The Influence of the Labeller's Regional Background on Phonetic Transcriptions: Implications for the Evaluation of Spoken Language Resources

Evie Coussé, Steven Gillis, Hanne Kloots, Marc Swerts

Centre of Dutch Language and Speech, University of Antwerp
Campus Drie Eiken, Universiteitsplein 1, B-2610 Antwerpen, Belgium
{evie.cousse, steven.gillis, hanne.kloots}@ua.ac.be
m.g.j.swerts@uvt.nl

Abstract

Phonetic transcriptions of spoken language corpora are not an exact written reproduction of the speech signal. They are influenced by a variety of factors such as the transcriber's native categorical perception. What remains unexplored is to what extent variation of perception *within* the same language exerts any influence on phonetic transcriptions. We report a case study of the labelling of vowel quality performed by native speakers of Dutch from either The Netherlands or Belgium. An analysis of the distribution of vowel quality labels reveals that labellers from The Netherlands have an other preference for certain vowel quality labels than labellers from Belgium. The inter-labeller agreement between is higher between labellers from the same region than between labellers from different regions. From these results, the conclusion can be drawn that labellers from The Netherlands and Belgium have a different perception of vowel quality in Standard Dutch. Thus, the factor regional background of transcribers should be taken into account when evaluating phonetic transcriptions of spoken language resources.

1. Introduction

There is growing awareness that the transcriptions and annotations of spoken language resources need to be evaluated. To assess the reliability and consistency of a transcription it is common to make some form of a reference transcription based on the transcription of more than one transcriber. This reference transcription is considered the 'best' transcription against which individual transcriptions can be evaluated (e.g. Shriberg et al, 1984). Yet, all human transcriptions – even reference transcriptions – are not an exact written reproduction of the speech signal since they are influenced by a variety of factors. Human transcribers are known to be affected by speech variables such as speech style and the length of an utterance but also by the amount of training they have had and their mother tongue (Cucchiarini, 1993).

In this paper, we focus on the potential impact of the transcriber's language background on phonetic transcriptions. It is often claimed in the literature that the perception of speech stimuli is determined by the category boundaries of the transcriber's sound system (e.g. Delattre et al, 1952). A straightforward example from cross-linguistic research is the problems native speakers of Japanese encounter learning the /r/-/l/ contrast of for instance English, which has no phonemic status in the Japanese sound system (Strange & Jenkins, 1978). The impact of categorical perception can also be observed in comparing more closely related languages. Speakers of English, German and Dutch for example tend to calibrate the voice-onset continuum differently in order to distinguish voiceless from voiced consonants (Abramson & Lisker, 1967). These cross-linguistic examples indicate that the categorical perception of a transcriber is inevitably filtered through his or her native sound system. Thus, transcribers with a different mother tongue produce deviant transcriptions of the same utterance. What remains unexplored, however, is to what extent variation of perception *within* the same language exerts any influence on phonetic transcriptions. To our knowledge, only compilers of dialect atlases have signalled the possible impact of the regional background of transcribers on the reliability of transcriptions (Jaberg & Jud, 1927; Hotzenköcherle, 1962; Ringaard, 1964; Pée, 1971).

To assess the influence of the labeller's regional background on phonetic transcriptions, we explore as a case study the labelling of unstressed vowels in a subset of iambic words. According to Booij (1995) and Ernestus (2000), these vowels are highly sensitive to *vowel reduction.* In this paper, the process of vowel reduction comprises the shortening of phonologically long vowels, the reduction of a vowel to schwa and the complete deletion of a vowel. Previous research (Kloots et al, 2003) has shown that the preference for certain types of vowel reduction varies in the Dutch spoken in The Netherlands and Belgium. Although speakers from both countries share the same standard language, the national border between The Netherlands and Belgium inevitably acts as a linguistic border (De Schutter, 1994). In this paper, we focus on potential dissimilarities in perception of vowel quality between labellers from The Netherlands and Belgium.

