interspeech 2008 incorporating SST 08

Preparing a Corpus of Dutch Spontaneous Dialogues for Automatic Phonetic Analysis

Barbara Schuppler[1], *Mirjam Ernestus*[1,2], *Odette Scharenborg* [1], *Lou Boves*[1]

[1]Center for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
{B.Schuppler, O.Scharenborg, L.Boves}@let.ru.nl, Mirjam.Ernestus@mpi.nl

Abstract

This paper presents the steps needed to make a corpus of Dutch spontaneous dialogues accessible for automatic phonetic research aimed at increasing our understanding of reduction phenomena and the role of fine phonetic detail. Since the corpus was not created with automatic processing in mind, it needed to be reshaped. The first part of this paper describes the actions needed for this reshaping in some detail. The second part reports the results of a preliminary analysis of the reduction phenomena in the corpus. For this purpose a phonemic transcription of the corpus was created by means of a forced alignment, first with a lexicon of canonical pronunciations and then with multiple pronunciation variants per word. In this study pronunciation variants were generated by applying a large set of phonetic processes that have been implicated in reduction to the canonical pronunciations of the words. This relatively straightforward procedure allows us to produce plausible pronunciation variants and to verify and extend the results of previous reduction studies reported in the literature.

Index Terms: corpus creation, conversational speech, spontaneous dialogues, reductions, pronunciation variants, automatic phonemic transcription

1. Introduction

When studying different types of speech corpora, one will encounter specific phenomena that are characteristic for the considered speech style. In spontaneous speech, words are often reduced compared to their canonical representations. Extremely reduced forms, which imply multiple syllable deletions, are quite frequent in conversational speech. The following Dutch examples demonstrate cases where only the stressed syllable survived the reduction processes [1]:

eigenlijk	'E+k	actually
bijvoorbeeld	'vOlt	for example
natuurlijk	'tyk	naturally

These examples may seem to be unintelligible, but psycholinguistic experiments show that listeners are capable of restoring the missing acoustic information when hearing extreme reduced words in context [2]. However, much higher in frequency are less extremely reduced forms. A study on American English shows that, while syllable deletions occur in 6% of the words, segment deletions occur even in every fourth word [3]. Recent linguistic research confirms the high frequency of reductions and focuses on the various different levels of reductions, from shortening of segment duration and lenition processes [4] to the deletion of syllables [5] to the complete absence of words [6].

The present study reflects a strong interaction of linguistic research with automatic speech recognition (ASR). On the one hand, our method for analyzing reductions makes use of an ASR system, which facilitates the exploration of large speech corpora and allows to quantify reduction trends. Also, a better understanding of reductions can improve automatic speech recognition itself. The acoustic characteristics of many words in conversational speech do not match their canonical representations in pronunciation lexicons. Saraçlar et al. showed that pronunciation variability correlates with recognition error rate. They recorded and transcribed conversational data, which then afterwards was, for a second recording, read by the same subjects. The error rate for the conversational data was more than 50% higher than for the read version [7]. Obviously, reduction implies diminished discrimination (or higher confusability). Adding reduced variants to an ASR lexicon without appropriate measures for dealing with increased confusability has adverse effects. Adding variants can only help in conjunction with accurate estimates of the conditions under which specific reductions are likely to occur. And to obtain such estimates we need large amounts of data about conversational speech.

The aims of this paper are two-fold. First, we describe the steps that were needed to make a corpus of 15 hours of spontaneous Dutch dialogues (collected by Ernestus [1]) accessible for automatic phonetic research using an ASR system. Secondly, initial results of testing the frequency of reduction rules in spontaneous speech will be presented as an example of the possibilities offered by automatic processing. In the literature different approaches can be found for analyzing reductions in spontaneous Dutch. While Ernestus chose an impressionistic method for her study [1], Van Bael used mixed effect models to analyze which factors may affect phone and syllable deletion [8]. The approach proposed here extends the data-driven approach of Van Bael, in that we include reduction phenomena that can only be accounted for by information about the stress patterns of the words. As Van Bael, we generate pronunciation variants by applying reduction rules to the canonical pronunciations of all words occurring in the corpus. Subsequently, a forced alignment is carried out with both a lexicon of canonical pronunciations and a lexicon including the pronunciation variants. The comparison of these two alignments gives information about which pronunciation variants were realized by the speakers.

