Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 18 (1994), 87-94.

TRANSLITERATION OF THE DUTCH SPEECH STYLES CORPUS

Els A. den Os *

Abstract

In this paper we present the transliteration procedure that has been used for the so-called Dutch Speech Styles Corpus. Some decisions made before and after the process of transliterating which are related to Dutch, are mentioned. Some recommendations are given for future transliterations of large speech corpora containing spontaneous speech.

1 Introduction

In recent years a number of American-English (telephone-)speech corpora have been published (e.g. TIMIT, 1986; Resource Management, 1988; Switchboard, 1992; ATIS, 1992; Macrophone, 1994), as well as a number of European ones (e.g. Eurom.1, to appear in 1995; Polyphone, 1994; Rafael.0, 1994; TED, 1994). Speech corpora can be used for research and development in many different fields, such as speech recognition, speech understanding, speaker recognition, speech synthesis, basic speech research, psycholinguistics, sociolinguistics, and audiology. A speech corpus never is an aim in its own right; it is always intended to serve independently motivated purposes. However, a well-documented and annotated corpus may very well be suitable for other, additional (research) purposes.

Transcription of the speech is a very important part of the annotation of a speech corpus. Without a transcription a corpus has almost no practical value. For large corpora the transcription will most often be limited to a transliteration (i.e. an orthographic transcription); a more precise phonemic or even phonetic transcription can only be provided for smaller corpora. However, TIMIT, a medium sized corpus of 630 speakers each reading 10 sentences, is provided with phonemic transcription and labelling.

In order to make a large corpus optimally suitable for e.g. initialisation in speech recognition it is necessary that all audible events in time are represented. Thus, the speech events as well as the non-speech events (all kinds of noises) must be written down. Only when this has been done, training of speech recognition systems or automatic segmentation programmes will get useful results. Especially for the development of speech and speaker recognition systems large speech corpora are very important. This means that a transliteration must be used that is based on an orthographic representation of the speech (a phonemic or phonetic transcription would be too time-consuming and too expensive), combined with an optimally precise account

* Also Speech Processing Expertise Centre (SPEX)

of all acoustic events occuring in time. Thus, a *balance* must be found between what is still feasible for a transcriber and the exactness of a representation.

Most corpora mentioned above were transcribed orthographically. TIMIT, Resource Management, and parts of ATIS and Macrophone contain *read* speech. Switchboard and parts of Macrophone and ATIS contain *spontaneous* speech. In ATIS, the spontaneous speech relates to information requests for flights. Switchboard consists of recordings of telephone conversations by a large number of couples who agreed to carry on a conversation on one of a small number of topics. Next to the orthographic transcription, ATIS and Switchboard used conventions to indicate audible events that may occur during speaking. These conventions consist of different kinds of brackets with or without additional information, e.g. [uh] for a hesitation pause, [laughter] for laughing, angle brackets for verbally deleted words (<show> <me> show me the flights to Boston).

Below we will discuss the transcription procedure used for a medium sized corpus of Dutch spontaneous speech known as 'Speech Styles'. This procedure very much relies upon the ones used in ATIS and Switchboard. An adaptation was made for Dutch. When people are speaking spontaneously, they frequently use reduced word forms and cliticised words. By 'reduced word forms' we mean reduced forms (mostly vowel reduction has taken place) of single words, not resulting in syllable deletion (e.g. *mijn* [mEin] -> *m'n* [m@n]). By clitisation we mean that two words behave as one phonological word, and syllable deletion may take place (e.g. *het is* [hEt Is] -> *'t is* [tIs], but also *maak ik* [ma:k?Ik] -> [ma:gIk]. There were two reasons why we treated these forms differently, i.e. not obeying the demands of orthographic representation in the Dutch dictionary of Van Dale (1992). The first reason is that we suppose that automatic segmentation programmes and speech recognition programmes get into trouble if these reduced forms are not provided in the transliterations, e.g. if syllables are missing as in English *it's* for *it is* or Dutch *da'k* for *dat ik* ('that I'). The other reason was that the reduced or cliticised words are frequently occurring in spontaneous speech. For the sake of convenience for the transcriber, who also might find it counterintuitive to write down full forms for reduced forms, these forms were allowed to be transcribed as reduced forms (although these were sometimes not present in the dictionary). It turned out that the number of these words was limited. Furthermore, the transcribers indicated whether filled pauses were attached to preceding words (*en[uh]* 'and[uh]') or not (*en [uh]* 'and [uh]'). In the first case, transcribers were allowed to write *enne*, *datte*, etc. This too is against the orthographic demand, but also here the frequency of occurrence of these forms is supposed to justify this decision.

In this paper, we will first give a short description of the Speech Styles corpus, then we will discuss the transcription system, and we will present the list of reduced and cliticised forms. In addition we will give some information about the number of occurrences of these forms.

