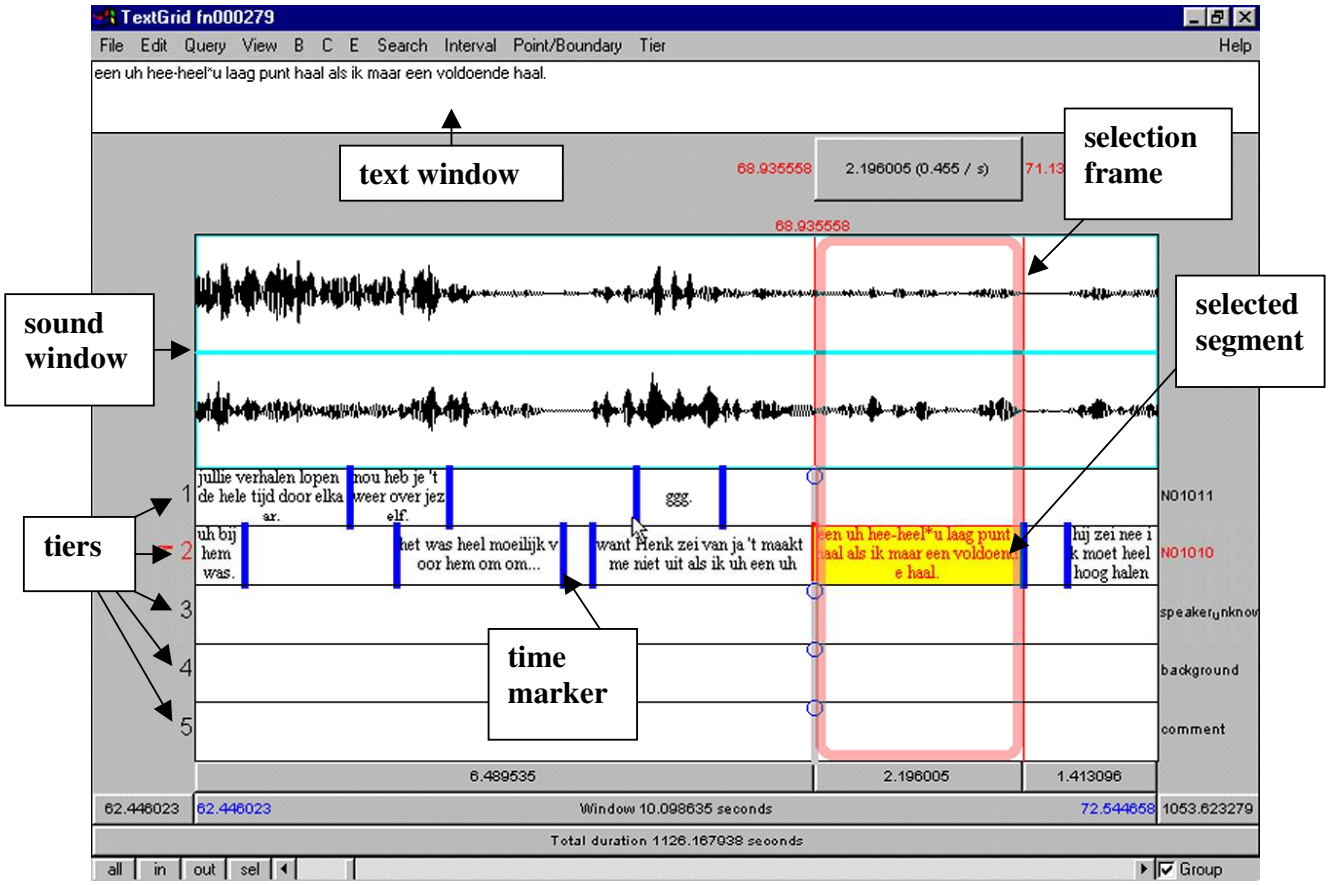

Figure 1: Screen dump of the transcription interface of PRAAT

allowed to make remarks about the recording in general and about background noises that characterise the recording as a whole.