The rest of this paper is organized as follows. In Section 2, the corpus of Dutch spontaneous dialogues, is described. The methodological approach is outlined in Section 3, where the focus will be on the preparation of the corpus for ASR systems and on the building of pronunciation variants. Then, in Section 4, the first results obtained with our method are presented. In

Accepted after peer review of full paper

	Duration	Nb of Chunks
Total length	123,840 s	21,542
Empty chunks	75,160 s	741
Speech	48,658 s	11,799
Overlap of speakers	4,916 s	8,253
Speech experimenter	3,414 s	1,091
Overlap with experimenter	182 s	734
Total nb of word tokens	229,704	
Total nb of word types	9,270	
Nb of hapax legomena	4,886	
Frequency of the word *ja* (yes)	6,185	

Table 1: *Factual data of* CORPUS ERNESTUS.

Section 5, future steps with adapting the corpus for ASR use and the preliminary results of the reduction analysis will be discussed. This paper closes with concluding remarks.

2. Material

The speech material used was the CORPUS ERNESTUS, a corpus of Dutch spontaneous dialogues [1], consisting of ten conversations of approximately 90 minutes each. The corpus contains 229,702 word tokens and 9,270 word types. An orthographic transcription was available in PRAAT LONG TextGrid format [9], where different tiers were used for the different speakers. The speech chunks are up to 15 seconds long, and 53 % of the chunks are shorter than 3 seconds, as is usual in the transcription of conversations, especially if these are made with a view to investigating discourse phenomena (cf. Table 1).

Characteristic for this corpus is the homogeneity in geographical and social background of the speakers. All 20 speakers were male native speakers of Dutch. They lived in the western provinces of the Netherlands and all had academic degrees. The speakers are between 21 and 55 years old. The set-up for achieving spontaneous dialogues was as follows:

- Pairs of colleagues or friends talked with each other, seated some 1.5 m from each other at a table in a soundproof room.
- Recordings were made with two Sennheiser MD527 supercardioid microphones on Sony DAT tapes.
- The speakers chose the topic for the first 40 minutes of the conversations freely.
- The second part was a role-play, where the speakers negotiated about the purchase of camping goods. Both speakers were told the goals they should reach before the role-play separately; no further specific instructions were given.
- The experimenter was only present during the first part, but hardly interfered in the conversations.

As the speakers experienced a relaxed atmosphere, i.e., speaking with a friend about everyday issues, the dialogues have a casual, chatty style. The spontaneous conversational speaking style is reflected not only by the high number of hesitations and broken words, but also by the high frequency of backchannel-like words such as *ja* (yes), *maar* (but), *nou* (now, well) and *nee* (no) with 6,185, 2,649, 1,617 and 1,242 occurrences, respectively. These four types already account for 5% of all word tokens. The relatively high proportion of word tokens that occur only once (53%) in Table 1 is due to the fact that all free conversations were about different topics.

Besides the characteristic word frequencies, also the big amount of overlapping speech is typical for spontaneous speech. Table 1 shows details about general speech duration and overlapping speech. Even though 70% of the chunks contain some overlapping speech, with the inevitable cross-talk in the recordings that comes with it, the effective overlap time in most chunks affects only a small proportion of the speech.

3. Method

To make the corpus accessible for phonetic research, a phonetic transcription is needed. Since human transcriptions are prohibitive, we decided to generate a broad phonemic transcription by means of a forced alignment using HTK [10]. The first step in this process is building a lexicon.

3.1. Building the lexicon

For building a lexicon the words in the orthographic transcriptions were looked up in the TST-lexicon. The TST-lexicon is a Dutch-language lexical database containing 361,163 word tokens. It was compiled by merging lexical resources such as CELEX, RBN and CGN [11]. Phonemic representations use 46 SAMPA phoneme symbols with no diacritics. Only 8,149 of the 9,270 word tokens could be found in the TST-lexicon. This high number of mismatches is partly caused by the use of different spelling conventions in the CORPUS ERNESTUS and also to spelling mistakes, which were subsequently corrected in the text grids. Furthermore, broken words were flagged in the original transcriptions. Another large proportion of the words not found in the TST-lexicon were proper names and foreign words like *Tatort, PHD-student, honeymoon, correctness, come-back* and *Bond-film*.

