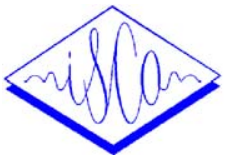
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Investigating Speech Style Specific Pronunciation Variation in Large Spoken Language Corpora

Christophe Van Bael, Henk van den Heuvel, Helmer Strik

Radboud University Nijmegen, the Netherlands
e-mail:{c.v.bael,h.v.d.heuvel,w.strik}@let.kun.nl

Abstract

In the past, linguistic research was typically conducted on relatively small datasets that were specifically designed for the research at hand. Whereas to date many large spoken language corpora have become available, the usefulness of these corpora is still not fully established in linguistic research. The research reported on in this paper was conducted to illustrate the potential of large multi-purpose spoken language corpora for linguistic research.

The possibility was investigated of identifying phonetic regularities in different speech styles. To this end, a data-driven study was conducted with a large multi-purpose spoken language corpus comprising a manually corrected broad phonetic transcription of the data. Our results show that speech style specific pronunciation processes can indeed be found in such a large corpus. This indicates that large multi-purpose spoken language corpora can contribute to linguistic research, if only for the purpose of hypothesis generation and verification.

1. Introduction

In the past, linguistic research was typically conducted on handcrafted data sets that were specifically designed for the research at hand. Such data sets can cover many details with regard to the topic under investigation, but most of these sets are small because they are very expensive to produce. Inevitably, the small size of the typical data set in linguistic research raises questions about the extent to which results can be generalised. Moreover, the manual annotation of data is prone to errors and inconsistencies, and specific data sets are often not suitable for use in other research.

From the eighties onwards, researchers have investigated the possibility of annotating large amounts of speech in a (semi-) automatic fashion. These annotations typically lack the degree of detail present in the annotations of handcrafted corpora. However, the automatic procedures with which these large corpora are generated guarantee them to be much cheaper, more consistent, and better suitable for a much larger variety of research than handcrafted corpora.

Because the use of large multi-purpose spoken language corpora is still not fully established in linguistic research, an attempt was made to investigate the potential of one such corpus for linguistic research, viz. the Spoken Dutch Corpus (Corpus Gesproken Nederlands - CGN [1]). To this end, a research task was defined in which pronunciation differences were studied between three different speech styles with varying degrees of spontaneity: read speech (RS), public lectures (PL), and telephone dialogues (TD).

A separate subcorpus was available for each speech style. Each of these subcorpora came with a manually corrected broad phonetic transcription of the data. In order to characterise pronunciation differences between the speech styles, the manual phonetic transcription was aligned with a canonical reference transcription of the data. In this way, context-sensitive re-write rules were obtained at the phone level. Subsequently, a measure was defined for the probability that rules apply given the presence of their phonetic context in the reference transcription. This measure was called the Rule Application Probability (RAP). By statistically comparing the RAPs of the most prominent rules in all speech styles, and by investigating the rules, interesting pronunciation differences between the three speech styles were discovered.

This paper is organised as follows. In section 2, the general method of the experiment is presented, as well as the material used in the experiment. In section 3, the results of the experiment are presented and discussed. Finally, in section 4, general conclusions are presented, as well as our plans for future research.

2. Method and material

2.1. Method

The programme ALIGN [2] was used to align the Manual Phonetic Transcription (MPT) with the canonical Reference Phonetic Transcription (RPT) of the data (see section 2.2). ALIGN is a dynamic programming algorithm that finds the optimal alignment of two strings of phonetic symbols on the basis of a feature matrix in which distances between phonetic symbols are defined.

Each time ALIGN yielded a mismatch between the RPT and the MPT, the relevant phones in the RPT and in the MPT were selected, as well as the two phonetic symbols to the left and the two phonetic symbols to the right of the phone in the RPT. In this way, observations of the context-sensitive optional rewrite rule [3]

$$X \rightarrow Y \,/\, L_2L_1 - R_1R_2 \qquad (1)$$

were generated. Subsequently, the RAP was computed for each rule in every speech style. The RAP of a rule was defined as follows:

$$RAP = (N_{rule} \,/\, N_{context}) \qquad (2)$$

where N_{rule} was the number of times a rule applied, and $N_{context}$ the number of times the context for the rule was encountered in the RPT. Only the rules of which the contexts occurred frequently in all speech styles were selected for further investigation. In doing so, we normalised for the differences in corpus size. For each corpus, the threshold for determining whether a context occurred frequently was dependent on the size of the corpus. Rules of which the contexts did not occur frequently in at least one of the three speech styles were disregarded. This selection procedure

10.21437/Interspeech.2004-675

enabled us to study the RAPs of the same rules in three different speech styles [4].

