WHAT DO TRANSCRIPTION AGREEMENT INDICES SAY ABOUT TRANSCRIPTION ACCURACY?

Catia Cucchiarini
Centre for Language and Migration, Louvain, Belgium

ABSTRACT

This paper deals with the drawbacks of a common measure of transcription agreement, percentage agreement. It is argued that this metric does not give a realistic representation of transcription similarity and that it can be easily inflated by adopting a higher level of abstraction (one involving fewer categories) than the one recorded in transcriptions, when calculating agreement. An alternative measure of transcription (dis)similarity is presented and its advantages over percentage agreement are discussed.

INTRODUCTION

In the last few years the issue of transcription reliability has received considerable attention in the literature (for a review see [1]). Since it is known that phonetic transcriptions tend to contain an element of subjectivity, it is now common practice to check the objectivity and accuracy of transcription data before using them for research. To give an indication of the accuracy of the transcriptions on which their findings are based, researchers usually provide so-called transcription reliability or agreement indices. The most used measure for this purpose is percentage agreement, which is computed by comparing two transcriptions symbol by symbol and by taking the percentage of identical symbols in the two strings. Although this index in reality expresses agreement between transcriptions, the term reliability index is often used instead [1]. However, this is not correct given that phonetic transcription involves classification into categories (phonetic symbols) which are not ordered. In other words, the variables have the properties of measurement at the nominal level. At this level there can be no "proportionality of ratings" [2], a notion which is crucial to reliability. For these reasons, the term agreement index will be used in the present paper (for further details, see [2] and [3]).

In general, no standards or levels of significance are available for transcription agreement indices. Although it seems that indices should be as high as 75% [4] or 85% [1, 5] in order to be acceptable, the plausibility of these agreement values has never been considered, let alone demonstrated. Consequently, it is not clear whether high percentages of agreement really correspond to high degrees of transcription accuracy. This point is addressed in the following section.

THE IMPACT OF CHANCE AGREEMENT

One of the things that tend to be overlooked in the literature is that the value of an agreement index does not only depend on the degree of accuracy of the transcriptions in question, but also on the number of categories on which two transcribers, or one and the same transcriber on different occasions, have to agree. The number of categories involved in the judgement partly determines the impact of chance agreement, that part of agreement that is determined by chance alone. It can be stated that agreement indices tend to be higher for simple judgements such as correct / incorrect than for more complex decisions involving a greater number of

categories like, for instance, the precise description of the vowels produced by a speaker. Consequently, agreement indices are expected to be higher for broad transcriptions than for narrow transcriptions. This means that providing an agreement index is not sufficient to give a precise idea of the degree of accuracy of transcriptions. One should also mention the number of categories out of which transcribers could choose (level of abstraction).

Unfortunately, the impact of chance agreement on transcription agreement is often overlooked in the literature (see for instance [1], [4]), with the result that in computing agreement researchers often reduce the number of categories in order to achieve the longed for 75% or 85% of agreement. However, as has been pointed out [6] "higher coefficients of agreement that result from use of simpler observation codes do not guarantee that observer recordings are accurate".

The fact that agreement indices are so sensitive to the number of categories involved has to do with the way in which transcription differences are treated, when it comes to determining the degree of agreement. In general it is assumed that certain transcription deviations are more serious than others. For instance, transcribing [b] instead of [p] would be considered to be less serious a mistake than transcribing [l] instead of [p]. Similarly, differences concerning diacritical marks are assumed to be less serious than differences concerning basic symbols.

However, these differences in gravity are usually neglected when transcriptions are compared symbol by symbol. The only thing researchers look at is whether the symbols and the accompanying diacritics are identical in the two transcriptions or not. The result is that any difference will affect the agreement index in the same way, regardless of its degree of gravity. In turn the agreement index will be extremely sensitive to the degree of detail recorded in the transcriptions. This also means that this kind of index can easily be inflated by reducing the degree of detail, not when making the transcriptions, but when calculating agreement.

In addition to making percentage agreement so subject to manipulation, this procedure is also unrealistic. It is obvious that a measure of transcription agreement should take account of the various degrees of (dis)similarity between speech sounds. Moreover, when diacritics are present, one should not merely check whether the same diacritic is used or not, as is sometimes done [1]. As a matter of fact, a diacritic is an integral part of the phonetic symbol, since it partly determines its meaning, so it would be wrong to consider them as separate elements. Furthermore, different diacritics used with different basic symbols could represent very similar speech sounds. For example, the two vowel symbols [o] and [ɔ] can be made more similar by adding appropriate diacritics for 'height' properties as follows: [o̝] [ɔ̝]. The higher degree of similarity between these two transcriptions would not be reflected by percentage agreement. In fact, in this metric the two differences would be combined thus obtaining a very low agreement index.

AN ALTERNATIVE APPROACH TO TRANSCRIPTION EVALUATION

In the previous section I have argued that percentage agreement is no adequate measure of transcription similarity, because it is too sensitive to the level of abstraction of transcriptions and because it treats agreement between phonetic symbols in an all-or-none way. In an attempt to overcome these problems, an alternative measure of transcription (dis)similarity was developed, which does take account of the various degrees of similarity (or difference) between speech sounds and of the effect of

diacritics on basic symbols [3]. This metric is called average distance because it gives an indication of the mean distance between the vowels and/or the consonants of two transcription strings.

