Regional Bias in the Broad Phonetic Transcriptions of the Spoken Dutch Corpus

Evie Coussé & Steven Gillis

Centre of Dutch Language and Speech (CNTS), University of Antwerp
Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium
evie.cousse@ugent.be steven.gillis@ua.ac.be

Abstract

In this paper, we assess an aspect of the quality of the broad phonetic transcriptions in the Spoken Dutch Corpus (CGN). The corpus contains speech from native speakers of Dutch originating from The Netherlands and the Dutch speaking part of Belgium. The phonetic transcriptions were made by transcribers from both regions. In previous research, we have identified regional differences in the transcribers' behaviour. In this paper, we explore the precise sources of the regional bias in the CGN transcriptions and we evaluate its impact on the phonetic transcriptions. More specifically, (1) the regional bias in the canonical transcriptions that served as the basis for the verification task of the transcribers is critically analysed, and (2) we verify in an experiment the regional bias introduced by the transcribers themselves. The possible effects of this inherent regional bias in the CGN transcriptions on subsequent linguistic analyses are briefly discussed.

1. Introduction

The recently completed Spoken Dutch Corpus (*Corpus Gesproken Nederlands*, henceforth CGN)[1] offers a wealth of empirical data for linguistic research of Standard Dutch as currently spoken in The Netherlands (NL) and the northern part of Belgium (B). The CGN consists of nine million words which are all transcribed orthographically and are enriched with part-of-speech information. In addition, a selection of one million words is supplied with a syntactic annotation and a broad phonetic transcription. The phonetic transcription forms the empirical basis of our ongoing research into pronunciation variation in Standard Dutch (Swerts et al., 2003).

In the course of our study, we were faced with inevitable methodological questions about the reliability and consistency of the broad phonetic transcription. Within the CGN project, several small scope evaluations were made for transcriptions of speech fragments originating from The Netherlands (Binnenpoorte et al., 2003; Goddijn & Binnenpoorte, 2003). However, as we are interested in the regional variation of spoken Standard Dutch, we also need information about the reliability and consistency of transcriptions originating from Belgium. It is well possible that the available evaluations of NL transcriptions do not hold for B speech fragments as the transcription procedure applied in both regions deviates to some extent.

The transcription procedure for the broad phonetic transcription in the CGN is semi-automatic. The basis of the phonetic transcription is a orthographic transcription of the speech fragments. Through a simple look-up procedure every word form in the orthographic transcription was replaced by its corresponding canonical phonetic transcription of either the pronunciation lexicon Celex (Baayen et al., 1995) or Fonilex (Mertens & Vercammen, 1998). In order to anticipate pronunciation differences in The Netherlands and Belgium, NL speech fragments rely systematically on the Netherlandic Dutch pronunciation of Celex, whereas B speech fragments are made on the basis of the Belgian pronunciation in Fonilex. These canonical transcriptions form the starting point for the human transcribers. Transcribers were instructed to listen to the speech signal and correct the canonical transcription in accordance to what they heard. Since the CGN project was organised nationally, transcribers only corrected canonical transcriptions of speech fragments originating from their own region. Thus, NL fragments were corrected by NL human transcribers whereas B fragments were corrected by B transcribers.

In this paper, we investigate to what extent the deviant canonical transcription and the different regional background of the transcribers induces a regional bias in the phonetic transcriptions. Previous research (Coussé et al., 2004) has shown that the regional background of transcribers exerts a considerable impact on the labelling of vowel quality in spoken language resources: labellers from The Netherlands and Belgium tend to categorise vowels differently in the experiment, though they speak the same standard language. We expect that the phonetic transcriptions of the CGN suffer from a similar regional bias. The presence of such a regional bias has important repercussions for investigations of Standard Dutch on the basis of the CGN phonetic transcriptions, and a fortiori on studies of regional pronunciation variation. Moreover, this study is of interest for any researcher working with large speech corpora: when analyzing such corpora the exact format of the transcription procedure as well as the (regional) background of transcribers should be closely scrutinized.