2 Speech Styles Corpus

The Dutch Speech Styles Corpus was collected to investigate the voice quality of speakers of standard Dutch. The speech material was designed by R. van Bezooijen and the speech recordings were made by J. van Rie and R. van Bezooijen. The corpus contains three different speech styles: spontaneous speech (monologues), semi-spontaneous speech (picture descriptions), and read speech. The speech was always recorded in the presence of an 'interviewer', who only intervened if the monologues flagged. In all three styles the speech contents refer to domestic topics, eating habits, and food. The vocabulary in the three styles is therefore very much identical. There are

127 speakers, 60 male and 67 female, in three age groups: under twenty (30 speakers), between twenty and sixty (45 speakers), and over sixty (52 speakers). The speech has been transliterated by one transcriber, and has been completely checked by another person. In total the duration of the speech is more than 19 hours (4.66 hours of monologues, 10.35 hours of picture descriptions, and 4.19 hours of reading texts). The total number of word forms that were transcribed in the corpus amounts to about 118,000. There are about 6,300 different word forms. For example, the first person singular of the verb 'to walk' (*loop* [lo:p]) and the third person singular of 'to walk' (*loopt* [lo:pt]) are counted as two word forms, whereas the homograph *zijn* (1st, 2nd, and 3rd person plural of the verb 'to be' and the full verb 'to be', as well as the possessive pronoun 'his') is counted as one word form.

3 Transliteration conventions

The transliteration of the Speech Styles Corpus is a word level transliteration of what the speakers said. A normal orthographic notation is used. This necessarily implies a compromise between the sounds heard and what has to be written down. We decided to use next to the full words listed in the dictionary, a controlled (but extendable) set of word forms frequently used in spontaneous speech. Before the transliteration started we did not know whether this set of reduced forms contained the right and only forms. We also did not know what types of clitisation would occur. We therefore used this corpus to develop a list of 'special' forms. We explicitly asked the transcriber and the person who corrected the transcriptions to take notice of these forms. In section 4 we will discuss the set of forms in more detail.

The transliteration was performed in two steps. Firstly, the transliteration was done at the lexical level. Secondly, labels for pauses, filled pauses, non-speech sounds, lengthening of sounds, etc. were applied. The Speech Styles Corpus has been labelled at the utterance level (i.e. a time stamp between the utterances is provided that allows to access the speech in the files). The notion of what constitutes an utterance in spontaneous speech is necessarily an arbitrary one. We chose for a rather global segmentation, based on pauses, syntax, and semantics. An utterance was defined as a number of words being semantically consistent and containing at least a subject and a verb. In addition, this string had to be preceded and followed by a clear acoustic pause.

3.1 Markings

a) Case
Transcriptions are in lower case. Proper names begin with a capital. The beginning of an utterance is *not* indicated by a capital.

b) Spelling
Standard spelling is used. No hyphens are used (in Dutch spelling hyphens are used in a great number of compounds, but in the corpus hyphens are used to indicate cliticised words, see section 4). Abbreviations and numbers are spelled out as pronounced.

c) Punctuation
No standard punctuation is used. Punctuation is used to indicate prosodic factors: periods, preceded and followed by a space indicate silent pauses; periods may occur within words, but then the period is surrounded by hyphens (-.-). Exclamation marks are used to indicate emphatic stress on the following syllable. Commas are used to

indicate intonational separation (continuation rise). The colon is used to indicate lengthening of the preceding sound. This type of prosodic information is incorporated, since it may help the automatic segmentation and recognition programmes.

d) Mispronunciations
Obviously mispronounced words that are nevertheless intelligible are marked with stars, *blabla*.

e) Deletions
Only for the read texts: words in the written texts which are not spoken are put between round brackets. For spontaneous speech, it is almost impossible to know whether a speaker deleted words.

f) Verbally deleted words
Words verbally deleted by the subject are enclosed in angle brackets. Verbal deletions are words spoken by the speaker but which are superseded by subsequent speech explicitly (show me the <flights <I> <mean> fares) or implicitly (show me the <flights> fares to Boston).
Word fragments are also indicated with angle brackets <fl> flights.

g) Words spoken by the interviewer
Words spoken by the interviewer are indicated with curly brackets { }.

h) Simultaneous speech by speaker and interviewer
Simultaneous speech by speaker and interviewer is indicated with # before and after the words that are spoken at the same time.

i) Unintelligible words
Partly unintelligible words are placed between double brackets (()). If a word is completely unintelligible, the brackets are typed with an empty space in it: (( )).

j) Filled pauses
Filled pauses are indicated by [uh], [um], [naa], or [mm].

k) Non-speech sounds
Non-speech sounds can be: [laughter], [smack], [cough], [loud_breath], [grunt], or [noise]. Noise is meant to indicate non speech noise; since this corpus was recorded under fairly good conditions, it was not necessary to divide noise into subcategories, as is often done for telephone speech. For the above mentioned events the beginning and end relative to spoken words are indicated as follows:
[laughter/]I don't like[/laughter] tomato soup
([laughter/] indicates the beginning of laughter and [/laughter] the end).

