Staccato: Segmentation Agreement Calculator according to Thomann

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

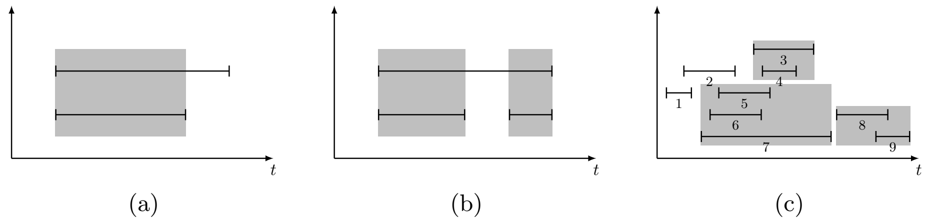
[1] Goethe-University Frankfurt am Main
[2] Bielefeld University, CRC 673, B1
luecking@em.uni-frankfurt.de
{sebastian.ptock,kirsten.bergmann}@uni-bielefeld.de

No matter what your favorite definition of co-verbal "hand and arm gesture" is, at the basic kinematic level a hand and arm gesture is in any case manifested by some spatio-temporal body movement. Accordingly, a prerequisite of any empirical, data-driven account to gesture is to locate those segments in the stream of hand and arm movements that make up a co-verbal hand and arm gesture. Following [3], a gesture basically has a tripartite movement profile, that consists of a preparation, a stroke, and a retraction phase. In addition to these three basic phases, temporary cessations of motion might occur either before or after the stroke might occur in terms of pre- and post-stroke holds [5]. In the annotation of video data we have to deal with the *segmentation problem*, that is, the demarcation of the temporal arm and/or hand movements that constitute gestures.

Like all annotations, gesture segmentation has to be evaluated with regard to its reliability (see [1]for an introduction into this topic). The standard method for gauging reliability of annotations are chance-corrected assessments of the agreement between multiple annotations of the same material – see [6] for the set-up of a careful evaluation of detailed gesture annotations. However, the reliability of gesture segmentation cannot be assessed by these methods. For standard agreement measures – like the wide-spread kappa statistic [2] – are applicable only to annotation data gained in the a test design that consists in classifying a fixed set of given items (gestures, in our case) into a predefined response categories. The demarcation of movement segments, by contrast, first of all determines *what the items are* (which then have to be classified according to response categories). Segmentation, therefore, is a precondition for item-based annotations. But how, then, is the reliability of gesture segmentation to be evaluated?

Procedures proposed in this regard 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 proposals of [4] and, at least partly, [7].

Such frequentist analyzes, however, fail to capture higher order structures like the number or the allocation of marked segments. The core of the assessment problem becomes obvious by the following examples, illustrated in subfigures (a) and (b) of Figure 1. In each of the subfigures segmentations of two annotators are displayed as horizontal lines, indicating the length of the respective segmentation according to the temporal x-axis. In subfigure (a) the annotators agree on the



(a) (b) (c)

Fig. 1: Configurations of segmentations. (a) and (b) are undecidable in frequentist terms. (c) illustrates *nuclei* (partially reproduced after Figure 1 of [8, p. 341]).

occurrence of a single gesture 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 kept separate an 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, assuming that both pairs of segmentations show the same amount of overlapping – as is the case in illustrations (a) and (b) – we would nonetheless expect that an assessment of the reliability of gesture segmentation account for this kinds of deviation properly and assigns a higher degree of agreement in case of (a) than in case of (b). Frequentist metric measurement, however, is not able to tell (a) and (b) apart, since both pairs of segmentations show the same amount of overlapping (see the gray area in the subfigures). What a segmentation assessment has to account for is 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 [8].[3] Instead of simple frequentist measures, Thomann utilizes graph-theoretical techniques that take structural features like number of items and allocations into account. The rationale of the Thomann method is illustrated in subfigure 1(c). Each row of segments has been produced by a different annotator, that is, 1(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 organization – the operationalization of agreement – is derived. A nucleus is defined in terms of internal homogeneity and external heterogeneity and indicated by gray boxes in subfigure 1(c). The first condition simply requires the segments in a nucleus to mutually overlap. According to this requirement, 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

[3] We would like to thank Timo Sowa for pointing us at the work of Thomann.

overlapping relations of all segments of a nucleus to be indistinguishable. As we can see in (c), segment 2 overlaps with segments 5, 6 and 7, but not with segments 3 and 4. Thus, external relations of 3 and 4 on the one hand and 5 to 7 on the other hand are distinguishable [8, p. 343]. Applying both conditions yields the nuclei depicted in the illustration. 7 out of 9 segments are organized 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 due to chance. To this end, the absolute degree of agreement is normalized against a random baseline. The random baseline constitutes the reference value for nuclei formation that are expected by chance alone. The resulting values of this normalization is the *degree of organization* [8, p. 343] that lies in the interval $(-1, 1)$. A value of 0 means that the empirically found number of segment nuclei equal the number in random configurations. Note that the degree of organization makes different configurations of segmentations (say, of various studies) comparable.