The resulting RAPs were submitted to an Analysis of Variance (ANOVA). We used a split-plot design in which the factors Rule type (substitutions, deletions, insertions) and Speech style (read speech, public lectures, telephone dialogues) were crossed. The block factor for the repeated measures was the rule, nested under Rule type, Speech style being the within-subjects factor. Huynh-Feldt sphericity corrections were applied to the degrees of freedom. Subsequent ANOVAs were carried out for the various rule types.

2.2. Material

The MPTs were taken from the Spoken Dutch Corpus (Corpus Gesproken Nederlands – CGN [1]). Statistics of the data are presented in Table 1.

Speech Style	# words	# phones
Read Speech (RS)	17,011	86,830
Public Lectures (PL)	3,473	17,037
Tel. Dialogues (TD)	8,558	35,027

Table 1: Number of words and phones in the transcriptions

The broad phonetic transcription delivered with the Spoken Dutch Corpus is a manually corrected version of an automatically generated canonical transcription. Transcribers got the instruction to modify the automatic phonetic transcription only if they were sure that their changes would yield a transcription that was significantly closer to the actual speech. Due to this procedure, a bias towards the original canonical transcription is to be expected in the MPT.

The RPTs were generated through a lexical lookup procedure with the orthographic transcriptions of the data and CELEX [5], a validated canonical lexicon comprising 381K lexemes and their Dutch pronunciation. All obligatory word-internal processes [6] were applied in CELEX.

3. Results and discussion

The selection of the rules of which the contexts occurred frequently in the three speech styles resulted in 154 rules. These rules were divided into a set of substitution rules, a set of deletion rules and a set of insertion rules.

Table 2 shows that the mean RAPs of the rules in the RS and the PL were similar, whereas phones were more frequently substituted and deleted in the TD than in the RS and the PL. Phones were most frequently inserted in the RS.

Rules	# rules	RS	PL	TD
Substitutions	N=81	.15	.12	.17
Deletions	N=46	.04	.06	.16
Insertions	N=27	.07	.05	.02
All	N=154	.10	.09	.14

Table 2: Mean RAPs of the three rule types (sub, del, ins)

The ANOVA of the data underlying Table 2 revealed that the factor Speech style was significant ($F(1.90, 287.47) = 12.58$, $p < .001$) together with its interaction with Rule type ($F(3.80, 287,47) = 10.49$, $p < .001$). This means that at least two speech styles were different with regard to the mean RAPs of the rules, but that the differences were dependent on the rule type. A pairwise comparison showed that the mean RAPs of the rules in the RS (mean RAP = .10) and in the PL (mean RAP = .09) were very similar, and that these RAPs were significantly lower ($p < .02$) than the mean RAP of the rules in the TD (mean RAP = .14). (For all pairwise comparisons Bonferroni adjustments were applied). Closer investigation of the rules revealed that it made sense to make further distinctions within the sets of substitution, deletion and insertion rules.

3.1. Substitution rules

Table 3 presents the mean RAPs of the different types of substitution rules in the three speech styles. Mean RAPs are given of rules in which consonants were substituted for other consonants, in which vowels were reduced to schwa, in which vowels were substituted for other vowels, and in which schwas were substituted for vowels.

Rules	# rules	RS	PL	TD
Consonant	N=22	.43	.37	.40
Vowel to schwa	N=22	.06	.06	.14
Vowel	N=22	.01	.01	.05
Schwa to vowel	N=15	.09	.04	.03
All	N=81	.15	.12	.17

Table 3: Mean RAPs of the substitution rules

An ANOVA of the data underlying Table 3 revealed that at least two speech styles differed significantly with respect to their mean RAPs of the substitution rules ($F(1.96, 156.43) = 4,61$), $p < .02$). A pairwise comparison showed that significantly more phones were substituted in the TD (mean RAP = .17) than in the PL (mean RAP = .12)($p < .05$).