4 List of reduced and cliticised forms

4.1 Reduced forms

An orthographic transcription means that all words must be written down in their dictionary form. For read speech, this is mostly not problematic. However, for spontaneous speech, in which reduced forms often occur, this may be inconvenient for transcribers. For example, they have to write down *eens* [e:ns] for *'s* [@s].

Sometimes, the reduced form is no longer felt to be a reduced form of a full form, but it is felt to be more or less lexicalised, like *'t* instead of *het*. Thus, it would be pleasant for transcribers if they could use a list of reduced forms. Which forms have to occur in this list and which not? We defined two criteria:

1) the forms are more or less lexicalised, i.e. they are present in the dictionary (as reduced or cliticised forms) and/or do appear in written texts.
2) the forms frequently occur in spontaneous speech.

The forms we defined in the Speech Styles Corpus are presented in table 1.

Table 1. List of reduced word forms. In the first column, the forms (with the phonemic form) used in the corpus are given. In the second column the full form is given. The third column gives the number of occurrences in the corpus of reduced as well as full forms, and the fourth column presents the English translation. (In the case of *z'n*, we cannot decide on the frequency of occurrence of the full form, since this form is ambiguous (*zijn* can be the verb 'to be', as well as the possessive pronoun 'his').

Reduced word form	Full form	Number of occurrence reduced - full	English
ie [i]	hij	39-104	he
m'n [m@n]	mijn	216-180	my
z'n [z@n]	zijn	223-?	his
effe [Ef@]	even	7-103	quickly
as [As]	als	278-100	when, if
'r [@r]	er	313-173	there
d'r [d@r]	daar/haar	986-319	there/her
't [@t]	het	3143-104	it (article)/personal pronoun
'm [@m]	hem	21-18	him
's [@s]	eens	51-12	

It can be observed that the length of this list is limited (10 items). Since the form *effe* does not occur very often, the list can be reduced to 9 items. In most cases there also exists an (official) spelling for these reduced forms (not for *as*), and these forms occur mostly more frequently than the full forms. In these reduced forms, syllables are not deleted, but the vowels are mostly reduced to schwa. For *'r*, *'t*, *'m*, and *'s* the spelling as given in the table cannot be used in the transliteration, because it would suggest syllable deletion which is not true. Therefore, these words are transcribed as *ur*, *ut*, *um*, and *us*. This list of reduced forms is not the 'final' one. Other corpora containing spontaneous speech, e.g. dialogues, will certainly add new forms to this list.

4.2 Cliticised forms

Next to the list of reduced forms, we gave the transcribers the possibility to transcribe cliticised words. The idea behind this was, that these forms, which do not occur in the dictionary, occur frequently in spontaneous speech. Especially the cases in which syllables are deleted when words are spoken together are of significance in relation to automatic segmentation programmes and training of speech recognizers. Since we were not sure which clitic groups would occur, we gave the transcribers the possibility to indicate clitisation by means of a hyphen between the two full words. Clitisation not always implies syllable deletion, also voicing of final obstruents of the first words or glide insertion may indicate clitisation. The transcribers could also use these indications to decide on the presence of clitisation. On second thoughts, we think that only the

cases in which syllables are deleted must be transliterated in this way, since these cases are important for automatic segmentation. In table 2, we present the five cases which occurred most often in the corpus. There are about 570 cliticised forms, 429 of these involve syllable deletion; however, a number occurred only once or twice.
Especially the forms including the personal pronoun *ik* (I), the verb form *is* (is), and the personal pronoun *het* (it) involve syllable deletion.

Table 2. Cliticised words and the frequency of occurrence in the corpus of 118, 000 words

Clitic group	frequency of occurrence
xx-ik *(die-ik (die'k), kan-ik (kan'k))*	67
ik-xx *(ik-weet ('kweet), ik-heb ('kheb))*	45
xx-is *(dat-is (da's), dit-is (di's), wat-is (wa's))*	141
xx-het *(aan-het (aan't), als-het (as't))*	29
het-xx *(het-is ('tis), het raam ('t raam))*	147

4.3 Pause forms

A distinction was made between filled pauses which were uttered in isolation and those which were attached to words. For a filled pause to be isolated, a pause must occur before it and a glottal stop may be heard. In the transliteration the [uh] is transcribed preceded and followed by a space. In the second case, the uh-sound is uttered directly after the word. In the transcription the uh-sound is directly typed after the preceding word. It turned out that this almost only happened after function words, not after content words.

