Assessing Agreement on Segmentations by Means of *Staccato*, the *Segmentation Agreement Calculator according to Thomann*

Andy Lücking[1], Sebastian Ptock[2], and Kirsten Bergmann[2]

[1] Goethe-University Frankfurt am Main
[2] Bielefeld University, CRC 673, B1
luecking@em.uni-frankfurt.de
{sebastian.ptock,kirsten.bergmann}@uni-bielefeld.de

Abstract. Staccato, the *Segmentation Agreement Calculator According to Thomann*, is a software tool for assessing the degree of agreement of multiple segmentations of some time-related data (e.g., gesture phases or sign language constituents). The software implements an assessment procedure developed by Bruno Thomann and will be made publicly available. The article discusses the rationale of the agreement assessment procedure and points at future extensions of Staccato.

1 Segmentation and Annotation in Gesture Studies

No matter what your favourite definition of co-verbal "hand and arm gesture" is, at the basic kinematic level gestures are manifested by spatio-temporal bodily movements. However, not every body movement constitutes gestural behaviour. Accordingly, a prerequisite of any empirical, data-driven account to gestures is to locate those segments in the continuous stream of hand and arm movements that make up the gestures. Furthermore, gestures are structured in themselves: a gesture basically has a tripartite movement profile, consisting of a preparation, a stroke, and a retraction phase [5]. Therefore, gestural hand-and-arm movements not only are to be identified, but are also to be subdivided into phases. In addition, temporary cessations of motion might occur either before or after the stroke phase – the so-called pre- and post-stroke holds [7].

The demarcation of the temporal parts of a motion that constitutes a gesture and its phases is referred to as the *segmentation problem* throughout this article. The segmentation problem is not confined to gesture studies, of course; it also prevails in sign language research, and, generically, in accounts to temporally organised behaviour in general (think, e.g., of phoneticians that identify phonemes in a sound stream, choreographers that decompose dance figures, speech pathologists that demarcate deglutition phases, or traffic authority employees studying rush hour traffic). We focus on empirical work on gesture, however, where segmentation is typically carried out on video recordings, in which the significant movements are to be demarcated with respect to the video's time line.

The segmentation problem has to be kept apart from the *annotation problem*. In an annotation session, the annotator (or coder, or rater) has to classify items according to a set of response categories. A common example is that of nurses that have to assess the psychological health state of patients, say, in terms of "happy", "stoic", and "sad".[3] In case of gesture studies, the "patients" are gestures that are classified according to categories defined in a classification scheme. For instance, in case of the Speech-and-Gesture Alignment corpus (SaGA [8]) one set of response categories has been derived from the so-called *representation techniques* [9], that is, from the displaying functions gestures employ in order to depict something, like, for instance, *drawing* or *locating*.

A logical primacy relation obtains between segmentation and annotation: gesture annotation presupposes segmentation. Figure 1 visualises the connection between segmentation and annotation: the items defined by segmenting the continuous data stream are the objects of subsequent annotation. However, the subordinate response categories may influence the superordinate segmentation of the observable movements: since exact demarcation might be relative to a functional or Gestaltist perspective of annotation labels, a holistic understanding of a given gestural movement may impose top-down constraints on lower-level segmentation of this movement. For instance, whether a complex motion pattern makes up just one intricate shape-depicting gesture or manifests two consecutive locating gestures may well be a matter of interpreting the movement in its context, most notably, in the context of its accompanying speech.

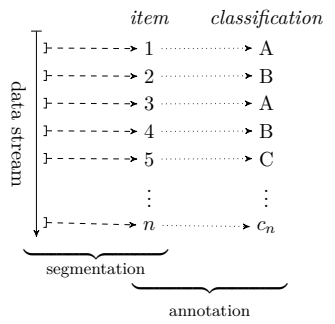
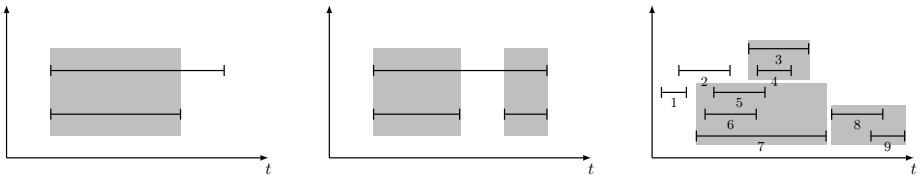


Fig. 1: The relationship between segmentation and annotation. Firstly, the segmentation of the data stream of movements results in a set of n items. Secondly, these items are classified according to a set of response categories.