The three speech styles were not significantly different with regard to the substitution of consonants. Nineteen consonant substitution rules described a process in which a /t/, /d/, /s/, /z/, /f/, /v/, or /p/ was voiced or devoiced: 15 times at a word boundary of a monosyllabic word, and 4 times in other contexts. The remaining 3 substitution rules defined substitutions of the /n/ in the monosyllabic indefinite article 'een' (/ən/, a(n)) with another nasal because of regressive assimilation of place.

Vowels were significantly more often reduced in the TD (mean RAP = .14) than in the RS (mean RAP = .06) and in the PL (mean RAP = .06)($p < .05$). Our set of 22 vowel reduction rules comprised 17 rules that occurred most frequently in the TD. All but one reduction rule defined a vowel reduction in a frequent monosyllabic word. Reduction rules were found for the vowels /ɪ/, /ɛ/, /ə/, /ɑ/, /a/, /o/, /ɔ/ and the diphthong /ɛi/, which is the only Dutch diphthong that can be substituted for /ə/ [6,7].

Twenty-two vowel substitutions for shorter, longer, or simply different vowels were found in the three speech styles. Twenty of these rules did occur in frequent mono-syllabic words. Significantly more vowels were substituted in the TD (mean RAP = .05) than in the other two speech styles (mean RAPs = .01)($p < .01$). Our set of 22 vowel substitution rules comprised 16 rules that were encountered most frequently in the TD. Moreover, only 2 of the 22 vowel substitution rules encountered in the TD occurred in the PL (of which 1 rule did

only occur in the PL), and only 6 of the 22 rules were present in the RS (of which 5 rules did occur more often in the RS than in any other speech style).

The RAPs of the schwa substitution rules were highest (though not significantly) in the RS. The set of 15 schwa substitution rules comprised 12 substitutions of /ə/ for /ɛ/ or /e/. All 15 rules occurred in realisations of the indefinite article 'een' (/ən/, a(n)) or in realisations of the definite articles 'de' (/də/, the) and 'het' (/ət/, the).

Our results suggest that vowel reductions and substitutions are more probable in more spontaneous speech (TD) than in less spontaneous speech (RS and PL). This phenomenon is usually attributed to a speaker's natural tendency to reduce articulatory effort in his or her speech [7]. Moreover, most substitutions occurred in frequent monosyllabic words. The fact that the schwa in the indefinite and the definite articles in Dutch is often expanded to a full vowel, implies that speakers also show a tendency to devote more articulatory effort to well-prepared speech (RS and PL), than to more spontaneous speech (TD).

3.2. Deletion rules

Table 4 shows the mean RAPs of the deletion rules in the three speech styles. The mean RAPs of the rules are presented in which consonants were deleted, in which vowels were deleted, and in which schwas were deleted.

Rules	# rules	RS	PL	TD
Consonant	N=23	.07	.08	.20
Vowel	N=13	.00	.02	.07
Schwa	N=10	.01	.07	.21
All	N=57	.04	.06	.16

Table 4: Mean RAPs of the deletion rules

An ANOVA of the data underlying Table 4 showed that the mean RAPs of at least two speech styles differed significantly ($F(1.34, 60.27) = 24.22$, $p < .001$). A pairwise comparison revealed that significantly more phones were deleted in the TD (mean RAP = .16) than in the RS (mean RAP = .04) and in the PL (mean RAP = .06)($p < .001$). The mean RAPs of the deletion rules in the RS and the PL were not significantly different. These results once more indicate that phones are more frequently deleted in more spontaneous than in less spontaneous speech.

