EasyDIAg: A Tool for Easy Determination of Interrater Agreement

Henning Holle[1] & Robert Rein[2]

[1] Department of Psychology, University of Hull, U.K.

[2] Institute of Health Promotion and Clinical Movement Science, Department of Neurology, Psychosomatic medicine, and Psychiatry, German Sport University Cologne, Germany

Manuscript accepted for publication in: *Behavior Research Methods*

Correspondence should be addressed to Henning Holle, h.holle@hull.ac.uk, Phone +44-(0)1482 – 466152

Acknowledgments: We are grateful to Hedda Lausberg for sharing the annotation data. The acquisition of this data set was supported by a grant awarded to Hedda Lausberg from the DFG (LA 1249/2-1). The EasyDIAg toolbox described in the paper can be downloaded free of charge from http://sourceforge.net/projects/easydiag/.

Abstract

Reliable measurements are fundamental for the empirical sciences. In observational research measurements often consist of observers categorizing behavior into nominal-scaled units. Since the categorization is the outcome of a complex judgment process, it is important to evaluate the extent to which these judgments are reproducible, by having multiple observers independently rate the same behavior. A challenge in determining inter-rater agreement for timed-event sequential data is to develop clear objective criteria to determine whether two rater's judgments relate to the same event (the linking problem). Furthermore, many studies currently only report raw agreement indices, without considering the degree to which agreement can occur by chance alone. Here, we present a novel, free and open-source toolbox (EasyDIAg) designed to assist researchers with the linking problem, as well as providing chance-corrected estimates of inter-rater agreement. Additional tools are included to facilitate the development of coding schemes and rater training.

Introduction

Obtaining precise measurements is fundamental to the empirical sciences. In observational research, measurements often consist of observers categorizing behavior into units. Because the categorization is the outcome of complex judgment processes, it is important to determine the extent to which these categorizations are reproducible (Cohen, 1960). Without such demonstrations of inter-observer agreement, we are left with individual narratives of unknown reliability (Bakeman & Quera, 2011). For example, in the field of gesture research, determining inter-observer agreement typically involves at least two raters independently assessing video recordings of a gesturing person. Reliability in this context is demonstrated when one video coded independently by two raters produces essentially identical results.

Provided the stream of behavior can be segmented into meaningful units based on an objective external criterion, standard methods of determining inter-observer reliability can be applied. For instance, one could segment a recorded video into 10-second chunks, followed by raters individually coding each 10-second chunk for the type of behavior observed. Each unit can then be coded as belonging to one of *K* different categories, depending on the research question. This type of data has been termed event sequential data (Bakeman & Quera, 1992). The *K* different categories should be mutually exclusive and exhaustive, which means that categories should not conceptually overlap with each other and that all categories taken together should describe all possible types of behavior (Cohen, 1960).

However, spontaneous behavior often cannot be segmented meaningfully into chunks of equal length, because (i) the time between multiples instances of spontaneous behavior is often variable, (ii) the duration of the units of spontaneous behavior varies, and (iii) the units of interest (e.g., gestures) tend to be intermixed with other behavioral phenomena that are outside the researcher’s interest (e.g., self-touches). Thus, categorizing spontaneous behavior often involves making decisions on two levels (Bavelas, Kenwood, & Phillips, 2002). First, the units about which coding decisions are to be made need to be identified in the video stream. This could involve, for instance, determining the onset and offset of each episode of submissive behavior in a chimpanzee (Kaufman & Rosenthal, 2009). Only after this initial segmentation can raters make more specific decisions about what kind of behavior each unit represents (e.g., whether the submissive behavior consisted of the type ‘crouch’ or ‘present’). Thus, segmentation is part of the decision making process and can contribute to rater disagreement. Any procedure that aims to quantify the amount of inter-observer agreement for such so-called *timed*-event sequential data (Bakeman & Quera, 2011; Bakeman, Quera, & Gnisci, 2009) needs to take segmentation into account, either by providing separate indices for segmentation and categorization agreement (Bavelas, Gerwing, Sutton, & Prevost, 2008), or by means of an overall agreement index that jointly considers segmentation and categorization (Holle & Rein, 2013).