However, the most frequent category of missing words, with 650 word types, is formed by compounds. The creation of compounds is licensed by Dutch morphology and by the spelling conventions and novel compounds abound in spontaneous speech. A semi-automatic method was used to generate the missing canonical phonemic transcriptions. The compounds were split up into their parts by hand and looked up in the TST-lexicon. If they were found, the canonical phonemic transcriptions were concatenated. Subsequently, degemination rules were applied to the transcriptions of the compounds and stress-marks and syllable-boundaries were hand-checked. The following examples show some of these compounds and their canonical transcription:

antiquariaatcatalogus	An-ti-kwa-ri-'jat-kA-'ta-lo-GYs
bergsportvakantie	'bErx-'spOrt-va-'kAn-si
sinterklaassurprises	sIn-t@r-'kla-sYr-'pri-s@s

Finally, for the rest of the words, canonical transcriptions were handmade by the authors.

3.2. Preparing the corpus for automatic processing

Since the corpus was collected more than ten years ago for the purpose of studying reduction phenomena with an impressionistic method, automatic processing by ASR systems could not be anticipated. Therefore, the annotations in the text grids, which are the basis for all further processing, did not adhere to de facto standards for orthographic transcription that were developed in the CGN project. For example, hyphens were used within compounds, to separate the letters of words that a speaker was spelling and to mark parenthesis, in all cases with or without white spaces to separate hyphens from words. Another example

is the use of carriage returns within transcriptions that belonged to one single chunk, which interferes with automatic processing of the text grids in PRAAT [9]. The orthographic transcriptions have been adapted to the CGN standards, partly automatically, partly by hand.

Next, chunks with loud speaker and background noises need to be discarded for phonetic analysis. The same holds for chunks containing unintelligible speech for which no orthographic transcription can be made. Since transcribers were not supplied with explicit rules for transcribing these special cases, speaker noises and unintelligible parts were transcribed in different ways. Concerning background noise, overlapping speech is an important issue when dealing with a corpus of conversations. Even though the speech files were recorded on two separate channels, both speakers can still be heard on both channels. Therefore, for the experiments presented in this paper, chunks which contain overlap are not taken into account, as ASR systems do not deliver reliable results in these conditions. Finally, the wave files were re-sampled to the same sampling rate of the speech database that was used to train the acoustic models used.

3.3. Building a phonemic transcription

In a first experiment, a forced alignment was carried out with the HTK Speech Recognition Toolkit [10], using canonical phonemic transcriptions. For the forced alignment 37 32-Gaussian tri-state monophone acoustic models [12] were trained on the Dutch library of the blind of the CGN (Corpus Gesproken Nederlands) [11]. The goals of this exercise was to identify and discard chunks containing phenomena that prohibit automatic alignment.

For the second experiment, reduction rules were applied to the canonical representations of the words in order to create the pronunciation variants. The transcriptions of the TST-Lexicon already include frequently occurring connected speech processes, e.g. Sandhi rules [13], as well as frequent reductions. Word-final /n/ deletion after /@/ is applied to all words in the database; we restored these forms to their full version as starting point for applying the set of reductions rules. The rules applied in our first experiment are shown in Table 2.

The processes in lines 5-7 are well-studied for Dutch and they have been used before for the automatic generation of phonemic transcriptions [8]. Reduction rules 1-4 and 8-15 resulted from research on voice assimilation and segment reduction in casual Dutch [1]. Processes 1-4, the vowel and schwa reductions were applied to stressed syllables only. Due to the fact that canonical pronunciations of compounds in the TST-lexicon contain only one primary stress per word, compounds with a high number of syllables result in pronunciation variants which are very unlikely. To solve this problem, secondary stress was marked by hand in all compounds.

The reduction rules were then applied in 47 cycles on the full pronunciation form, where the output of one cycle was the input for the following. After every time a consonant reduction in word medial position (Table 2, lines 7-12) was applied, degemination was carried out before going on to further reduction rules. For every resulting pronunciation variant all rules that applied were documented. Finally, extremely reduced forms found in a previous study on reductions were added to the pronunciation dictionary [1]. After all, on average 2.91 pronunciation variants were generated in addition to the full canonical representations. The maximum number of pronunciation variants per word type was nine.

Then, forced recognition was carried out with this extended lexicon. Comparing the forced recognition results with the canonical representation provides a first glimpse of the reduction phenomena in the corpus.