2 Accounting for the Reliability of Segmentations

Like all informational enrichments of primary data, gesture segmentation has to be evaluated with regard to reliability (see [2] for an introduction into the topics of reliability and validity). The standard method for gauging reliability of annotations are chance-corrected assessments of the agreement between multiple annotations of the same material, either by one annotator (*intra*-annotator agreement) or by various annotators (*inter*-annotator agreement). However, the reliability of gesture segmentation cannot be assessed by these methods. For standard agreement measures – like the wide-spread kappa statistic [3] – are applicable only to *annotation* data, that is data gained in a test design that consists

[3] This particular example is taken from [10, p. 5].



(a) (b) (c)

Fig. 2: Configurations of segmentations. (a) and (b) are undecidable in frequentist terms. (c) illustrates *nuclei* (partially reproduced after Figure 1 of [11, p. 341]).

in classifying a fixed set of given items – cf. Section 1 above. The demarcation of movement segments, by contrast, first of all determines *what the items are*. But how, then, is gesture segmentation to be evaluated?

Procedures proposed to assess the reliability of segmentations try to measure the degree of agreement between segmentations of different annotators in terms of some metric gauges. As a metric reference value, the time line or the number of frames of the video film containing the recorded data has been used – cf. the *time slices* proposal of [6].

Such frequentist analyses, as we will call them, however, fail to capture higher order structures like the number or the allocation characteristics of marked segments. The core of the assessment problem is exemplified by the segmentation patterns depicted in subfigures (a) and (b) of Figure 2. In the figures, segmentations of two annotators are displayed as horizontal lines, indicating the length of their respective segmentation relative to the temporal x-axis. In subfigure (a) the annotators agree on the occurrence of a single gesture (or gesture phase) but merely assign it a different length. In subfigure (b) the annotators identify a different number of gestures. Both cases exemplify two poles of agreement that each pertinent reliability assessment has to account for:

1. annotators in example (a) share a reasonably common view of how the observed gestures have to be segmented;
2. annotators from example (b) have no shared understanding of the observed movements.

Accordingly, notwithstanding that both pairs of segmentations show the same amount of overlapping, we would nonetheless expect that an method for assessing the reliability of gesture segmentations assigns a higher degree of agreement in case of (a) than in case of (b). A frequentist metric measurement, however, is not able to tell (a) and (b) apart, since the shared extent of markings is the same for both pairs of segmentations (see the gray area in the subfigures). A segmentation assessment has to be able to deal with the (b)-case of demarcations, namely demarcations that coincide in temporal terms but differ in the number of demarcated items.

We propose to employ a method that has been developed by [11].[4] Instead of simple frequentist measures, Thomann utilises graph-theoretical techniques that take structural features like the number of items and their allocations into account. The rationale of the Thomann method is illustrated in Figure 2(c). Each row of segments has been produced by a different annotator, that is, 2(c) assembles the markings of five annotators. To what extent do the annotators agree? In order to prepare an answer to this question, we have to introduce the notion of *nucleus*, that is, an aggregation of segmentations from which a measure of the degree of organisation – the operationalisation of agreement – is derived. A nucleus is defined in terms of internal homogeneity and external heterogeneity and is indicated by gray boxes in Figure 2(c). The first condition simply requires the segments in a nucleus to mutually overlap. According to this requirement alone, segments 3 to 7 would form a nucleus – what they actually do not. Segments 3 and 4 are excluded by the second condition, which constrains the external overlapping relations of all segments of a nucleus to be indistinguishable. As we can see in 2(c), segment 2 overlaps with segments 5, 6 and 7, but not with segments 3 and 4. Thus, the external relations of 3 and 4 on the one hand and of 5 to 7 on the other hand are distinguishable [11, p. 343]. Applying both conditions yields the nuclei depicted in the illustration, where 7 out of 9 segments are organised in nuclei, that gives an absolute degree of agreement of $^7/_9 \times 100 = 77.78$.