There were four possible filled pauses: [uh], [um], [mm] and [naa]. In table 3 the number of isolated filled pauses and attached filled pauses are presented. For the more frequent functions words, like *en*, *wat*, etc. the transcribers were allowed to write *enne*, *watte*, instead of *en[uh]* or *wat[uh]*. The number of these forms are presented in the last colmun of the table.

Table 3. Number of isolated filled pauses, of attached filled pauses, and of frequently occurring function words ending in schwa.

	isolated	attached	frequently occurring function words ending in schwa
uh	1558	1354	632
um	694	57	
naa	182		
mm	147	72	

It can be observed that the most frequently used filled pause is [uh]. If we consider the cases like *enne*, and *watte* to be the same as filled pauses attached to words (which seems reasonable), the number of [uh]'s attached to words exceeds the number of [uh]'s in isolation. Furthermore, it was noticed that pauses *within* words were seldom uttered: only 20 times.

5 Recommendations for the transliteration of spontaneous Dutch speech

- Use as much of the orthographic standard spelling of the language as possible. In the present case we asked NOT to use the hyphen in the usual way, but to reserve this for cliticised forms. This was rather difficult for the transcribers. They frequently used the hyphen to indicate compound words (as is usual for a number of compounds in Dutch), whereas they had been instructed <u>not</u> to use hyphens in this case, and to write two words separated by a space.

- Use a list of reduced forms. These forms are frequently used in spontaneous speech and are more or less accepted in the spelling.

- If two words are cliticised and one or more syllables are deleted, this must be indicated by a single quotation mark (as is usual in Dutch). If no syllable deletion is involved, the orthographical forms or the reduced forms are used.

- Since filled pauses are frequently uttered directly after a word (no pauses), it is better to indicate these pauses differently from filled pauses in isolation. It is not necessary to handle words like *enne*, *watte*, *danne*, etc. differently.

In the appendix we present examples of transliterations.

Acknowledgments

The comments of Lou Boves and Louis Pols are kindly acknowledged.

References

Bernstein, J., Taussig, K. & Godfrey, J. (1994): "Macrophone: An American English telephone speech corpus for the Polyphone project", *Proceedings ICASSP'94*, Adelaide: I-81 - I-83.

Damhuis, M., Boogaart, T., In 't Veld, C., Schelvis, W., Bos, L. & L. Boves (1994): "Creation and analysis of the Dutch Polyphone Corpus", *Proceedings ICSLP'94*, 4: 1803-1806.

Hirschman, L. (1992): "Multi-site data collection for spoken language corpus", *Proceedings ICSLP'92*. 2: 903-906.

Lamel, L.F., Schiel, F., Fourcin, A., Mariani, J. Mariani & Tillman, H. (1994): "The Translaaguage English Database (TED)",*Proceedings ICSLP'94*, 4: 1795-1798.

Price, P., Fisher, W.M., Bernstein, J. and D.S. Pallet (1988): "The DARPA 1000-word resource management database for continuous speech recognition", *Proceedings ICASSP'88* : 651-654.

Rosenbeck, P., Baungaard, B., Jacobson, C. & Barry, D (1994): "The design and efficient recording of a 3000 speakers Scandinavian telephone speech database: Rafael.0", *Proceedings ICSLP'94*, 4: 1807-1810.

Van Dale, (1992): *Groot Woordenboek der Nederlandse taal*. Van Dale Lexicografie, Utrecht-Antwerpen.

Weatley, B., Doddington, G., Hemphill, C., Godfrey, J., Bolliman, E.C., McDaniel, J. & Fisher, D. (1992): "Robust automatic time alignment of orthographic transcriptions and unconstrained speech", *Proceedings ICASSP'92*, 1: 533-536.

Appendix

Examples of transliterations:

en bovenaan . <de> <etalage> <aan> <ut[uh]> . <bovenkant> vastgemaakt hangen . rechts een worst in ut midden . ook een worst <en> <aan> <ut> [um] . <nee> <sorry> . [loud_breath] !links een worst . in ut midden een rookworst . [loud_breath] en aan de rechterkant een ham

[smack] daarnaast . maar'k weet niet precies wat dat zijn twee kleinere taartjes twee tulbanden ofzo . [loud_breath] en daarnaast liggen twee: . stokbroden

<in> <de> <zuivelwinkel> misschien wil je ut woord melkboer weten . weet ik niet . [loud_breath] [um]

de: linker vrouw staat . half afgekeerd . en kijkt de andere vrouw aan en !die vrouw kijkt <ons> . ons weer aan <kijk>

e:n[uh] . ik heb rooie gordijnen met: . roo-.-d geel en paarse: . blokken durop

d:ie koop ik . puur . voor . bot-.-ontkalking tegen te gaan