Nb	Reduction Process	%
1	Transition from long to short vowels	1.4%
2	Transition vowels to schwa in unstressed syllables	4.4%
3	Short vowel deletion between voiceless obstruents	0.3%
4	Schwadeletion before liquid following obstruent	1.5%
5	Degemination	0.1%
6	/n/ deletion after schwa in wordfinal position	8.1%
7	/r/ deletion after schwa	4.6%
8	Deletion of bilabial plosives before /m/	0.1%
9	/n/ deletion after vowels before /s/	0.3%
10	/r/ deletion after vowels	9.4%
11	/t/ deletion: word final & in consonant clusters	6.7%
12	Consonant deletion in /n/ cluster	0.1%
13	Suffix -lijk reduced to /k/	0.6%
14	/h/-deletion in forms of hebben	0.8%
15	Transition from voiced to unvoiced consonants	3.4%

Table 2: *Reduction processes and their frequencies in relation to all word tokens.*

4. Results

In computing the results presented here 109,737 word tokens and 7,714 word types were used. Only half of the material can be used for straightforward automatic phonetic analysis, since the complete corpus is not yet available and chunks containing more than 10% overlap resulted in unreliable alignments. Furthermore, we excluded the high frequency word *ja* (yes). Since *ja* is never reduced, including it would bias the results of automatic processing the reduction analysis and compromise the comparability with corpora that do not comprise a large proportion of backchannel-like utterances.

Comparing the forced alignment with the pronunciation variants with the canonical transcriptions showed that 23.7% of all word tokens in the corpus are affected by at least one of the processes shown in Table 2. Considering the lexicon, 57.8% of the word types occur with one of the added pronunciation variants. This is quite a high number, because 52.7% of the word tokens are hapaxes.

7.1% of all word tokens suffered from syllable deletion; 97.2% of these were single syllable deletions and the maximum of 3 syllable deletions occurred in 0.1%. Rules that led to segment deletion are shown in Table 2 from lines 3 to 14, including vowel, schwa and consonant deletions. In total, 22.8% of all words in our data were reduced by at least one segment; 37.4% of these were single segment deletions, 0.2% were deletions of 9 to 14 segments.

Lines 1-4 in Table 2 show the applied vowel and schwa reduction rules. 7.6% of all words are affected by these reductions. Of the consonant reduction rules, /n/ and /t/ deletion are the most prominent, affecting 14.8% of all words.

The number of extremely reduced words is remarkable. The examples shown in Section 1, as extreme as they may seem, occurred in at least half of the cases when those words were used: *natuurlijk* with 54.2%, *eigenlijk* with 31.0% and *bijvoorbeeld* with 74.5%. A more in-depth analysis of these findings is left for future research.

5. Discussion

5.1. Pronunciation variants and reductions

Up to nine pronunciation variants were generated by applying reduction rules to canonical representations. Having a large speech corpus does not guarantee that we can estimate the relative frequency of the variants or the conditions in which specific variants tend to occur. On the one hand the word *hebben* (to have) consists of only 2 syllables, so that only three pronunciation variants were generated. It occurs 463 times in the analyzed material and the pronunciation variants occur with nearly equal frequency, while the canonical form occurs only in 3% of the tokens. This case would allow us to conclude on actually used pronunciation variants. On the other hand, for a long word like *wortelkanaalbehandeling* (apiceoctomy) a higher number of variants is generated, but as this word only occurs three times in the corpus, no conclusions can be drawn. Variants chosen by the alignment are marked with a star.

hebben	wortelkanaalbehandeling
'hE-b@n*	'wOr-t@l-ka-'nal-b@-'hAn-d@-lIN*
'hE-b@*	'wOr-t@l-kA-'nal-b@-'hAn-d@-lIN*
'hE-p@*	'wOr-t@l-kA-'nal-p@-'hAn-t@-lIN
'E-b@*	'wO-t@l-kA-'nal-p@-'hAn-t@-lIN
	'wO-t@l-k@-'nal-p@-'hAn-t@-l@N
	'wO-t@l-k@-'nal-p@-'hAn-t-l@N*

Although we are not yet in the position to draw conclusions about the relative frequency of reduction processes, the procedure for generating plausible reduced forms can already be improved. In the future we will use a tree-structured algorithm rather than the sequential one used for the experiments presented here. This will result in more, but also better pronunciation variants.