Of course, nucleus formation might to a certain degree be just due to chance. To this end, the absolute degree of agreement is normalised against a random baseline. The random baseline constitutes the reference value for nuclei formation that are expected by chance alone. The resulting value of this normalisation is the *degree of organisation* [11, p. 343]. It ranges in the interval $(-1, 1)$. A value of 0 means that the empirically found number of segment nuclei equals the number expected in random configurations. Note that the degree of organisation makes different configurations of segmentations (say, of various studies) comparable.

Needless to say that the determination of nuclei is too painstaking to do it by hand. Though there is an algorithm implemented by Thomann himself, no publicly available tool for calculating agreement of segmentation is at disposal. We offer such a software tool, called (somewhat will-fully) *Staccato* (***S**egmen**t**ation **A**greement **C**alc**u**lator **a**ccording to **T**h**o**mann*). Staccato is written in platform-independent Java. This stand-alone software will become a component of the standard multimodal annotation tools Elan[5] and Anvil[6].

3 Example Calculation

To illustrate the usage of Staccato, let's assume the situation that several annotators finished their segmentations of gestures using Elan (see Figure 3(a),

[4] We would like to thank Timo Sowa for pointing us at the work of Thomann.
[5] www.lat-mpi.eu/tools/elan
[6] http://www.anvil-software.de/

which is adopted from Figure 2(c) above). In order to analyse agreement between participants, the data is loaded as a CSV file (exported from the annotation software). Subsequently, parameters for the agreement calculation are to be set, namely (1) the number of Monte-Carlo-Iterations, i.e., how often a Monte-Carlo-Simulation (MCS) will be processed for generating random outcomes of the Thomann method, (2) the granularity for annotations' length to adjust the duration of annotations randomly generated from MCS to the appearance of durations in the annotated data, and (3) the level of significance to reject null hypothesis of chance-based agreement.

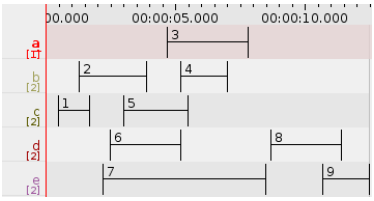
The result of running the agreement calculation with 10 000 MCS, a granularity for annotation length of 10, and $\alpha = 0.05$ is given in Fig. 3(b). The *Degree of Organisation* of 0.549 24 signifies participants' agreement to be much higher than chance (see explanation above). To get a graphical overview of the results, a CSV file can be exported from Staccato and imported into the annotation software (see Figure 3(c)). The result not only shows the Degree of Organisation but also *Fields*, *NucleusNominations* and *Nuclei*. Fields are subsets where all annotations are connected via overlaps. NucleusNominations are nominations that potentially might form a nucleus (see [11, p.44] for a definition). Figure 3(b) also shows normalised data for the MCS outcome.

4 Discussions

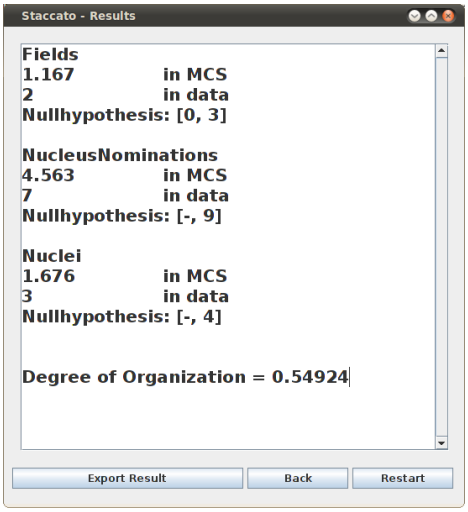
The preceding sections introduced the segmentation problem and argues in favour of the *nucleus*-based account of [11] for assessing the degree of inter- or intra-coder agreement on annotations. The procedure has been implemented in the Staccato tool. Implementation follows the independence assumption between segmentation and annotation, as advocated in Section 1. Given top-down influences on segmentations, there nevertheless might be some reasons do perform a classification-sensitive reliability assessment of segmentations. How this is accomplished within Staccato will be the topic of the discussion in Section 4.1. Subsequently, we contrast our approach to a related one which has recently been proposed.

