AGREEMENT IN CONSENSUS TRANSCRIPTIONS OF TRAINED AND UNTRAINED TRANSCRIBERS

W.H. Vieregge* and A.P.A. Broeders**

*Department of Language and Speech, University of Nijmegen, Netherlands
**National Forensic Science Laboratory, Rijswijk, Netherlands

ABSTRACT

Consensus transcriptions were made by trained as well as untrained transcribers of several segmental variables in Dutch. A randomly selected subset of these variables was transcribed twice by both groups. Two hypotheses were tested: the degree of agreement between non-contemporary consensus transcriptions is a measure of their validity; trained transcribers reach higher consistency levels than untrained transcribers.

1 INTRODUCTION

In her discussion of the meaning of the terms validity and reliability as applied to phonetic transcription, Cucchiarini [1] suggests that, in the absence of a proper benchmark for the estimation of the validity of a transcription, the consensus transcription may serve as a viable alternative. The consensus transcription is often proposed as a procedure which will reduce errors in transcriptions and increase agreement among transcribers (Shriberg et al. [2]). We have found that the consensus transcription can serve as a suitable format for the analysis of intra- and interspeaker variation in the realization of certain segmental variables in Dutch (Vieregge and Broeders [3]). However, we are not aware of the existence of studies in which the agreement between consensus transcriptions was examined to see if this would produce a more satisfactory measure of transcription validity.

2 AIM OF THE STUDY

The main aim of the investigation was to look into the possibility of testing two hypotheses, both of them inspired by our experience with the consensus transcription and following from the claim that this transcription procedure tends to reduce errors due to inattention, and leads to greater agreement between transcribers (Ting et al. [4]). If this is true, the degree of agreement found in consensus transcriptions made at different points in time should provide a good measure of the validity of these transcriptions. In other words, we hypothesize that consensus transcriptions are more valid as they are replicated with greater consistency.

On the assumption that trained transcribers may be expected to be more competent than untrained tran-scribers, a second hypothesis can be formulated, viz. that trained transcribers will reach a higher degree of consistency than untrained transcribers.

In order to test these hypotheses consensus transcriptions made as part of a study of inter- and intraspeaker variation in the realization of segmental variables in Dutch were used.

3 THE SPEAKERS

The speech samples were produced by 7 educated speakers of Dutch, hailing from various parts of the country. The amount of regional variation in their speech varied from hardly any to quite marked. All speaker were male, with ages ranging between 25 and 50. The speech style could be described as quasi-spontaneous: all seven speakers were asked to give a description of what they saw in three drawings, showing a street scene, some shops and a living-room respectively. The duration of their descriptions varied from 2 minutes to 2 minutes and 45 seconds. The material forms part of a larger corpus collected for a different purpose by our colleague Van Bezooijen, who kindly made the recordings available to us.

4 THE VARIABLES

The segmental variables used in this investigation form a random subset of the larger set of variables transcribed as part of a study to look into the inter- and intraspeaker variation of certain segmental variables in Dutch (Broeders and Vieregge [5]. They are presented in Table 1.

Table 1. Variables used in the investigation (N: the number of tokens per variable in the subset).

Variable	N
/x/	21
/z/	14
/v/	14
schwa-insertion after /r,l/	13
assimilation of voice before /b,d/	14
n-deletion after schwa	14

The variables themselves were selected as part of the earlier study on the basis of their expected variability in Dutch. The subset of tokens used in the present study was picked at random.

5 THE TRANSCRIBERS

Consensus transcriptions were made by two trained transcribers, the present writers, and nine pairs of untrained transcribers. The latter were all Language and Speech Pathology students of the University of Nijmegen, all of them qualified speech therapists, who made the transcriptions in part fulfilment of the requirements of a 120-hour course in phonetic transcription taught by the first author. They were instructed to produce a consensus transcription in accordance with the IPA conventions [6], which they were told would later be assessed by their teacher.

6 PROCEDURE

The trained transcribers made the second transcription of the random subset several months after the first. For the untrained transcribers both transcriptions were made as part of a single transcription assignment but the work was structured in such a way that, unlike the trained transcribers, they may be assumed to have been unaware of the fact that they were transcribing (some of) the variables twice.

7 RESULTS

The results are presented in Table 2. Transcriptions were considered to be in agreement if the same phonetic symbol plus any of a limited number of diacritics was used on both occasions. They are expressed as the percentage agreement reached per variable. The percentages given for the untrained transcribers are averaged for the 9 pairs.

Table 2. Variables used in the investigation (U: untrained, T: trained transcribers; N: number of tokens per variable in the subset).

Variable	U	T	N
/x/	64.6	76.2	21
/z/	78.6	92.9	14
/v/	69.0	92.9	14
schwa-insertion	80.3	92.3	13
assimilation	69.0	71.4	14
n-del	90.5	85.7	14

8 DISCUSSION

It appears that, with the exception of the last variable, trained transcribers achieve considerably more agreement than untrained transcribers. The difference in the amount of agreement found between trained and untrained transcribers is significant ($t = -2.44$; $p < 0.05$; one-tailed).