Including segmentation when determining inter-observer agreement for timed-event sequential data is therefore desirable, but it also creates challenges. When observers make coding decisions about pre-segmented units, the only kind of error that can occur is that of disagreement. However, when observers independently segment the continuous video stream into units before coding, there can also be errors of commission and omission (Bakeman & Quera, 2011). An error of commission by observer A with respect to observer B occurs when A identifies an event that B did not. Vice versa, an error of omission by observer A with respect to observer B occurs when A fails to identify an event that B does. In the absence of a gold standard, it is of course an arbitrary decision whether to consider errors of this type as commission or omission. Furthermore, raters may qualitatively agree in detecting a unit, but differ in the onset and offset times of that unit (Rein, 2013). Thus, the following questions need be addressed by any new procedure that aims to take segmentation into account when determining inter-rater agreement: First, on which criteria should units be linked when evaluating inter-rater agreement? Second, how much misalignment should be tolerated?

Deciding which criteria should be used to link units is less trivial than one might think. Figure 1 provides an illustration of the linking problem. Here, Rater 1 has identified one long event, whereas Rater 2 has identified two shorter events. There are at least three different ways in which these events could be linked. First, the short unit seen by Rater 2 could be linked with the long unit of Rater 1, based on their very similar onset time, whereas the second unit of Rater 2 would remain unlinked. A second possibility might be to link the long unit of Rater 2 with the unit seen by Rater 1 (based on their substantial overlap), leaving the first shorter unit of Rater 2 as unlinked. Finally, one could allow multiple linking so that a unit from one rater can be linked to multiple units from a second rater.

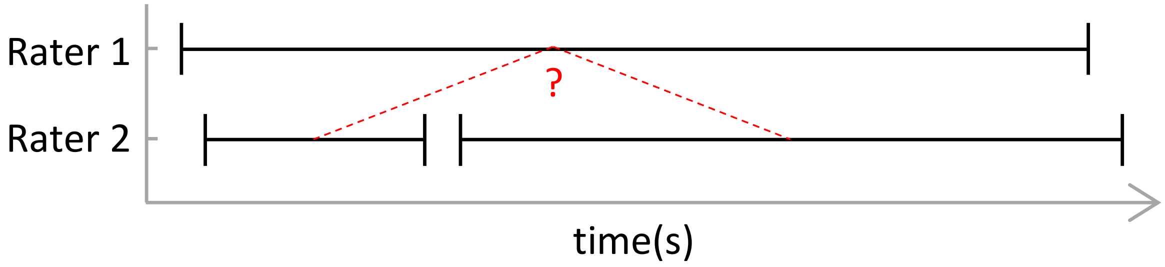


Figure 1 Illustration of the linking problem. Here, two raters have independently identified onset and offset of events of interest.

One motivation for developing our toolbox was to develop an algorithm to tackle this linking problem. Relying on an algorithm when deciding whether two observers' decisions relate to the same event, rather than visual inspection, should enable more consistency in determining inter-observer reliability across studies. A second motivation was to provide researchers with a simple way to obtain chance-corrected agreement indices for timed-event sequential data. When we reviewed a random sample of 32 research papers from the domain of gesture production, three studies did not address inter-observer reliability in their published text. Fifteen studies provided only raw agreement indices. However, raw agreement does not take into account how much the raters agree by chance alone. For instance, when two raters are coding whether units are of either type 'A' or 'B', and 'A' is much more frequent than 'B' (e.g., 90% vs. 10%), then chance agreement is $0.9^2 + 0.1^2 = 0.81$. Thus, a raw agreement of 81% in this case does not indicate good inter-observer reliability, but simply how much raters can be expected to agree by chance alone. In contrast, a chance-corrected agreement index, such as Cohen's kappa, normalizes the observed agreement by the amount that could be expected by chance alone (Cohen, 1960). A recent review in the field of observational research on animal behavior has found that out of 100 published studies, 96 did not address inter-observer reliability in their published text (Kaufman & Rosenthal, 2009), further underlining the need for a tool to facilitate the determination of inter-rater agreement.