The average distance is based on the feature matrices defining vowels and consonants that are presented in [7]. These matrices were obtained by combining results of experiments on proprioceptive speech sound dissimilarity with phonetic knowledge. The values contained in these matrices make it possible to express the degree of dissimilarity between all possible pairs of sounds in numerical form. Each speech sound is assigned a numerical value for each of the defining features in the matrices. Dissimilarity values for pairs of speech sounds can be determined by calculating city-block distances between them. This is done by comparing two speech sounds feature by feature and by summing the individual differences. Overall dissimilarity values for the vowels and consonants contained in each transcription pair are obtained by computing the mean for all vowel and consonant pairs, respectively.

One of the advantages of this method is that it gives a more realistic impression of the degree of (dis)similarity between two transcriptions. For instance, with this metric it is possible to indicate that there is more similarity between [b] and [p] than between [l] and [p]. In other words, this metric goes beyond the mere appearance of phonetic symbols (are they identical or not?) and takes account of their meaning (which speech sounds do they represent and how are they related to each other?). Moreover, in this metric it is possible to discount the impact of diacritics on basic symbols before computing the distance between two corresponding symbols. Also in this case the meaning of diacritics is considered (what is the effect of adding this specific diacritical mark to this basic symbol?) and not merely their presence or absence.

This brings us to another advantage of this metric, namely that in computing agreement one can take account of all the details that have been recorded in transcriptions in a realistic way. It does not make sense to carry out transcriptions at a certain level of abstraction, for instance narrow transcription, and then compute agreement at a higher level of abstraction, i.e. broad transcription, in order to achieve acceptable percentages of agreement. If researchers make narrow transcriptions there must be a reason for this, i.e. the details are relevant to their research. It is therefore important to know to what extent transcribers agree at this level of specificity. It is obvious that the more details transcribers record, the less likely they are of agreeing with each other. However, one should avoid using a measure such as percentage agreement which penalizes detailed transcriptions in an unwarranted way.

That the average distance is a more appropriate measure than percentage agreement was also revealed by the results of an evaluation test described in [3]. For 50 transcription pairs the overall dissimilarity between vowels and consonants was computed by means of the two metrics, i.e. the average distance and percentage disagreement, the complement of percentage agreement. The values thus obtained were compared with the dissimilarity judgements expressed by 19 experienced phoneticians for the same transcription pairs. The phoneticians were asked to assign a mark varying between 1 (no similarity) and 10 (no difference) to the vowels and consonants of each transcription pair. The reliability coefficient computed for these judgements (formula for composite ratings with raters as a random factor) appeared to be high (0.97).

It turned out that the average distance better reflected the phoneticians' judgements than percentage agreement. As a matter of fact, the correlation coefficient was higher in the former case ($r = -0.86$, $df = 48$, $p < 0.01$) and lower in the latter ($r = -0.68$, $df = 48$, $p < 0.01$). The two coefficients also appeared to be significantly different ($t_{47} = -2.94$, $p < 0.01$). It should be noted that the correlation coefficients are negative in both cases because the phoneticians' judgements indicate similarity and the other two measures dissimilarity.

On the basis of this test it seems that when phoneticians judge the degree of (dis)similarity between pairs of transcriptions, they do not limit themselves to establishing whether the symbols in the two strings are identical or not, but try to determine to what extent they are similar. Apparently, phoneticians consider agreement between phonetic symbols to be gradual, not all-or-none. This is precisely what happens when the average distance is calculated (for a fuller account of this method and of its advantages over percentage agreement, the reader is referred to [8]).

CONCLUSIONS

The new measure of transcription agreement proposed in this paper, the average distance, differs from the more common percentage agreement, because it makes it possible to indicate different degrees of (dis)similarity between corresponding phonetic symbols. Only two of the advantages of this method are discussed here. First, the average distance gives a more correct representation of transcription (dis)similarity because it takes the meaning of phonetic symbols and diacritics into account. Second, since different degrees of (dis)similarity between phonetic symbols can be distinguished, this measure is less sensitive to the level of abstraction of transcriptions. For these reasons it seems that the average distance provides a more appropriate measure of transcription agreement than percentage agreement. This was also confirmed by experimental results.

ACKNOWLEDGEMENTS

The present research was supported by the Linguistic Research Foundation, which is funded by the Netherlands Organization for Scientific Research, NWO. I am indebted to M. Biemans for the realisation of the transcription evaluation experiment described in this paper.

REFERENCES

[1] Shriberg, L.D. & Lof, L. (1991), "Reliability studies in broad and narrow phonetic transcription", *Clinical Linguistics and Phonetics*, vol. 5, pp. 225-279.

[2] Tinsley, H.E.A. & Weiss, D.J. (1975), "Interrater reliability and agreement of subjective judgments", *Journal of Counseling Psychology, Vol.* 22, pp. 358-376.

[3] Cucchiarini, C. (1993), *Phonetic transcription: a methodological and empirical study* (PhD thesis, University of Nijmegen).

[4] Henderson, F.M. (1938), "Accuracy in testing the articulation of speech sounds", *Journal of Educational Research*, vol. 31, pp. 348-356.

[5] Pye, C., Wilcox, K.A. & Siren, K.A. (1988), "Refining transcriptions: the significance of transcriber 'errors'", *Journal of Child Language*, vol. 15, pp. 17-37.

[6] McReynolds, L. & Kearns, K.P. (1983), *Single-subject experimental designs in communicative disorders* (Austin, TX: Pro-Ed).

[7] Vieregge, W.H., Rietveld, A.C.M. & Jansen, C.I.E. (1984), "A distinctive feature based system for the evaluation of segmental transcription in Dutch", *Proceedings of the Xth International Congress of Phonetic Sciences*, Utrecht.

[8] Cucchiarini, C. (1996), "Assessing transcription agreement: methodological aspects", to appear in *Clinical Linguistics and Phonetics*, vol. 10.