2. Experimental design

To assess the regional bias on the broad phonetic transcriptions of the Spoken Dutch Corpus, we set up a transcription experiment. We selected a sample of 18 speech fragments from the CGN of about one minute each, taken from interviews with teachers of Dutch who originate from different regions in The Netherlands and Belgium. The varying regional background of the speakers in the sample reflects the regional variation present in the speech fragments of the CGN.

[1] More information about the CGN can be consulted online on http://lands.let.kun.nl/cgn/ehome.htm & http://www.tst.inl.nl/cgn.htm.

For this speech sample, two canonical transcriptions were generated automatically: one transcription reflecting the NL pronunciation of Celex, the other showing the B pronunciation found in Fonilex. Consequently, differences between canonical transcriptions were no longer exclusively linked to a particular regional background of the speaker. This disconnection enables us to detect regional differences between the two experimental canonical transcriptions based on identical speech fragments.

Six experienced transcribers of the Spoken Dutch Corpus were recruited from the original pool of CGN transcribers. They were instructed to listen closely to the speech sample and to correct the canonical transcription if necessary. To assess the impact of the regional background of the transcribers on their transcription task, we engaged three native speakers of Dutch from The Netherlands and three from the Dutch speaking part of Belgium. The NL transcribers had to verify the NL canonical transcription and the B transcribers the B canonical transcription.

To compare the transcriptions of all transcribers, the symbols needed to be aligned. This process was automated by means of a script using a minimal edit distance algorithm that calculates the appropriate cost for every substitution, deletion and insertion between two transcriptions on the basis of an articulatory feature matrix for vowels and consonants.

3. Results

In this section, the results of our experiment are reported. First, we investigate the differences between the six experimental transcriptions as an exploratory evaluation (3.1). Then, we analyse the impact of the regional bias on the canonical transcriptions (3.2) and the different regional background of the transcribers (3.3) on the phonetic transcription separately.

3.1. Variation in phonetic transcriptions

In this section, we explore the variation and consistency in the transcriptions of our experiment. We have aligned the transcriptions in pairs so that we can compare each symbol and determine the exact amount of agreement between all transcriptions. Table 1 shows the percentages of identical symbols between the six transcriptions.

	NL1	NL2	NL3	B1	B2	B3
NL1	-	-	-	-	-	-
NL2	87.1	-	-	-	-	-
NL3	85.6	88.9	-	-	-	-
B1	84.3	86.4	87.3	-	-	-
B2	86.0	88.1	88.3	90.9	-	-
B3	83.2	85.8	85.5	88.2	89.7	-

Table 1: Inter-transcription agreement (in %, n = 10,696)

Table 1 shows an average agreement of 85.6% (SD = 2.1, Median = 87.1%). The percentage agreement is very high, thus pointing at a high degree of consistency for all experimental transcriptions. These results can be compared with previous evaluations of the Spoken Dutch Corpus. Goddijn & Binnenpoorte (2003) report on the consistency of transcriptions made by four transcribers from The Netherlands of 16 minutes of CGN speech fragments produced by speakers from The Netherlands. In their experiment, similar high results (between 87.7% and 91.7% agreement) have been attained for transcriptions of interviews, the speech style that most resembles our sample of speech fragments.

The percentages in Table 1 appear to diverge according to the origin of the transcribers. The highest agreement is reached among B transcribers and varies between 88.2% and 90.9%, on average 89.6%. The other pairs of transcriptions attain a markedly lower agreement: NL transcribers reach an average agreement of 87.2% and the average agreement percentage between NL and B transcribers is 86.1% (range: 83.2 – 88.1). The differences between the B-B percentage agreements, the NL-NL agreements and the B-NL agreements are almost significant, as assessed by a Kruskal-Wallis test ($\chi^2 = 5.9$ $p = 0.053$). At the moment, we can only speculate about the precise reasons for these differences. The variation might be attributed to differences in the experimental canonical transcriptions or to different transcription tendencies of the transcribers. In the next sections, we will analyse these factors separately.