There is some previous work on the problem of determining inter-rater agreement for timed-event sequential data and the linking problem. One possible solution has been described by

Haccou and Meelis (1994). For their algorithm, the authors set up a cascade of matching rules. Based on these rules, every annotation from two raters is matched during several data passes. The output is an agreement table of size *K* by *K* where every annotation is matched. As has been pointed out by Bakeman et al. (2009), the rules used to match annotations in the Haccou-and-Meelis algorithm result in some annotations from one rater being matched with more than one annotation from the other rater. Such multiple linking leads to more events showing up in the agreement table than actually have been coded by the raters, potentially biasing the results. A different algorithm for computing observer agreement was later developed by Jansen and co-workers (2003) and implemented in The Observer (www.noldus.com). This algorithm allows the user to specify a tolerance window within which annotations are time-shifted with respect to each other during the matching process. This approach therefore addresses the problem of slight onset and offset differences between raters. Despite this improvement, that algorithm still suffers from the same multiple linking problem as the original Haccou-and-Meelis algorithm (Bakeman et al., 2009). Furthermore, both algorithms do not allow for errors of commission or omission. The INTERACT program (www.mangold.de) includes an algorithm that allows for errors of omission and commission, by including a final processing step where any remaining unlinked events of a rater are linked to a *nil* event of the other rater. However, because multiple annotations can still be matched to a single annotation from the other rater, the results are potentially biased towards overestimating agreement (Bakeman et al., 2009).

A different approach has been taken by Bakeman and colleagues (Bakeman & Quera, 2011; Bakeman et al., 2009; Quera, Bakeman, & Gnisci, 2007). In Quera et al. (2007), an algorithm was introduced that uses a sequence alignment algorithm originally developed to classify DNA sequences. The algorithm obtains an optimal sequence alignment and allows errors of omission and commission. However, time is not explicitly taken into account but only the sequence of the events (see also Dijkstra & Taris, 1995). Thus, much useful information readily accessible in the data is not being used by the algorithm. Finally, Bakeman et al. (2009) developed a new approach based on dynamic programming algorithms to explicitly deal with timed-event sequences. The algorithm remains based on sequence information but allows the user to specify parameters to account for variations of event onset and overlap. However, given that in many fields of research (including gesture research), different categories tend to have very different event lengths (see also Figure 3), we considered such algorithms with hard-coded temporal offset parameters as unsuitable for our problem.

Taken together, the current approaches with respect to calculating inter-rater agreement for timed-event sequential data in general and gesture annotations in particular are somewhat unsatisfying, because they either come with hard-coded temporal offset parameters or unnecessarily inflate the number of tallies in the agreement table because of multiple linking. Below, we describe a novel approach that allows a more accurate determination of inter-rater agreement. Our approach takes into account all decisions during segmentation and coding of video material. It is based on a simple cascading rule scheme, similar to those applied by Haccou and Meelis (1994), but at the same time avoids the problem of multiple linking. The algorithm is implemented in MATLAB as an open-source toolbox called EasyDIAg (**Easy D**etermination of **I**nterrater **Ag**reement). The EasyDIAg toolbox can be downloaded free of charge (http://sourceforge.net/projects/easydiag/). The algorithm takes as input annotation data from two raters that have independently segmented and coded one or more videos, using a coding system consisting of *K* mutually exclusive and exhaustive categories (Cohen, 1960). The output is a (*K*+1) by (*K*+1) agreement table, as well as supplementary agreement indices. By jointly considering both temporal overlap and agreement of classification labels, the algorithm provides an estimate of inter-rater agreement. The algorithm is also symmetric in the sense that it does not matter which of two raters is labeled as rater 1 or rater 2. To enable more effective training of raters or to identify specific problems in inter-rater agreement, additional functions are included in the toolbox. These functions help to identify those movie clips and coding categories that are the biggest sources of disagreement between raters.

Assumptions

The EasyDIAg toolbox assumes that two raters have independently annotated one or more video clips[1]. Each segment of interest has a defined onset and offset time as well as a category label. The categories should be nominally scaled, mutually exclusive and exhaustive (Cohen, 1960).

[1] It is of course also possible to analyze annotations of audio clips.

We recommend using ELAN for the annotation process, which is freely available software (http://tla.mpi.nl/tools/tla-tools/elan/). Users can choose between downloading the toolbox as MATLAB code (if they have a working installation of MATLAB), or as a stand-alone application (for users that do not have a MATLAB license). Further technical information can be found in the accompanying documentation of the toolbox.

Description of the matching algorithm

The aim of the matching algorithm is to generate an agreement table, based on timed-event sequential rating data from two raters (for an example data set, see Figure 2). The resulting agreement table will have as many rows and columns as there are categories in the coding system, plus one additional column/row for commission/omission errors (*nil*, see Bakeman & Quera, 2011, as well as Table 1). Agreements are tallied on the main diagonal of the table from the top left to bottom right. Disagreements are tallied on the respective off-diagonal cells. Since observers usually do not code the absence of a behaviour (e.g., *not* having seen a gesture of type x), the combination *nil-nil* cannot logically occur in the absence of a gold standard (i.e., it is a structural zero, see Bakeman & Quera, 2011).