At first sight, the results seem to confirm the second hypothesis that trained observers reach higher consistency levels than untrained transcribers. However, inspection of the actual transcriptions suggests that there are one or two complicating factors at work whose effects, while undeniably present, are difficult to quantify. On the one hand, there is the fact that some of the variables are essentially binary (n-deletion, schwa-insertion). Obviously, all other things being equal, agreement is likely to be higher if the number of options is small and vice versa. On the other hand, there are variables like /x/ that easily run into as many as 5 different symbolizations, each combining with several diacritics. Of course, in principle this embarras de choix applies to trained and untrained transcribers alike. In practice, however, it must be expected to work against the trained transcribers, as their greater familiarity with the phonetic symbol set and greater experience as trained listeners should make more options available to them. By the same token, untrained listeners are likely to reach higher agreement between transcriptions because they have a smaller set of symbols to choose from. On balance though, the results lend support to our second hypothesis: agreement between consensus transcriptions is higher for trained than for untrained transcribers.

However, in the course of the discussion we have seen that there are strong indications that our first hypothesis is not tenable as it stands. Agreement per se is a necessary but not a sufficient criterion for validity. It is simply not the case that the consensus transcription that happens to show the highest degree of agreement is for that reason also the more valid one. What is essential of course is that the consensus transcriptions are made by competent transcribers. If agreement is high between non-contemporary replications of consensus transcriptions by experienced transcribers it is reasonable to assume that these can be used as a criterion against which the quality of other transcriptions can be measured.

9 A VALIDITY CRITERION

If we revise our hypothesis in the light of these observations, we are in a position to judge the quality of the consensus transcriptions made by the pairs of untrained transcribers, using the consensus transcriptions of the trained transcribers as our criterion (Vieregge [7], p. 31). Obviously, this will only be possible for those cases where the trained transcribers produced identical transcriptions in the two consensus sessions. While it is clear that this introduces a degree of inaccuracy in those cases where the trained transcribers disagree between the two sessions, it is safe to assume that the effect of this is marginal. After all, for most variables the agreement scores reached by the trained transcribers are quite high, and what discrepancies do arise will by and large occur in respect of the transcription of the rather more problematical variables, on which untrained transcribers would be unlikely to do better in the first place.

10 THE VALIDITY CRITERION APPLIED

If we apply the above criterion to the transcriptions made by the untrained transcribers this yields two types of information. First, we can calculate the score for each variable averaged over the nine pairs of untrained transcribers. This figure expresses the extent to which the untrained transcribers, on average, produced transcriptions that are identical to those of the trained transcribers. It gives an indication of how well the variable in question was transcribed by the untrained transcribers. The results are presented in Table 3, which also specifies the number of tokens for each variable transcribed identically by the trained transcribers and used in the validity criterion. It is worth noting that on average the transcription of the variables /x/, /v/ and *assimilation* deviates in the majority of cases from that of the trained transcribers, which may be taken as an indication of the difficulty these variables present.

Table 3. Variables used in the investigation (Mean: average score per variable; N2: number of tokens per variable used in validity criterion; N1: total number of tokens per variable in the subset).

Variable	Mean	N2	N1
/x/	46.9	16	21
/z/	70.1	13	14
/v/	41.4	12	14
schwa-insertion	67.8	12	13
assimilation	43.9	10	14
n-deletion	90.3	12	14

Alternatively, we can calculate the performance of the separate pairs of untrained transcribers for each variable, again using the identical transcriptions of the trained transcribers as our criterion. The results are presented in Table 4. It appears that average performance scores vary between 53 and 69%.

Table 4. Average scores per pair over all the tokens used as part of the validity criterion (For reasons of space, numbers are rounded off to the nearest integer; P: pair; V: variable; s'a: schwa-insertion; ass: assimilation; n-del: n-deletion).

P\V	/x/	/z/	/v/	s'a	ass	n-del
1	41	62	50	63	20	83
2	28	81	42	42	75	83
3	28	73	17	88	70	92
4	50	62	38	75	60	100
5	44	81	38	83	25	63
6	50	58	58	88	45	100
7	84	73	46	79	30	100
8	53	89	42	25	40	100
9	44	58	42	67	30	92

11 CONCLUSION

The results of the study lend support to our hypothesis that trained transcribers reach higher consistency levels in replicated consensus transcriptions than untrained transcribers.

It also appears that, while agreement between consensus transcriptions is not a good validity criterion per se, high agreement between non-contemporary consensus transcriptions made by trained transcribers can be used as a measure of transcription validity.

REFERENCES

[1] Cucchiarini, C. (1993) *Phonetic Transcription: A Methodological and Empirical Study*, Nijmegen.

[2] Shriberg, L.D. et al. (1984) 'A procedure for Phonetic transcription by Consensus: A Research Note', *Journal of Speech and Hearing Research* 27, 456-465.

[3] Vieregge, W.H. and Broeders, A.P.A. (1993) 'Intra- and Interspeaker Variation of /r/ in Dutch', in: *Proceedings of Eurospeech 93*, 267-270.

[4] Ting, A. et al. (1970) 'Phonetic Transcription: A Study of Transcriber Variation', *Report*, Madison: Wisconsin University.

[5] Broeders, A.P.A. and Vieregge, W.H. (1991) 'Intraspeaker Variation on the Segmental Level: a Transcription-based Approach', in: *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence: Université de Provence, Vol 5: 46-49.

[6] (1993) 'The International Phonetic Alphabet' *Journal of the International Phonetic Association* 23(1), center pages.

[7] Vieregge, W.H. (1987) 'Basic aspects of Phonetic Segmental Transcription', in: Almeida, A. and Braun, A. (eds.) *Probleme der phonetischen Transkription*', Stuttgart: Franz Steiner Verlag.