4.1 Classification-Sensitive Assessment of Segmentation Reliability

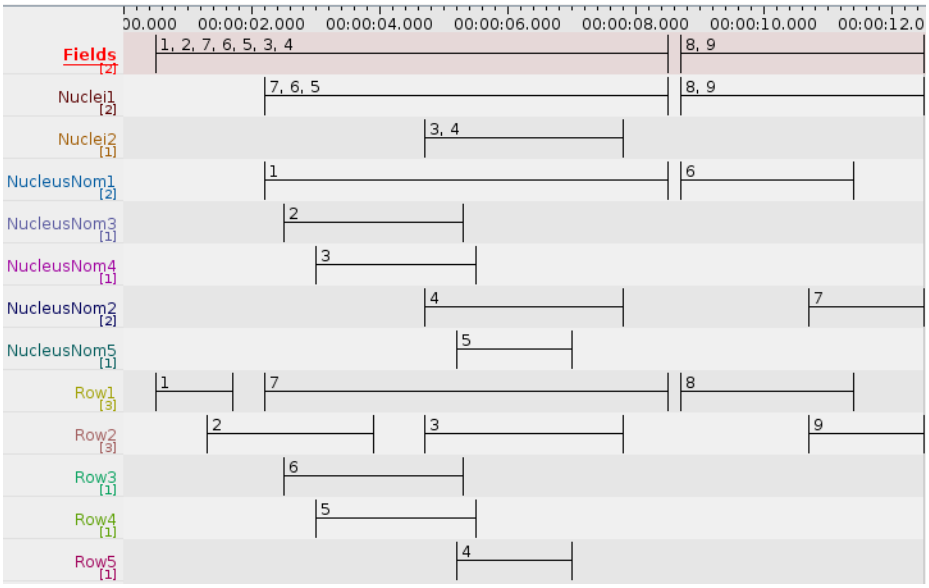
In section 1 we acknowledged a potential top-down influence of annotation on segmentation, but the Staccato treatment so far ignores any annotation information. How can we account for annotation labels in the evaluation of segmentations? There is a straightforward answer to this question: the Staccato procedure is to be made label-aware. That means, that all segmentations are grouped according to the type they are assigned to. The nuclei-based algorithm then operates on this groups instead of the set of segments *in toto*. Each result can then be compared to its random baseline, given an indicator for serious (i.e., not chanced-based) agreement for each class defined by the annotation. This way, an assessment of the combined segmentation/annotation for each labelled



(a)



(b)



(c)

Fig. 3: Annotated data (a), result of the Thomann method in Staccato (b), and output exported from Staccato (c): some tiers are split into further tiers using indices to avoid overlaps. The original data appears in a sorted order as tier *Row* that was generated by Staccato for the purpose of higher performance.

segment is achieved. In order to estimate overall reliability, the mean of the single results can be taken.

Depending on the kind of annotation in question, simply averaging the Staccato outcomes for different annotation groups of segments might be to coarse-grained, probably disguising some inner structure among the annotation labels. Consider an annotation scheme where hypernym labels introduce hyponym sub-labels, for instance an annotation scheme for dialogue acts like the taxonomy of [1]. There, the *Information-providing function* called *Inform* is a super-type for the three dialogue acts *Answer*, *Agreement*, *Disagreement* (where *Answer* and *Disagreement* introduce sub-labels themselves). Intuitively, if in a two-rater study the one rater decides for the hypernym label *Inform* while the other one chooses *Agreement*, their disagreement is not so strong since *Inform* and *Agreement* are compatible due to their hyponym relation. Compatibility vanishes if one rater chooses *Agreement* while the other chooses *Disagreement*, however. Treating both cases alike in an evaluation would result in an unbalanced reliability assessment. What is needed in this case is a weighting function that reflects the relationships among annotation labels and thereby balances evaluation measures. Such a weighting can easily be added: the Staccato result for each annotation group contributes its value modified by certain weighting factor to the average overall outcome. In its present stage of development Staccato does not account for annotations. We will extend Staccato with the option of label-sensitive calculation in the near future.