Step 1: Creating an overlap matrix

As a first step, the algorithm indexes all annotations for each rater individually in their temporal order (see Figure 2). This results in two sets, one for each rater, indexing each annotation together with their respective onset- and offset-times.

$$R_1 := \{a_{11}, a_{12}, ..., a_{1n_1}\}$$
$$R_2 := \{a_{21}, a_{22}, ..., a_{2n_2}\}$$
$$a_{ij} := (onset_{ij}, offset_{ij}, k_{ij})$$

n_1 := number of annotations made by Rater 1, n_2 := number of annotations made by Rater 2, k_{ij} ε 1, .., K.

Next, the algorithm evaluates which annotations are temporally overlapping. For each annotation, it checks whether there is an overlapping annotation from the other rater. If there is an overlap, the percentage of temporal overlap is calculated according to the following equation:

$$\%overlap_{ij} := \frac{\min(offset_{1i}, offset_{2j}) - \max(onset_{1i}, onset_{2j})}{\max((offset_{1i} - onset_{1i}), (offset_{2i} - onset_{2i}))}$$

i := annotation from Rater 1, j := annotation from Rater 2

Thus, the $\%_{\text{overlap}}$ between two annotations is calculated relative to the length of the longer annotation. The $\%_{\text{overlap}}$ values are written to an n_1 by n_2 overlap matrix. This matrix is conceptually similar to a distance matrix used during a cluster analysis (Kaufmann & Rousseeuw, 1990). Given that most of the annotations will overlap with only a few other annotations, the overlap matrix will contain mainly zeros and can therefore be implemented using sparse matrix algorithms. Rater-specific annotations that do not have any temporal overlap with annotations from the other rater are stored separately.

Step 2: Transferring values from the overlap matrix to the agreement table

Subsequently, the $\%_{\text{overlap}}$ values are sorted. Starting with the two annotations that yield the greatest $\%_{\text{overlap}}$, the algorithm checks whether the $\%_{\text{overlap}}$ is greater than the user-specified overlap threshold. If yes, the two annotations are linked (see red connecting lines in Figure 2) and are henceforth treated as one unit. This unit is then tallied in the respective cell of the agreement table. It is tallied as an agreement on the main diagonal if the category labels are identical, and tallied as a disagreement in the respective off-diagonal cell if the category labels are not identical. At the same time, the two annotations constituting the unit are removed from the original overlap matrix, to avoid them being counted multiple times.

If the $\%_{\text{overlap}}$ is below the user-specified threshold, the two involved annotations are not linked, but are tallied independently as commission/omission errors in the respective *nil*-column and *nil*-row on the agreement table. The algorithm then proceeds with the next-largest $\%_{\text{overlap}}$ value from the overlap matrix, until all overlapping units have been tallied to the agreement table.

Finally, isolated annotations that have no temporal overlap are tallied as commission/omission errors to the respective *nil*-column or *nil*-row.

After termination of the tallying step, an agreement table with *K*+1 rows and *K*+1 columns is obtained (for an example, see Table 1). Please note that the entry *nil-nil* represents a structural zero, since it cannot occur empirically. Accordingly, Cohen's kappa cannot be calculated using the standard formula (Bakeman & Robinson, 1994) in this case. Instead, the EasyDIAg toolbox uses an iterative proportional fitting algorithm (Deming & Stephan, 1940) to obtain a kappa value from the global agreement table (see also Bakeman & Quera, 2011).

Once the global agreement table is available, separate agreement statistics are calculated for each annotation type k ε 1, .., *K*, using a 2x2 agreement table (for an example, see Table 2). To this end, the number of agreements for type k is written into the top-left corner, the sum of all false positive errors for type k of rater 1 (column k excluding entry$_{k,k}$) is written into the bottom-left corner, the sum of all false positives for type k of rater 2 (row k excluding entry$_{k,k}$) is written into the top-right corner, and the sum over all remaining cells is written into the bottom-right corner. This is performed for all *K* different types such that *K* 2x2 agreement tables are obtained. From these agreement tables standard Cohen's kappa statistics are calculated including κ, κ_{max} (Cohen, 1960), and positive agreement p_{pos} (Cicchetti & Feinstein, 1990).