2. Method

Six native speakers of Dutch listened independently to a small subset of the Spoken Dutch Corpus[1]. The subjects were trained transcribers with a linguistic background and originate from the cross-border dialect region 'Brabant' in The Netherlands (NL) and the Dutch speaking part of Belgium (B). Three Belgian and three Dutch labellers participated in the experiment. The set of stimuli consisted of 894 instances of the iambic words *moment* (moment), *probeer(t)* (to try), *manier* (manner) and *docent* (teacher), taken from the component 'spontaneous speech' of the Spoken Dutch Corpus and produced by teachers of Dutch. The labellers had to assign a vowel quality label to each target vowel: 'long', 'short', 'schwa', 'zero', their

[1] More information on http://lands.let.kun.nl/cgn/ehome.htm

intermediate values as well as the label 'unintelligible'. The transcription task was monitored by the internet application WWStim developed by Theo Veenker (Utrecht University)[2]. The stimuli were presented via the audio channel of a computer and the subjects had to click on one of the labels shown on the screen, after which the responses were further processed automatically.

3. Results

In order to assess the effect of the labeller's regional background on phonetic labelling, two research questions are explored. (1) Does the appreciation of vowel quality differ between the labellers from The Netherlands and Belgium? (2) What is the degree of agreement between the labellers from the same region as opposed to labellers from different regions?

3.1 Vowel quality labelling

In this first section, we examine which labels the transcribers assigned to the selected target vowels. In particular, we are interested in the frequencies of the labels used by the NL and B scores in order to reveal their preferences. Preferences in labelling can be attributed to two factors: (1) either the preference reflects an inherent quality of the stimuli, (2) or the preference reveals an inherent property of the labellers, i.e. perception. In table 1, the relative frequencies of the vowel quality labels are given for NL and B labellers separately and for all labellers together.

	NL labellers (n = 2682)	B labellers (n = 2682)	All labellers (n = 5364)
Long	34.8	6.8	20.8
Long/short	5.8	12.5	9.2
Short	30.5	45.8	38.2
Short/schwa	1.2	7.7	4.4
Schwa	9.3	7.3	8.3
Schwa/zero	0.7	2.9	1.8
Zero	10.2	9.7	10.0
Unintelligible	7.5	7.2	7.3

Table 1: Relative frequencies of vowel quality labels for all stimuli (%)

Table 1 shows quite some variation in the relative frequencies of the labels. When we examine the distribution of vowel quality labels used by all transcribers, the label 'short' is assigned most frequently to the stimuli, followed by the label 'long'. All other labels do not reach a frequency of more than 10%. In general, the transcribers show a clear preference for the label 'short' and 'long' in order to score the target vowels, which are all phonologically long. This preference could reflect the actual presence of many short and long vowels in the stimuli. However, the regional perception of the labellers also may have played a role of importance. Therefore, we investigate whether there are dissimilarities in the labelling of NL and B transcribers. An exploratory statistical analysis reveals that the label frequencies prove to be significantly dependent on the regional background of the labellers ($\chi^2 = 824.57$, $p < 0.01$). When we consider the regional frequencies for each label separately, the dependency appears to remain significant ($p < 0.01$) for all labels but the labels 'zero' and 'unintelligible'. Especially the frequencies of the labels 'short' and 'long' deviate considerably. B labellers perceived around 45% of the target vowels as 'short' whereas the NL labellers only did so for about 30% of the same stimuli. The reverse is true for the label 'long': the NL scorers classified more than a third of the stimuli as a long vowel while the B labellers showed no such preference. These marked dissimilarities in labelling of the same stimuli directs at a distinct perception of vowel quality relative to regional background.