The decision not to categorise background noises and not to transcribe less conspicuous background noises was taken in full appreciation of the fact that it may hinder the automatic alignment of the speech signal and the transcriptions, as well as the automatic generation of phonemic transcriptions. For reasons of cost efficiency however, it was not possible to include such detailed transcriptions of non-speech events.

6. Transcription Procedure

6.1. The interface

Transcriptions are made with the help of the interactive signal-processing tool PRAAT[2]. Figure 1 shows a screen dump of the transcription interface of this program.

As can be seen in the figure, transcribers are presented with a visual representation of the acoustic signal (in this case a stereo signal) next to the auditive representation. Below the acoustic signal, five tiers can be distinguished in the figure. The first two tiers are meant to contain the transcriptions of the two speakers in this speech fragment. The third tier, named *speaker_unknown*, is meant for transcriptions of incidental speech that cannot be attributed to one of the known speakers, or speech that the transcriber can not attribute with certainty to one of the known speakers. The fourth tier, named *background*, is meant for the transcription of clearly audible background noises, or background noises that clearly have an influence on the course of the conversation. An example of the last situation would be the sound of footsteps, followed by the sound of an opening door and a "new" voice, joining in the conversation. The last tier, named *comment*, is meant for remarks about the recording in general and about background noises that characterise the recording as a whole. Representing simultaneous or overlapping speech is no problem. As is apparent from figure 1, each speaker is assigned a tier and in each tier time markers can be inserted independently. The number of tiers is varable and corresponds to the number of speakers in a fragment. Note that a speaker code is inserted next to each speaker tier (in the figure: N01011 and N01010).

6.2. The procedure

As mentioned before, every transcription is made in two cycles. The exact nature of the cycles may vary ans is dependent onthe different types of speech. Read speech for instance, characterised by the fact that a prompt text is available and can be used (after some pre-processing) as a first transcription. The second cycle for this type of speech then consists of checking and improving the first transcription, inserting time markers and matching the words to the lexicon.

For the remaining types of speech, sometimes a script is available that may serve as a first transcription (for example the script for a lecture). More often, the fragments are to be transcribed from scratch. In that case, a full transcription according to the Protocol is made in the first cycle, including the insertion of time markers, the match of the words in the transcription against the lexicon and the assignment of the utterances to speakers. During the second cycle, an independent check (preferably by another transcriber) of all of these actions is performed.

The transcriptions are manufactured mainly by students, within one of the participating institutes. The students are not necessarily trained in language or speech.

For part of the data (the monologues and dialogues, not the multilogues), an external word processing agency has been subcontracted to create the first cycle transcriptions. This agency supplies transcriptions produced in a regular word processor, which means that they do not yet contain any time markers. The agency does have the CGN-lexicon at its disposal, so that a match of the words against the lexicon is included in their transcriptions. They also attribute the utterances to the different speakers. The second cycle for this kind of transcription therefore consists of the insertion of time markers, a check of the transcription, a check of the match of the words against the lexicon and a check of the attribution of the utterances to the different speakers.

7. Acknowledgements

This publication was supported by the project "Spoken Dutch Corpus (CGN)" which is funded by the Netherlands Organization for Scientific Research (NWO) and the Flemish Government (AWI).

8. References

Oostdijk N, 2000. *The spoken Dutch Corpus. Overview and first evaluation*. Proceedings LREC-2000 (this issue).

Gibbon D, Moore R, Winski R, 1997. *Handbook of Standards and Resources for Spoken Language Systems.* Mouton de Gruyter (Den Haag).

LDC, 1994. *Switchboard: a user's manual.* http://www.ldc.upenn.edu/readme_files/switchboard.readme.html

MacWhinney B, 1999. *The CHILDES Project : Tools for Analyzing Talk (2 ed.).* Hillsdale, NJ : Lawrence Erlbaum Associates

[2] For more information on PRAAT see http://www.praat.org/