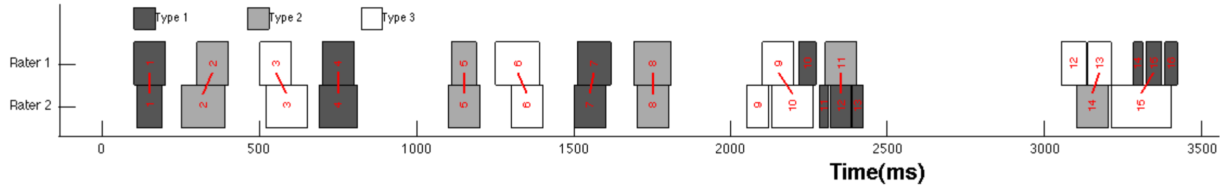


Figure 2 **Example of a timed-event sequential data set. Two raters independently categorized behavior into three types (1, 2 or 3). Annotations that are linked by the algorithm based on the overlap criterion are connected with a red line.**

Empirical validation

To validate the toolbox empirically, we used annotation data provided by Hedda Lausberg. In this study, a total of 44 participants were videotaped during the retelling of 'Mr. Bean' video clips. The total length of annotated video material was 53 min, with a mean video length of 72.40

s (SD=33.55). Two certified raters trained in the NEUROGES-ELAN coding system (Lausberg, 2013; Lausberg & Sloetjes, 2009) annotated the observed movement types, according to the six categories defined in Module 1 (Lausberg & Sloetjes, 2009, see also Footnote 2). The two raters placed a combined total of 1964 annotations, with at least 100 tags for each of the six categories. As can be seen in Figure 2, both the length and frequency of annotations vary considerably between categories. This allowed us to see whether such variations systematically influence the agreement indices provided by the algorithm and how the proposed algorithm deals with actual, rather than simulated, annotation data. The toolbox analyses data from each file separately and subsequently combines the agreements tables from all files to obtain a single agreement table across all data. This table is used to calculate the different kappa measures.

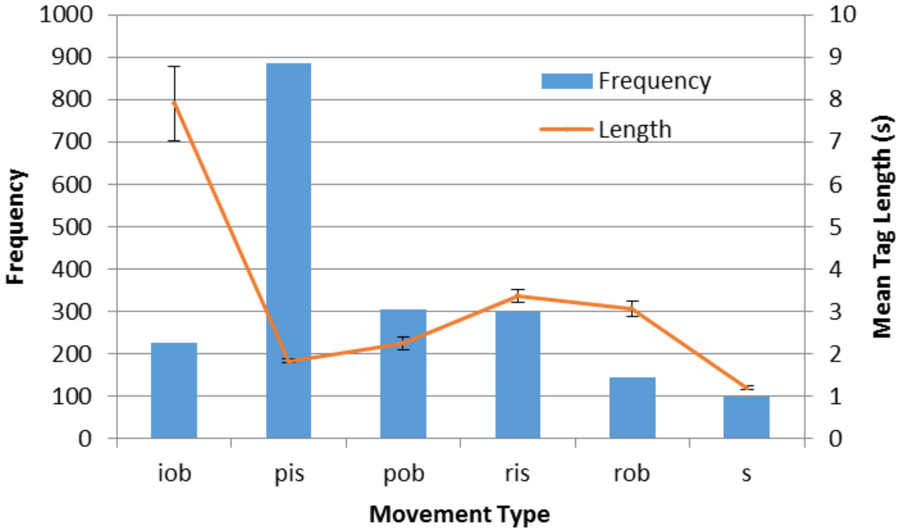


Figure 3 **Frequency and mean length (± *SD*) of annotations, for each of the six categories. Categories are defined in detail in (Lausberg & Sloetjes, 2009). Abbreviations: iob=irregular-on-body, pis=phasic-in-space, pob=phasic-on-body, ris=repetitive-in-space, rob=repetitive-on-body, s=shift.** [2]

[2] Please note the labels (but not the underlying concepts) of 3 of the 6 movement categories shown here has changed since the publication of the original NEUROGES-ELAN system (Lausberg & Sloetjes, 2009). The old terms 'continuous-on-body', 'phasic distant', and 'repetitive distant' (as used in Lausberg & Sloetjes, 2009) have been since been renamed to 'irregular-on-body', 'phasic-in-space', and 'repetitive-in-space', respectively. Another change that has occurred since the time the annotation took place in 2008 is that in the most recent update of NEUROGES-ELAN (Lausberg, 2013), the six categories mentioned above are now considered as distinct values within a superordinate category called 'StructureFocus'.