To distinguish the influence of the inherent quality of the stimuli from the interference of the transcriber's perception in vowel labelling, we compare the labelling of stimuli produced by speakers from either The Netherlands or Belgium. As already mentioned in the introduction, the use and the degree of vowel reduction in both regions differ considerably. In our corpus of iambic words, the factor regional background of the stimulus proves to be significant ($\chi^2 = 1328.41$, $p < 0.01$). Varying the inherent quality of the stimuli in this way provides us the following knowledge on the impact of the regional variation of the labellers: (1) either the labelling of NL and B transcribers differs according to the regional background of the stimulus. This indicates that the perception of the transcribers is of minor importance. (2) Or the labelling of NL and B transcribers remains similar to the labelling in table 1 irrespective the variation in vowel quality. Thus, the transcription must be heavily biased by perception. First, we discuss the vowel quality labels assigned to NL target vowels.

	NL labellers (n = 1425)	B labellers (n = 1425)	All labellers (n = 2850)
Long	33.4	10.4	21.9
Long/short	4.6	10.9	7.8
Short	14.7	24.6	19.6
Short/schwa	1.1	8.5	4.8
Schwa	14.0	12.5	13.2
Schwa/zero	1.1	5.3	3.2
Zero	17.7	17.1	17.4
Unintelligible	13.5	10.7	12.1

Table 2: Relative frequencies of vowel quality for NL stimuli (%)

Table 2 shows a large variety in labels assigned by all labellers to the target vowels. This label distribution may indicate a heterogeneous vowel quality present in the stimuli produced by speakers from The Netherlands. When we focus on the label frequencies of NL and B transcribers, there seem to be clear dissimilarities in the labelling patterns. This general impression is confirmed by the statistical analysis: vowel quality labelling proves to be significantly dependent on the regional background of the labeller ($\chi^2 = 372.32$, $p < 0.01$). A more fine-grained analysis reveals that the dependency remains significant ($p < 0.01$) for all labels but the labels 'schwa', 'zero' and 'unintelligible'. As in table 1, the largest differences in labels frequencies are reached for the label 'long' and 'short'. Again, NL labellers perceive most vowels as 'long' whereas their Belgian colleagues have a

[2] http://www.let.uu.nl/~Theo.Veenker/personal/projects/wwstim/doc/en/

clear preference for the label 'short'. In sum, NL and B labellers tend to label target vowels in a consistent manner irrespective of the variation present in the stimuli. In the next subsection, we investigate whether this conclusion can be extended to the labelling of B stimuli.

	NL labellers (n = 1257)	B labellers (n = 1257)	All labellers (n = 2514)
Long	36.4	2.8	19.6
Long/short	7.2	14.3	10.7
Short	48.4	69.8	59.1
Short/schwa	1.2	6.8	4.0
Schwa	4.1	1.5	2.8
Schwa/zero	0.4	0.2	0.3
Zero	1.7	1.4	1.6
Unintelligible	0.6	3.1	1.9

Table 3: Relative frequencies of vowel quality for B stimuli (%)

A first analysis of the label frequencies of all labellers indicates that almost 90% of the stimuli produced by Belgian speakers have a vowel quality that ranges between 'long' and 'short'. Labellers from The Netherlands and Belgium appear to diverge in the labels they have assigned to the stimuli. Not surprisingly, the statistical analysis reveals that the labelling of vowel quality depends significantly on the regional background of the transcribers ($\chi^2 = 528.10$, $p < 0.01$). This dependence remains significant ($p < 0.01$) for most labels separately except for the labels 'schwa/zero' and 'zero'. The B transcribers have an overwhelming preference for the label 'short' whereas the NL labellers still use the label 'long' quite frequently besides the label 'short'. As in table 2, we have to conclude that the vowel quality distribution for the B stimuli shows striking parallels with the overall pattern.

In sum, we have attested a marked dissimilarity in labelling between the labellers from The Netherlands and Belgium. The NL labellers have a clear preference for the label 'long' irrespective of the fact that the target vowels originate from The Netherlands or Belgium. The Belgian labellers on the other hand prefer the label 'short' in any case. This distribution must be the result of a deviating perception of vowel quality with labellers from The Netherlands and Belgium.