4.2 Comparison with Holle & Rein

At about the same time we prepared Staccato, [4] (H&R in the following) also approached the segmentation problem. H&R tried to adhere to the kappa statistic. The algorithm they present is a "time stretched" variant of Cohen's kappa for two raters. For each segment demarcated by one of the annotators it is looked whether the other annotator identifies a segment with sufficient overlap, too. H&R furthermore motivate by threshold testing that an overlap of 60% is empirically adequate. Its "conservative" character, that is, its being likened to standard annotation evaluations makes it an attractive procedure. However, we think that there are three shortcomings compared to the Staccato way:

1. H&R's procedure is constrained to the case of two raters. That means that the example presented in Figure 2(c) and our three-annotator segmentation data gained in the SaGA project [8] cannot be handled by the H&R approach. Since Staccato is the more general method, both can only be fairly compared in their overlapping domain, that is the case of two annotators as exemplified by Figure 2(c). H&R have explicitly taken care for the particular overlapping relation of segments drawn in the figure: in order to avoid perfect agreement in this case they have to invoke a 50+% rule of overlap measured at the longer one of segments. Due to that stipulation the configuration illustrated in Figure 2(c) counts as a case of disagreement for H&R. The same result is achieved with Staccato, where it follows naturally from

the internal homogeneity and external heterogeneity constraints. This finding corroborates that Staccato provides the more general procedure, which comprises H&R's approach as a special case.
2. The H&R algorithm does not provide a chance-correction for segmentations. There doesn't seem to be any reason to assume that there is no "segmentation under uncertainty" that gives rise to more or less randomly set markings as is the case in annotation tasks. Staccato incorporates a way to assess a random baseline from the start.
3. H&R binarise annotation label sets with more than two elements. However, binarisation leads to a distortion of results. In the appendix we provide a detailed example that shows exemplarily how binarisation of three annotation categories brings about better kappa coefficients. This means, that the H&R procedure systematically leads to inflated results.

There is also a more subtle difference between our conception of the relationship between segmentation and annotation and the one put forward in [4]. H&R argue for *independence* of segmentation and annotation, but model them in a *dependent* way (only labelled segments are considered). To the contrary, in Section 4.1 we argue for an interplay, but leave its concrete account open: one can choose between a separated evaluation of annotation and segmentation, or one can evaluate both in a combined way.

References

1. Bunt, H., Alexandersson, J., Carletta, J., Choe, J.W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., Traum, D.: Towards an iso standard for dialogue act annotation. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (5 2010)
2. Carmines, E.G., Zeller, R.A.: Reliability and Validity Asessment. Quantitative Applications in the Social Sciences, SAGE, Beverly Hills ; London (1979)
3. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46 (1960)
4. Holle, H., Rein, R.: Assessing interrater agreement of movement annotations. In: Lausberg, H. (ed.) The Neuropsychological Gesture Coding System. Peter Lang Verlag, Bern (forthcoming)
5. Kendon, A.: Some relationships between body motion and speech. an analysis of an example. In: Siegman, A.W., Pope, B. (eds.) Studies in Dyadic Communication, chap. 9, pp. 177–210. Pergamon Press, Elmsford, NY (1972)
6. Kipp, M.: Multimedia annotation, querying and analysis in ANVIL. In: Maybury, M. (ed.) Multimedia Information Extraction, chap. 19. IEEE Computer Society Press (2010)
7. Kita, S., van Gijn, I., van der Hulst, H.: Movement phases in signs and co-speech gestures, and their transcription by human coders. In: Wachsmuth, I., Fröhlich, M. (eds.) Gesture and Sign Language in Human-Computer Interaction, pp. 23–25. Springer, Berlin/Heidelberg (1998)

8. Lücking, A., Bergmann, K., Hahn, F., Kopp, S., Rieser, H.: The Bielefeld speech and gesture alignment corpus (SaGA). In: Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality. pp. 92–98. 7th International Conference for Language Resources and Evaluation (LREC 2010), Malta (5 2010)
9. Müller, C.: Redebegleitende Gesten. Kulturgeschichte – Theorie – Sprachvergleich, Körper – Kultur – Kommunikation, vol. 1. Berlin Verlag, Berlin (1998)
10. Stegmann, J., Lücking, A.: Assessing reliability on annotations (1): Theoretical considerations. Tech. Rep. 2, SFB 360 *Situated Artiicial Communicators*, Universität Bielefeld (2005)
11. Thomann, B.: Oberservation and judgment in psychology: Assessing agreement among markings of behavioral events. BRM 33(3), 339–248 (2001)