Despite the simplicity of our experiment, our results appear to confirm trends reported in the literature. Syllable and segment deletions in our data, 7.1% and 23.7% respectively, are very similar to the observations on a conversational corpus of American English by Johnson [3]. They are slightly higher than the results obtained by Van Bael, with 6.9% for syllable and 20.3% for segment deletions [8]. This may be due to the fact that Van Bael did not include the effects of vowel reduction in his study because he did not want to use information on word stress.

5.2. Future steps in the preparation of the corpus

To make the complete corpus accessible for research we will divide the present chunks into smaller ones of not more than three seconds. For this purpose, the outcome of the forced alignments will be used to set chunk boundaries such that overlapping speech and speaker noise will be separated from valuable speech material.

6. Conclusions

This paper presents a procedure to make already existing speech corpora, for whatever use they may have been collected, accessible for automatic phonetic research. The steps in preparing a corpus for automatic processing are explained and specific problems are illustrated. An automatic broad phonemic transcription was built using a forced alignment in two experiments. While in the first one, a lexicon of canonical phonemic representations of the words was used, the second experiment was carried out with a lexicon that had been enriched with pronunciation variants. These variants were generated by applying reduction rules to the canonical transcriptions of the words. The comparison of these two alignments allows us to deduce which reduction rules occurred with which frequency. Preliminary analysis of the results of a straightforward attempt to obtain automatic phonemic transcriptions show that the overall procedure is feasible. In future research we will extend the corpus, refine the transcriptions and analyze the results for improving our understanding of reduction processes.

7. Acknowledgements

This research was supported by the Marie Curie Project "Sound to Sense". Mirjam Ernestus was supported by a EURYI-award from the European Science Foundation and Odette Scharenborg by a Veni-grant from the Netherlands Organisation for Scientific Research (NWO). We would like to thank Eric Sanders for his support with Perl- and HTK- issues and for sharing his experience with large spoken language corpora.

8. References

[1] M. Ernestus, "Voice assimilation and segment reduction in casual dutch. a corpus-based study of the phonology-phonetics interface," Ph.D. dissertation, Vrije Universiteit te Amsterdam, Amsterdam, Juni 2000.

[2] M. Ernestus, R. H. Baayen, and R. Schreuder, "The recognition of reduced word forms." *Brain and Language*, vol. 81, pp. 162–173, 2002.

[3] K. Johnson, "Massive reduction in conversational American English," in *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, K. Yoneyama and K.Maekawa, Eds. Tokyo, Japan: The National International Institute for Japanese Language, 2004, pp. 29–54.

[4] E. Janse, S. G. Nooteboom, and H. Quené, "Coping with gradient forms of /t/-delition and lexical ambiguity in spoken word recognition," *Language and Cognitive Process*, vol. 22, no. 2, pp. 161–200, 2007.

[5] M. Adda-Decker, P. B. de Mareüil, G. Adda, and L. Lamel, "Investigating syllabic structures and their variation in spontaneous french," *Speech Communication*, vol. 46, pp. 119–139, 2005.

[6] K. J. Kohler, "The dissappearance of words in connected speech," *ZAS Working papers in Linguistics*, vol. 11, pp. 21–34, 1998.

[7] M. Saraçlar, H. J. Nock, and S. Khudanpur, "Pronunciation modelling by sharing gaussian densities across phonetic models," *Computer Speech and Language*, vol. 14, pp. 137–160, 2000.

[8] C. V. Bael, "Validation, automatic generation and use of broad phonetic transcriptions," Ph.D. dissertation, Radboud Universiteit Nijmegen, Nijmegen, October 2007.

[9] [Online]. Available: http://www.praat.org

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (version 3.2)," Cambridge University. Engineering Department., Tech. Rep., 2002.

[11] N. Oostdijk, W. Goedertier, F. V. Eynde, L. Boves, J. Martens, M.Moortgat, and H.Baayen, "Experiences from the spoken Dutch corpus project," in *Proceedings of LREC-2002*, vol. 1, 2002, pp. 340–347.

[12] A. Hämäläien, L. ten Bosch, and L. Boves, "Modelling pronunciation variation using multi-path hmms for syllables," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, 2007.

[13] W. Jongenburger and V. J. van Heuven, *Analysis and synthesis of speech: strategic research towards high-quality text-to-speech generation*. Mouton de Gruyter, 1993, ch. Sandhi Processes in natural and synthetic speech.