3.2. Regional bias in canonical transcriptions

One regional bias on the phonetic transcription of the Spoken Dutch Corpus might be the use of deviant sources for the canonical transcription of speech fragments originating from The Netherlands and Belgium. In this section, we will identify the sources of the differences between the two versions of the canonical transcription (3.2.1) and assess the impact on these canonical transcriptions (3.2.2).

3.2.1. Sources of regional bias

As mentioned in the introduction, one of the major sources of variation between the canonical transcriptions is the use of the pronunciation lexica Celex for NL speech fragments and Fonilex for B speech fragments. Hoste et al. (2004) provide a survey of the major systematic differences between the two pronunciation lexica. The differences mainly involve tendencies such as the (de)voicing of consonants and the confusion of tense and lax vowels. These pronunciation differences appear to coincide largely with the tendencies studied and described in linguistic comparative research of inter-regional pronunciation variation in standard Dutch (Booij, 1995). Another difference between in the canonical transcription of NL and B speech fragments involves the treatment of words in the speech fragments that were absent in the pronunciation lexica. For these out-of-vocabulary words, a different grapheme-to-phoneme conversion procedure was applied for both regions. For the B transcriptions, a memory-based learning algorithm was used to train a grapheme-to-phoneme converter with Fonilex as a training corpus (Hoste et al., 2000). Out-of-vocabulary items in NL transcriptions were transcribed by means of the rule-based grapheme-to-phoneme converter FONPARS (Kerkhoff & Rietveld, 1994). A final source of difference between the NL and B transcriptions is the implementation of assimilation and degemination rules in the canonical transcription for NL speech fragments. These (optional) word-external rules were not applied in B transcriptions.

3.2.2. Impact of regional bias

The use of different sources for the canonical transcription of speech fragments from The Netherlands and Belgium leaves its traces in the canonical transcription. We can track the differences in the canonical transcriptions by comparing the NL canonical transcription with the B canonical transcription, both made for the speech sample used in the experiment.

It appears that up to 7.3% of the symbols are not identical in both experimental canonical transcriptions. First of all, both transcriptions do not contain the same number of segments: 0.7% of the segments in the NL canonical transcription have no match in the B transcription whereas the reverse is true for 2.3% of the segments in the B transcription. This proportion implies that the B canonical transcription contains more segments than the NL transcription. Especially consonants which are present in the B transcription are absent in the NL transcription (2.2% of all symbols). Typically, the word-final /n/ (following a schwa) present in the B transcription is absent in the NL transcription, as is illustrated in examples (1) and (2).

(1) B komə**n** vs. NL komə 'to come'
(2) B bəvelə**n** vs. NL bəvelə 'to order'

The difference can be traced back to a different transcription option in the pronunciation lexica Celex and Fonilex. Fonilex represents the unreduced /n/ throughout the lexicon whereas Celex applies the deletion of /n/ consistently. Note that in Dutch the realization of /n/ after schwa is considered to be an optional phonological process (Booij, 1995). Apart from insertions and deletions in both experimental canonical transcriptions, the largest difference between the transcriptions can be attributed to substitution (4.4%). A frequent alteration between the two transcriptions is the substitution of voiced fricatives in the B transcription by voiceless fricatives in the NL transcription (1.8%). Typically, a word-initial voiced /ɣ/ present in Fonilex will have the voiceless counterpart /χ/ in Celex as can be observed in example (3) and (4).

(3) B **ɣ**ut vs. NL **χ**ut 'good'
(4) B **ɣ**raχ vs. NL **χ**raχ 'gladly'

Plosives on the other hand have more often a voiced quality in the NL transcription whereas they have a voiceless counterpart in the B transcription (0.5%). Unlike the above examples, these differences cannot be attributed to differences in the pronunciation lexica but to the implementation of an assimilation rule in the NL transcription. In examples (5) and (6), the regressive assimilation of the voiceless plosives /k/ and /t/ with the voiced plosives /b/ and /d/ is symbolized by '←'.

(5) B ɪ**k** bɛn vs. NL ɪ**g** ←bɛn 'I am'
(6) B mɛ**t** dɑ**t** dul
NL mɛ**d** ←dɑ**d** ←dul 'with that purpose'

We only mentioned some examples of differences to illustrate sources of variation in the canonical transcription. It is often hard to identify and quantify the influence of the different sources in the canonical transcription since they can interact. Several instances of variation that were discussed in isolation in the previous examples are brought together in example (7).