A detailed study revealed that consonants were more frequently deleted in the TD (mean RAP = .20) than in the RS (mean RAP = .07, $p < .01$) and in the PL (mean RAP = .08, $p = .001$). Only 2 out of 23 consonant deletion rules occurred more frequently in another speech style than the TD. Moreover, only 5 of the 23 rules actually occurred in the PL, and only 11 out of 23 rules were encountered in the RS. Most of the deletions (12 out of 23) occurred in monosyllabic words. All deletions occurred in common Dutch words. Seven rules described the reduction of the /t/ and the /d/ in word-initial or word-final position of frequent monosyllabic words, whereas 6 rules described the reduction of /n/ at the end of the indefinite article 'een' (/ən/, a(n)) or the conjunction 'en' (/en/, and). Five more rules defined processes that were literally transcribed in the research reported in [7]: the deletion of /l/ in the word 'als' (/ɑls/, if), the deletion of /r/ in coda position after a schwa, the deletion of /r/ after a low vowel, and the deletion of /d/ after a nasal before a schwa.

Vowels were more frequently deleted in the TD (mean RAP = .07) than in the RS (mean RAP = .00, $p < .05$) and the PL (mean RAP = .02, $p < .001$). The TD showed the highest RAPs for 11 out of 13 vowel deletion rules, because only 2 out of 13 rules occurred in the PL, and 2 other rules in the RS. All but one rule occurred in a monosyllabic word. All vowel deletions occurred in frequent words.

Likewise, schwas were more often deleted in the TD than in the other two speech styles. However, a pairwise comparison showed that the mean RAPs of the schwa deletion rules in the TD were not significantly higher than the average RAPs of the schwa deletion rules in the PL and in the RS. Five out of ten rules described the deletion of schwa in the definite article 'het' (/ət/, it), one rule described the deletion of schwa in the indefinite article 'een' (/ən/, a(n)), two rules described the deletion of schwa between an obstruent and a liquid. These are all plausible deletion processes in Dutch.

It can be concluded that far more phones were deleted in the most spontaneous speech style (TD). There seems to be a clear relationship between the degree of spontaneity of speech and the frequency of phone deletions. This tendency, as well as the majority of the deletion rules, is in line with findings reported in the linguistic literature [6,7].

3.3. Insertion rules

We encountered 27 different phone insertions in the data. However, 20 of these processes were probably due to the canonical transcriptions in the lexicon with which the RPT was generated. There are several moot points in the literature concerning the underlying representation of words. One such moot point is the underlying representation of words ending in an alveolar nasal after a schwa at the end of a morpheme that is not a verbal stem. In CELEX, the final /n/ was never transcribed. This means that we found /n/-insertion rules in our data. However, if the final /n/ had been transcribed in CELEX (and hence also in the RPT), many phone deletions would have occurred, especially in the more spontaneous speech. As we used CELEX to generate the RPT, 20 insertion rules reported on in this section can be equally interpreted as insertion processes or as the negation of optional deletion rules.

Table 5 presents the mean RAPs of the insertion rules in the three speech styles. The mean RAPs of the rules are presented in which an alveolar nasal was inserted in syllable-final position after a schwa at the end of a morpheme that is not a verbal stem. The remainder of the insertion rules were of a random nature. They could not be further divided into clear sets of rule types.

Rules	# rules	RS	PL	TD
Word-final /n/	N=11	.06	.03	.00
Remainder	N=16	.07	.06	.03
All	N=27	.07	.05	.02

Table 5: Mean RAPs of the insertion rules

The insertion rules showed opposite tendencies with regard to the majority of the substitution and the deletion rules. Whereas more phones were substituted and deleted in the more spontaneous speech style (TD), more phones were

inserted in the less spontaneous speech styles (RS and PL). An ANOVA of the data underlying Table 5 revealed that at least two speech styles differed with regard to the mean RAPs of their insertion rules ($F(1.65, 42.92) = 3.65$, $p < .05$). A pairwise comparison confirmed that significantly more insertions occurred in the RS (mean RAP = .07) than in the TD (mean RAP = .02)($p < .05$).