Overall agreement table and indices

The primary output data of the toolbox is an agreement table containing all categories of the coding system (plus the additional *nil* column/row mentioned above). Table 1 shows such an agreement table for the six movement categories of the empirical example. There is one additional 7th *nil* column/row for errors of commission/omission (marked in blue). A lot of useful diagnostic information is already contained in this agreement table. On the main diagonal are the agreement tallies for each of the six categories (marked in green). In the off-diagonal cells are the disagreement tallies (marked in red). For instance, one can learn from the table that category ‘s’ (shift) was only confused with categories ‘pis’ (phasic-in-space) and ‘pob’ (phasic-on-body), but not with the other three movement categories. Such information about which categories are confusable and which ones are not can be used to sharpen diagnostic criteria within a coding system.

Table 1 Agreement table for all six movement types, as determined by EasyDIAg, including totals and marginal probabilities p. The $\%_{\text{overlap}}$ parameter was set to 60%. Linked events can either be classified as agreement (shown in green) or disagreements (shown in red). Unlinked events are tallied as commission/omission errors in the respective *nil* column/row (shown in blue).

	Observer 2's codes								
Obs. 1’s codes	**iob**	**pis**	**pob**	**ris**	**rob**	**s**	**nil**	Total	*p*
iob	160	0	0	0	6	0	22	188	0.10
pis	0	684	20	22	6	6	80	818	0.42
pob	18	16	160	0	20	24	62	300	0.15
ris	2	18	2	220	8	0	21	271	0.14
rob	0	2	4	14	84	0	20	124	0.06
s	0	2	0	0	0	52	18	72	0.04
nil	32	76	31	29	10	13	–	191	0.10
Total	212	798	217	285	134	95	223	1964	1.00
p	0.11	0.41	0.11	0.15	0.07	0.05	0.11	1.00	0.69

The toolbox also provides several summary statistics for the overall amount of agreement between raters. Different researchers will be interested in different types of indices here. Some coding systems ask for composite agreement indices that reflect the sum of segmentation and categorization disagreements (e.g., Holle & Rein, 2013; Lausberg & Sloetjes, 2009). Users interested in such composite agreement indices should take note of the raw agreement (incl. *nil*), kappa (incl. *nil*) and kappa_max (incl. *nil*). For instance, the raw agreement (incl. *nil*) for Table 1 is the sum of all agreement tallies on the main diagonal (marked in green) divided by the total number of tallies (N=1964), and amounts to 69% in this case.

Other coding systems (e.g., Bavelas et al., 2008) ask the researcher to provide *separate* estimates for the amount to which raters agree in their segmentation and categorization, respectively. To obtain an estimate specifically for segmentation agreement, the toolbox sums the total amount of linked events, divided by the total number of tallies in the agreement table. In Table 1, this is equivalent to the sum of agreement and disagreement tallies (i.e., sum of all cells marked in green and red, but excluding the blue color-coded *nil* column and *nil* row, N=1550), divided by the overall sum of coding decisions (N=1964), giving an estimate of 79% for segmentation agreement. In other words, 21% of all units remain unlinked in the current analysis. To obtain indices that specifically reflect categorization agreement (and that exclude potential disagreements from segmentation being carried forward), the toolbox provides raw agreement, kappa and kappa_max (excl. *nil*) for linked units only. These pure categorization agreement indices are calculated using linked units only (i.e., the *nil* column and *nil* row are excluded here).

Category-specific agreement tables and indices

Researchers might additionally be interested in category-specific agreement tables (see Table 2). These tables are created by comparing agreement for the category of interest with collapsed data from all other categories. These 2 x 2 tables, which are also generated by our toolbox, allow not only computation of category-specific agreement indices (such as observed agreement and Cohen's kappa), but can also be used to detect rater bias (e.g., whether one rater is more likely to detect events of a certain type than another rater). For instance, in the present data set, one rater was much more biased towards coding events as 'pob' (phasic-on-body) than the other rater (see

Table 2). Such rater bias, which can also be statistically quantified by means of a McNemar test[3], may indicate a lack of rater training, or lack of clear diagnostic rules to distinguish between categories.