(7) B **ɣ**raχ bəvelə**n** **ɣ**eft 'gladly orders gives'
NL χraɣ ←bəvelə **χ**eft

We have demonstrated the differences between NL and B canonical transcriptions in our experiment. The divergence between both transcriptions can have a considerable impact on the ultimate phonetic transcription. Although transcribers are asked to correct the canonical transcription in accordance with the audio signal, we may expect some bias in the canonical transcription. First, human transcribers tend to suffer from loss of concentration and fatigue and thus run the risk of overlooking some symbols in the canonical transcription that do not correspond to the audio signal. Furthermore, the transcription protocol stipulates that in case of doubt the canonical transcription should be left unchanged. If these cases coincide with the deviant symbols in the NL and B canonical transcription, the verified phonetic transcription suffers from a regional bias of the canonical transcription.

3.3. Regional bias of human transcribers

In addition to the different sources for the canonical transcription, the regional background of the human transcribers can affect the phonetic transcription. In this section, we investigate whether NL and B transcribers correct the canonical transcription in a different way. To prevent interference of the differences between the two canonical transcriptions with the regional background of the transcribers, we only investigate identical symbols in both canonical transcriptions.

To asses the amount of corrections in the canonical transcription by the transcribers, we aligned the canonical transcription with the six transcriptions produced by our transcribers. All instances where the verified transcription deviates from the canonical transcription were identified. In Table 2, percentages of deletions, substitutions and insertions between the (reduced) canonical transcription and the six verified transcriptions are displayed.

	NL1	NL2	NL3	B1	B2	B3
Deletion	6.6	4.6	3.3	2.3	2.5	3.7
Substitution	8.7	6.4	6.5	5.1	4.5	5.9
Insertion	0.0	0.0	0.0	0.0	0.0	0.0
Corrections	15.3	11.0	9.8	7.5	7.1	9.7

Table 2: Disagreement between the canonical and the 6 verified transcriptions (in %, n = 10,085)

The percentages in Table 2 show rather few differences between the (reduced) canonical transcription and the experimental transcriptions. The transcribers have altered between 7.1% and 15.3% of the symbols in the canonical transcription. The NL transcribers appear to differ most from the canonical transcription and attain disagreement percentages between 9.8% and 15.3%, average 12.0%. Their B colleagues are more loyal to the canonical transcription and only change between 7.1% and 9.7% of the symbols, average 8.1%. The difference between the number of corrections made by the NL and the B

transcribers is quite remarkable. Perhaps the dichotomy indicates a different approach of the transcription task by NL and B transcribers. B transcribers may tend to attach greater value to the canonical transcription thus reassuring transcription consistency, whereas NL transcribers may have concentrated more on the detailed phonetic transcription of the speech signal. The effect of greater consistency among B transcribers can also be observed in Table 1.

In order to investigate whether NL and B transcribers not only change a different proportion of the canonical transcription but also use a different correction strategy, we have subdivided the corrections in the classes *deletion*, *insertion* and *substitution* (Table 2). Substitution is the most frequent change (between 4.5% and 8.7%) for all transcribers. Somewhat less frequent is the deletion of symbols in the canonical transcription (between 2.3% and 6.6%). We have not attested one single example of transcribers inserting a symbol in the canonical transcription. The use of deletion and substitution seems to be distributed proportionally among the six transcribers. This indicates that transcribers notice the same kind of pronunciation variation in the speech signal but vary to the extent they actually correct the canonical transcription.

The deviant transcription tendencies of the NL and B transcribers have a considerable impact on the phonetic transcriptions of the CGN. In the CGN project the verification of canonical transcriptions was organised 'nationally', i.e. B transcribers verified B canonical transcriptions and NL transcribers corrected NL canonical transcriptions. Hence, as our results indicate, NL phonetic transcriptions show more corrections and thus more variation in pronunciation than B transcriptions. This tendency in the transcriptions may give the non-validated impression NL speakers vary more in their pronunciation of Standard Dutch than B speakers do.