One of the most prominent phonological rules in Dutch is the rule of /n/-deletion in syllable-final positions after a schwa at the end of a morpheme that is not a verbal stem [6]. As already indicated, in CELEX, no occurrences of the /n/ in this position were transcribed, because they were not considered to be part of the underlying phonetic transcription. Therefore, no deletions of /n/ could be found in this position. However, we did find some interesting /n/-insertions in our data, suggesting that not all speech styles expose the same degree of /n/-deletion. Only 3 out of 11 /n/-insertion rules that occurred in the RS also occurred in the PL, and none of them occurred in the TD. It is therefore not surprising that the RS (mean RAP = .06) and the TD (mean RAP = .00) differed significantly with regard to the RAPs of the /n/-insertions ($p < .01$). There was no significant difference between the RAPs of the /n/-insertion rules in the RS (mean RAP = .06) and in the PL (mean RAP = .03). Five remaining insertion rules can be attributed to the fact that the canonical transcription of the definite article 'het' (it) was /ət/. Five insertion rules define the insertion of /h/ in word-initial position of the definite article. If the canonical transcription of the definite article had been /hət/, many deletion rules would have showed up in the more spontaneous speech. Five more insertion rules can be attributed to the canonical transcriptions in CELEX. Each of these 5 rules define phonetic processes reported on in [7]. One 'true' insertion rule was often encountered in all data sets: the insertion of schwa between the /l/ and the /f/ in the word 'elf' (/ɛlf/, eleven).

It can be concluded that phone insertions were most common in well-articulated speech (RS and, to a lesser extent, PL). However, 20 out of 27 rules can also be interpreted as negations of optional deletion rules frequently occurring in more spontaneous speech.

4. Conclusions and future research

Because the use of large multi-purpose spoken language corpora is still not fully established in linguistic research, we tried to illustrate the possible benefit of using such a corpus for linguistic research. In order to put our research into practice, a study was conducted in which we tried to define pronunciation characteristics in three different speech styles: read speech, public lectures, and telephone dialogues.

Our data-driven research provided interesting insights into pronunciation variation in Dutch data of different speech styles. In particular, our results show that vowels are more likely to be reduced to schwa or substituted for another vowel in more spontaneous speech than in less spontaneous speech. Phone deletions seem to be more common in spontaneous speech as well. Moreover, our results show that frequent monosyllabic words tend to be very vulnerable to substitution and deletion processes. This may be due to the fact that reduced pronunciations of these words are stored in our mental lexicon, as was suggested in [7]. The vast majority of our results closely resemble findings reported in the literature on connected speech processes in Dutch [6,7].

In our study, no significant differences could be found between the mean RAPs of the RS and the PL. This is not surprising, as both speech styles comprise prepared and well-articulated speech. This explains the significant differences found with the speech in the telephone dialogues, comprising more spontaneous and sloppy speech.

Our results support our belief that linguistic research can benefit from the use of large multi-purpose spoken language corpora, if only for the purpose of hypothesis generation and verification. Next, we will expand our research to automatic phonetic transcriptions. Firstly, we will try to automatically identify speech styles, based on knowledge gathered from this research. Secondly, we will generate automatic phonetic transcriptions for the same data material and compare the resulting rules and RAPs with the rules and RAPs of this research in terms of speech style differences and in terms of their usefulness for automatic speech style classification.

5. Acknowledgements

The work of Christophe Van Bael was funded by the "Stichting Spraaktechnologie" (Foundation for Speech Technology), Utrecht, the Netherlands.

6. References

[1] Oostdijk, N. "The Spoken Dutch Corpus: Overview and first evaluation", *Proceedings of LREC*, 887-893, 2000.

[2] Cucchiarini, C., "Assessing Transcription Agreement: Methodological Aspects", *Clinical Linguistics & Phonetics, Vol. 10(2), 1999, pp. 131-155.*

[3] Labov, W. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, 1972.

[4] Van Bael, C., Strik H. & van den Heuvel, H., "On the Usefulness of Large Spoken Language Corpora for Linguistic Research", *Proceedings of LREC*, to appear, 2004.

[5] Baayen, R.H., Piepenbrock, R., & Gulikers, L., "The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor], 1995.

[6] Booij, G., *The Phonology of Dutch*, Oxford University Press, New York, 1999.

[7] Ernestus, M., *Voice Assimilation and Segment Reduction in Casual Dutch. A Corpus-Based Study of the Phonology-Phonetics Interface*. Ph.D. thesis, University of Amsterdam, the Netherlands, 2000.