Appendix

The Holle & Rein procedure is worked through by means of a synthetic example, that shows that binarisation of annotation categories can distort the degree of agreement between raters. The synthetic data is presented in Table 1(a). Two of three possible binarisations are given in Tables 1(b) and 1(c). The corresponding contingency tables are given in Table 2.

The Kappa statistic is calculated according to the following formula:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

$P(A)$ is the proportion of superficial agreement, that is, the count of all items for which the raters have chosen the same annotation value. $P(E)$ is the chance estimator. For the Kappa statistics, the chance estimator is derived from raters' biases: $P(E)$ is the product of the sum of raters' marginal frequencies n_{i+} and n_{+1} as obtained by cross-tabulating the data, see formula (2) and the schema of a contingency table in Figure 4.

		R1		
		1	2	Total
R2	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n

Fig. 4: Schema of a contingency table for two annotators R1 and R2 and two response categories 1 and 2.

$$P(E) = \sum_{i=1}^{k} \frac{n_{i+}}{n} \cdot \frac{n_{+i}}{n} \tag{2}$$

The calculation of the Kappa statistic of the "original data" from Tables 1(a) and 2(a) is given in equation (3):

$$\kappa = \frac{0.6 - 0.34}{1 - 0.34} = \frac{0.26}{0.66} \approx 0.4 \tag{3}$$

where $P(E)$ is obtained as follows:

$$P(E) = 0.5 \cdot 0.3 + 0.4 \cdot 0.4 + 0.1 \cdot 0.3 = 0.34 \tag{4}$$

Likewise, the calculation for the first binarisation of the Kappa coefficient is given in (5):

$$\kappa = \frac{0.8 - 0.455}{1 - 0.455} = \frac{0.345}{0.545} \approx 0.633 \tag{5}$$

Item	R1	R2
1	A	A
2	A	A
3	A	A
4	A	B
5	A	B
6	B	B
7	B	B
8	B	C
9	B	C
10	C	C

(a) "Original" data

Item	R1	R2
1	a	a
2	a	a
3	a	a
4	a	b
5	a	b
6	b	b
7	b	b
8	b	b
9	b	b
10	b	b

(b) First binarisation: A $\mapsto$ a and B,C $\mapsto$ b

Item	R1	R2
1	a	a
2	a	a
3	a	a
4	a	a
5	a	a
6	a	a
7	a	a
8	a	b
9	a	b
10	b	b

(c) Second binarisation: A,B $\mapsto$ a and C $\mapsto$ b

Table 1: Outcome of a synthetic annotation session of two raters R1 and R2 who classified 10 items according to 3 response categories A, B, C.

		R1			
		A	B	C	Ttl.
R2	A	3			3
	B	2	2		4
	C		2	1	3
	Ttl.	5	4	1	10

(a) Contingency table for "original data"

		R1		
		a	b	Ttl.
R2	a	3		3
	b	2	5	7
	Ttl.	5	5	10

(b) Contingency table for first binarisation

		R1		
		a	b	Ttl.
R2	a	7		7
	b	2	1	3
	Ttl.	9	1	10

(c) Contingency table for second binarisation

Table 2: Contingency tables

and

$$P(E) = 0.5 \cdot 0.3 + 0.5 \cdot 0.7 = 0.455 \tag{6}$$

For the second binarisation we receive the following coefficient (7):

$$\kappa = \frac{0.8 - 0.66}{1 - 0.66} = \frac{0.14}{0.34} \approx 0.412 \tag{7}$$

and

$$P(E) = 0.9 \cdot 0.7 + 0.1 \cdot 0.3 = 0.66 \tag{8}$$

As can easily be verified by the reader, the Kappa value increases in case of binarisations (although just slightly for the second one). This means, that the binarisation method used by H&R may lead to systematic overestimation of the degree of agreement between two raters' segmentations.