Table 2 Six 2x2 agreement tables, obtained by collapsing the large agreement table shown in Table 1

	iob	**other**	Total
iob	160	28	188
other	52	1724	1776
Total	212	1752	1964

	pis	**other**	Total
pis	684	134	818
other	114	1032	1146
Total	798	1166	1964

	pob	**other**	Total
pob	160	140	300
other	57	1607	1664
Total	217	1747	1964

	ris	**other**	Total
ris	220	51	271
other	36	996	1032
Total	285	1679	1964

	rob	**other**	Total
rob	84	40	124
other	50	1790	1840
Total	134	1830	1964

	s	**other**	Total
s	52	20	72
other	43	1849	1892
Total	95	1869	1964

In addition to providing category-specific agreement tables, the toolbox also provides agreement indices in the form of raw observed agreement, Cohen's kappa, maximum kappa, and positive agreement (see Figure 4). The raw agreement indices, which are often used as sole agreement measure in gesture studies, are uniformly high (between 87% and 97%, see Figure 4). However, raw agreement does not take into account how much the raters agree by chance alone. Whenever agreement for an infrequent category is considered, chance agreement (shown in gray in Figure 4) will inevitably be high, because the marginal probabilities will be very unequal. The less frequent a category, the more uneven will be the marginal probabilities in the 2 x 2 table, resulting in greater chance agreement. For instance, the 97% observed agreement for category 's' (shift) is much less impressive, when considering that chance agreement in this case is 92%. The inverse relationship between category frequency and chance agreement can also be appreciated by comparing annotation frequencies shown in Figure 2 with chance agreement indices shown in Figure 4. This relationship between category frequency and chance agreement highlights the

[3] A McNemar Test for 2x2 contingency tables is provided by a freely available Excel®-Worksheet (Mackinnon, 2000, available online at www.mhri.edu.au/biostats/DAG_Stat)

need to report chance-corrected indices of agreement (such as Cohen's Kappa), rather than raw measures of agreement.

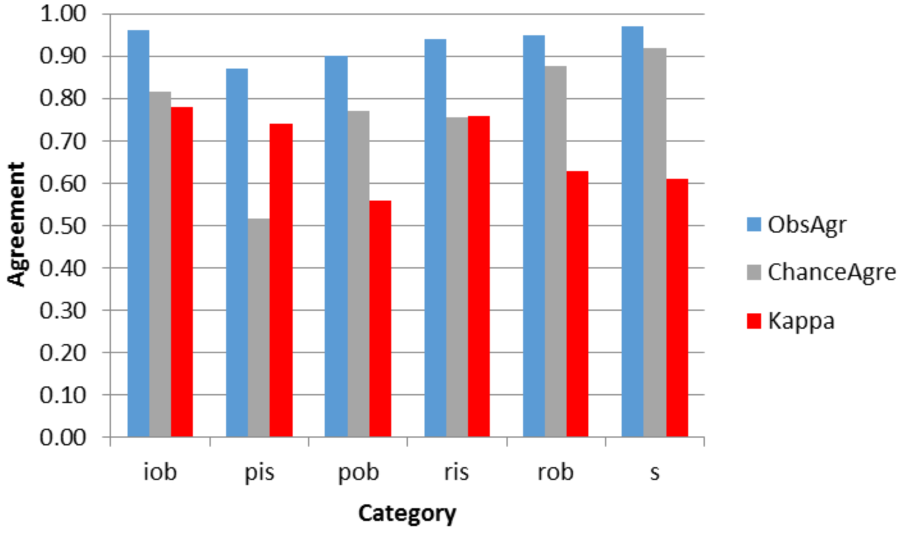


Figure 4 **Three agreement indices for the six categories: Observed Agreement (blue), Chance Agreement (gray) and Cohen's kappa (red).**

What is a sensible overlap threshold?

For the agreement tables described above, we used an overlap criterion of 60%. Of course, this can be criticized as an arbitrary setting, as any value between 51% and 100% could in principle have been used. In the toolbox, the user can specify values between 51 and 90%, with the default set to 60%. We further explored the extent to which this parameter setting influences the obtained kappa values, by plotting the amount to which kappa decreases as the overlap criterion is increased. As can be seen in Figure 5, there is a monotonic decrease in the observed kappa values, as more stringent overlap thresholds are applied to the annotation data. The sharpest decrease in agreement occurs for overlap values of 80% or above. As expected, kappa scores for categories that tend to have very long event durations (e.g., 'iob') are less affected by more stringent overlap criteria. In contrast, the drop in observed kappa scores as the overlap criterion is increased from 80% to 90% is most pronounced for the categories with the shortest event durations (i.e., categories 'pis' and 's').