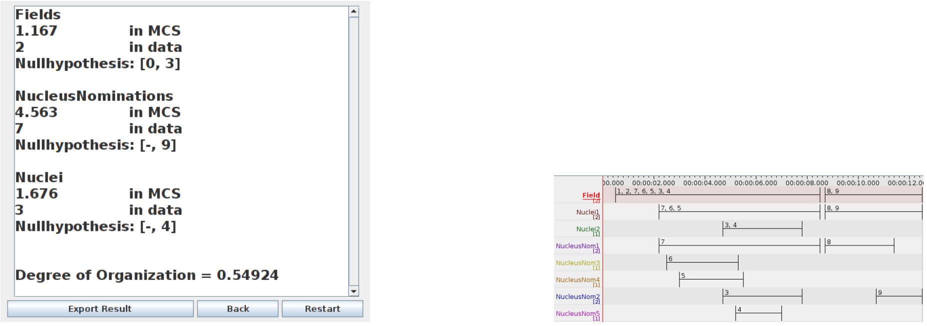
Needless to say that the determination of nuclei is too painstaking to do it by hand. Though there is an algorithm implemented by Thomann himself, no publicly available tool for calculating agreement of segmentation is at disposal. We offer such a software tool, (somewhat willfully) called *Staccato* (***S**egmen**t**ation **A**greement **C**al**c**ulator **a**ccording to **Tho**mann*), written in platform-independent Java. This stand-alone software will become a component of the standard multimodal annotation tools Elan[4] and Anvil[5].

Example Calculation To illustrate the usage of Staccato, let's assume the situation that several annotators finished coding specified events (an adoption from the first example in Fig. 1). To analyze 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: (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 nullhypothesis.

The result of running the agreement calculation with 10000 MCS, a granularity of 10 for annotation length, and *alpha*=0.05 is given in Fig. 2a. The *Degree of Organization* of 0.54924 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 Fig. 2b). The result not only comes with the Degree of Oraganization but also with *Fields*, *NucleusNominations* and *Nuclei*. Fields are subsets where all annotations are connected via overlappings. NucleusNominations are nominations that potentially might form a nucleus (see [8, p.44] for a definition). Fig. 2a also shows data and MCS-outcomes for each of the abovementioned as well as nullhypotheses according to MCS.

[4] www.lat-mpi.eu/tools/elan

[5] http://www.anvil-software.de/



```
Fields
1.167          in MCS
2              in data
Nullhypothesis: [0, 3]

NucleusNominations
4.563          in MCS
7              in data
Nullhypothesis: [-, 9]

Nuclei
1.676          in MCS
3              in data
Nullhypothesis: [-, 4]


Degree of Organization = 0.54924
```

(a)

(b)

Fig. 2: Result of the Thomann method in Staccato (a), and output exported from Staccato (b): 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.

Acknowledgements This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication".

References

1. Carmines, E.G., Zeller, R.A.: Reliability and Validity Asessment. Quantitative Applications in the Social Sciences, SAGE, Beverly Hills ; London (1979)
2. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46 (1960)
3. 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)
4. Kipp, M.: Multimedia annotation, querying and analysis in ANVIL. In: Maybury, M. (ed.) Multimedia Information Extraction, chap. 19. IEEE Computer Society Press (2010)
5. 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)
6. 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)
7. Rein, R., Holle, H.: Assessing interrater agreement of movement annotations: A work in progress. In: Gesture: Evolution, Brain, and Linguistic Structures. p. 127. 4th Conference of the International Society for Gesture Studies (ISGS), Europa Universität Viadrina Frankfurt/Oder (7 2010)
8. Thomann, B.: Oberservation and judgment in psychology: Assessing agreement among markings of behavioral events. BRM 33(3), 339–248 (2001)