3.2 Inter-labeller agreement

In the previous section, we have discussed the distribution of the vowel quality labels among the labellers from The Netherlands and Belgium. This method disregards the way each single stimulus is scored and the degree of agreement the labellers reach in scoring each stimulus. This knowledge, however, puts the impact of regional perception on vowel labelling in a new perspective: (1) high agreement among labellers indicates little variation in labelling and consequently points at the presence of a shared strategy the labellers can follow while performing the transcription task. (2) Low agreement indicates much variation among the labellers in assigning a certain label to each stimulus. To measure the degree of agreement among our set of labellers, we calculated the percentage agreement for each pair of labellers.

	NL1	NL2	NL3	B1	B2	B3
NL1	-					
NL2	49.1	-				
NL3	54.7	46.9	-			
B1	43.0	40.8	46.3	-		
B2	31.5	34.3	34.1	49.1	-	
B3	42.3	43.4	47.5	56.7	49.7	-

Table 4: Inter-labeller agreement for all stimuli (%, n = 894)

There is considerable variation in the degree of agreement between a pair of labellers in this experiment: the percentage agreements in table 4 range between 31.5% and 56.7%. On average, two labellers agree on 44.6% of the vowel quality labels they assigned to the stimuli. This rather low percentage can be partially attributed to the many labels the scorer could choose between. The highest inter-labeller agreements are reached between labellers from the same region. The average inter-labeller agreement for the NL labellers is 50.2% and for the B scorers 51.8%. Markedly more modest inter-labeller agreements occur with labellers from different regions. A comparison of percentage agreements across regions yields an average of 40.4%. Apparently, labellers follow divergent strategies to score vowel quality according to their regional background. This observation points towards a deviant perception of the stimuli by labellers from The Netherlands and Belgium. This conclusion concurs with the results from section 3.1.

When we separate the stimuli produced by the speakers from The Netherlands and those produced by Belgians, we can address the question if the labellers attain the same degree of agreement irrespective of the variation in the stimuli. Table 5 shows the inter-labeller agreement for stimuli produced by speakers from The Netherlands.

	NL1	NL2	NL3	B1	B2	B3
NL1	-					
NL2	52.4	-				
NL3	59.6	49.1	-			
B1	47.6	41.7	42.5	-		
B2	31.4	32.6	27.6	38.5	-	
B3	47.8	45.3	46.5	48.2	42.3	-

Table 5: Inter-labeller agreement for NL stimuli (%, n = 475)

As opposed to table 4, not all labellers from the same region reach a high degree of agreement. Solely NL labellers attain a pairwise agreement that ranges between 49.1% and 59.6%, with an average of 53.7%. Note that this average is higher than the 50.2% obtained in table 4. Agreement between B labellers, on the contrary, is more modest and varies from 38.5% to 48.2%, with an average of 43.0%. This average agreement is markedly lower than the value obtained in table 4 and is only slightly above the average percentage agreement for the comparison of NL and B labellers in table 5 (40.3%). Apparently, B labellers had much more trouble scoring the NL target vowels than the NL transcribers had. Perhaps the former did not feel familiar with the pronunciation of the target vowels produced by speakers from The Netherlands. The NL

labellers, on the contrary, have a clearer perception of the stimuli that reflect their native sound system and consequently reach higher labeller-agreement. In the next subsection, we verify whether this conclusion can be extended to stimuli produced by Belgian speakers.

	NL1	NL2	NL3	B1	B2	B3
NL1	-					
NL2	45.3	-				
NL3	49.2	44.4	-			
B1	37.7	39.9	50.6	-		
B2	31.7	36.3	41.5	61.1	-	
B3	36.0	41.3	48.7	66.3	58.0	-

Table 6: Inter-labeller agreement for B stimuli (%, n = 419)

The highest inter-labeller agreements displayed in table 6 are reached between B labellers and range from 58.0% up to 66.3%, average 61.8%. These values are substantially higher than the agreements among B labellers in table 4 and 5. NL labellers, on the contrary, attain a much lower agreement with percentage agreements which vary between 44.4% and 49.2%, average 46.3%. The lowest agreement percentages are reached by labellers with a different regional background with an average of 40.4%. These tendencies indicate that B labellers show less variation in classifying the stimuli produced by B speakers than the NL scorers do. Apparently, the NL labellers lack the categorical cues a B scorer can fall back on while transcribing stimuli for his or her own region.