4. Conclusion

In this paper, we evaluated the broad phonetic transcriptions of the Spoken Dutch Corpus. We conducted an experiment with six transcribers on 18 speech fragments under different regional settings in order to unravel the influence of regional bias of the canonical transcription and the regional background of the transcribers.

It appeared that the canonical transcription reflecting the NL pronunciation diverges considerably from the B canonical transcription. The major sources of these differences turned out to be the use of different pronunciation lexica, a different grapheme-to-phoneme conversion of out-of-vocabulary words and the implementation of word-external variation rules. These deviant sources induce a difference of 7.3% of the symbols in both canonical transcriptions.

Furthermore, the regional background of our transcribers had an impact on the transcription task. It appeared that transcribers from The Netherlands tend to correct more symbols in the canonical transcription than Belgian transcribers (12.0% vs. 8.0%). Although the transcribers from both regions vary to the extent they change the canonical transcription, they tend to change the transcription in the same way.

In sum, we have demonstrated experimentally that the phonetic transcriptions of the Spoken Dutch Corpus suffer to some extent from regional bias. These results have repercussions for linguistic research into the regional variation of Standard Dutch pronunciation based on the phonetic transcriptions of the CGN. There will always be the risk of interference of the regional bias in the transcription with the regional variation present in the speech material itself. First, some regional differences in the NL and B canonical transcriptions that are not supported by the speech signal may remain unchanged in the verified phonetic transcriptions. Furthermore, transcriptions made by NL transcribers show more corrections than transcriptions of B transcribers. As in the Spoken Dutch Corpus transcribers only transcribe native speech fragments, the greater variation in the NL phonetic transcription can give the false impression that NL speakers show more variation in their pronunciation than B speakers. Although we cannot remedy the interference of regional bias in the phonetic transcriptions, at least an explicit notice of the bias should be integrated in any discussion of regional variation in the phonetic transcription of the CGN.

Acknowledgements

This research was funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) within the project Flexible Large Vocabulary Recognition (FLaVoR).

References

Baayen, R., R. Piepenbrock & L. Gulikers (1995). *The CELEX lexical database.* Release 2. Philadelphia: University of Pennsylvania.

Binnenpoorte, D., S. Goddijn & C. Cucchiarini (2003). How to improve human and machine transcriptions of spontaneous speech. In *Proceedings ISCA & IEEE Workshop on SSPR*, pp. 1361-1364.

Booij, G. E. (1995). *The phonology of Dutch.* Oxford: Oxford University Press.

Coussé, E., S. Gillis, H. Kloots & M. Swerts (2004). The influence of the labellers' regional background on phonetic transcriptions: Implications for the evaluation of spoken language resources. In *Proceedings LREC*, pp. 1447-1450.

Goddijn, S. & D. Binnenpoorte (2003). Assessing manually corrected broad phonetic transcriptions in the Spoken Dutch Corpus. In *Proceedings 1st ICPhS*, pp. 147-150.

Hoste, V., S. Gillis & W. Daelemans (2000). A rule induction approach to modeling regional pronunciation variation. In *Proceedings ICML-2000*, pp. 327-333.

Hoste, V., W. Daelemans & S. Gillis (2004). Using rule-induction techniques to model pronunciation variation in Dutch. In: *Computer Speech and Language* 18, pp. 1-23.

Kerkhoff, J. & T. Rietveld (1994). Prosody in NIROS with FONPARS and ALFEIOS. In P. De Haan & N. Oostdijk (eds.) *Proceedings Department of Language and Speech, University of Nijmegen,* 18 pp. 107-119.

Mertens, P. & F. Vercammen (1998). *Fonilex Manual. Fonilex: a pronunciation database of Dutch in Flanders.* http://bach.arts.kuleuven.ac.be/fonilex.

Swerts, M., H. Kloots, S. Gillis & G. De Schutter (2003). Vowel reduction in spontaneous spoken Dutch. In *Proceedings ISCA & IEEE Workshop on SSPR*, pp. 31-34.