In sum, the exploration of inter-labeller agreement at the stimulus level showed that the highest agreement is reached between labellers with the same regional background. This observation brings us to the hypothesis that labellers from the same region have a shared perception that is deviant from the perception of other regions. Furthermore, inter-labeller agreement between labellers from the same region is higher when the labellers are asked to score stimuli from their native region. We assume that labellers have less trouble perceiving stimuli that reflect their native sound system the best.

4. Conclusion

In this paper, we have investigated the dissimilarities in the labelling of Dutch vowel quality performed by transcribers from The Netherlands and Belgium. With this case study, we have attempted to assess the influence of the labeller's regional background on phonetic transcriptions. The distribution of label frequencies and inter-labeller agreement indicate that labellers from The Netherlands and Belgium have a deviant perception of vowel quality in Standard Dutch. This finding has severe implications for the planning and evaluation of phonetic transcriptions of spoken language resources. Although we cannot remedy the inevitable interference of perception and especially regional background in phonetic transcriptions, it is of major importance to be very alert for this bias. Therefore, we strongly advise compilers of spoken language corpora to document the regional background of the human transcribers thoroughly so the future users of the phonetic transcriptions can control this transcriber variable according to their needs.

5. References

Abramson, A.S. & L. Lisker (1967). Discriminability along the voicing continuum: cross-language tests. In: *Proceedings of the VIth International Congress of Phonetic Sciences*: 569-573.

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

Cucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study.* Doctoral dissertation, Nijmegen.

Delattre, P.C., A.M. Liberman, F.S. Cooper & L.J. Gerstman (1952). An experimental study of the acoustic determinants of vowel color: observations on one- and two-formant vowels synthesized from spectographic patterns. In: *Word* 8: 195-210.

De Schutter, G. (1994). Dutch. In: E. König & J. Van der Auwera (eds.). *The Germanic languages*. London, Routledge: 439-477.

Ernestus, M. (2000). *Voice Assimilation and Segment Reduction in Casual Dutch. A Corpus-Based Study of the Phonology-Phonetics Interface.* Utrecht, LOT.

Hotzenköcherle, R. (1962). *Sprachatlas der deutschen Schweiz*. Bern, Francke Verlag.

Jaberg, K. & J. Jud (1927). Transkriptionsverfahren, Aussprache und Gehörsschwankungen (Prolegomena zum 'Sprach- und Sachatlas Italiens und der Südschweiz'). In: *Zeitschrift für romanische Philologie* 47: 170-218.

Kloots, H., G. De Schutter, S. Gillis & M. Swerts (2003). Verdoffende vocalen en klinkers die verdwijnen: een casestudy. In: *Nederlandse Taalkunde* 8: 231-254.

Pée, W. (1971). Blancquaerts reeks Nederlandse dialektatlassen. Een dringende toelichting. In: *Taal en Tongval* 22: 131-137.

Ringaard, K. (1964). The phonemes of a dialectal area, perceived by phoneticians and by the speakers themselves. In: *Proceedings of the Vth International Congress of Phonetic Sciences*: 495-501.

Shriberg, L.D., J. Kwiatkowski & K. Hoffman (1984). A procedure for phonetic transcription by consensus. In: *Journal of Speech and Hearing Research* 27: 456-465.

Strange, W. & J.J. Jenkins (1978). Role of linguistic experience in the perception of speech. In: R.D. Walk & H.L. Pick (eds.). *Perception and experience*. New York, Plenum Press: 125-169.

Acknowledgements

This research is partially funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) within the project Flexible Large Vocabulary Recognition and by the Fund for Scientific Research – Flanders (FWO).