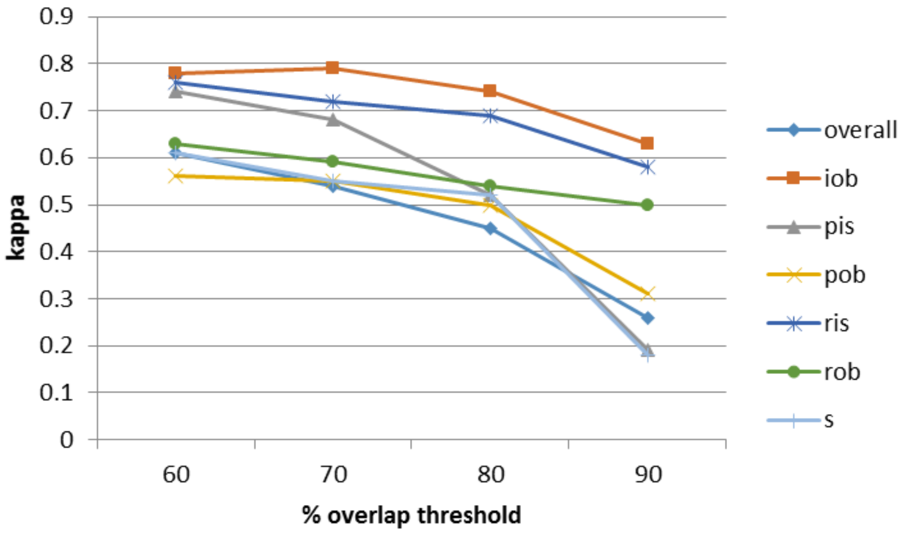


Figure 5 **Relationship between % overlap threshold and Cohen's kappa**

Given that kappa values are relatively stable across a range of overlap values we do not see a particular problem with assigning a default value of 60 percent. This value was chosen based on our personal experience with annotation gesture data. If researchers require a more conservative agreement estimate, they are free to increase the overlap threshold. In any case, researchers need to ensure that when reporting results of inter-rater agreement analyses carried out with EasyDIAg, the respective agreement indices are always accompanied by the chosen overlap threshold.

Comparison of EasyDIAg with GSEQ-DP

In order to further validate our algorithm, we compared EasyDIAg's performance in processing the above-mentioned data set with an alternative algorithm for timed-event sequential data, the Generalized Sequential Querier dynamic programming algorithm (GSEQ-DP), as implemented in GSEQ Version 5.1 (http://www2.gsu.edu/~psyrab/gseq/Download.html). GSEQ-DP has recently been found to compare favorable against other published algorithms in the field (Bakeman et al., 2009).

GSEQ-DP and EasyDIAg have in common that they both allow the user to calculate raw as well as chance-corrected agreement indices for timed-event sequential data. They differ with respect to their scope. Whereas GSEQ-DP provides not only event-based, but also time-unit based agreement indices and provides further data processing options for observational data, EasyDIAg has additional tools that should be helpful especially during the initial part of a behavioral research project (e.g., a rater disparity feature to identify which videos and categories are the biggest source of rater disagreement). The two algorithms also differ in the number of categories they provide in the agreement table, since GSEQ-DP by default includes an additional category representing the absence of an event.

To enable a fair like-for-like comparison, the overlap parameter was set to 60% in both algorithms. The Event Tolerance Parameter of GSEQ-DP was set to 0, since no equivalent exists in EasyDIAg. Milliseconds were specified as the time-unit for analysis in GSEQ-DP. Sessions where only one rater has placed annotations are automatically excluded from the analysis in GSEQ-DP. These sessions (containing 4 annotations in total, equivalent to 0.2% of the total data set) were therefore removed from the analysis file.

The overall agreement, as determined by the kappa value, was comparable to the value obtained with EasyDiag (0.59 for GSEQ-DP, 0.61 for EasyDIAg) suggesting that EasyDIAg is a valid tool for determining inter-rater agreement for timed-event sequential data. The full agreement table provided by GSEQ-DP can be found below (see Table 3).

Table 3 Overall agreement table for the empirical data set, as determined by GSEQ-DP (Version 5.1.17)

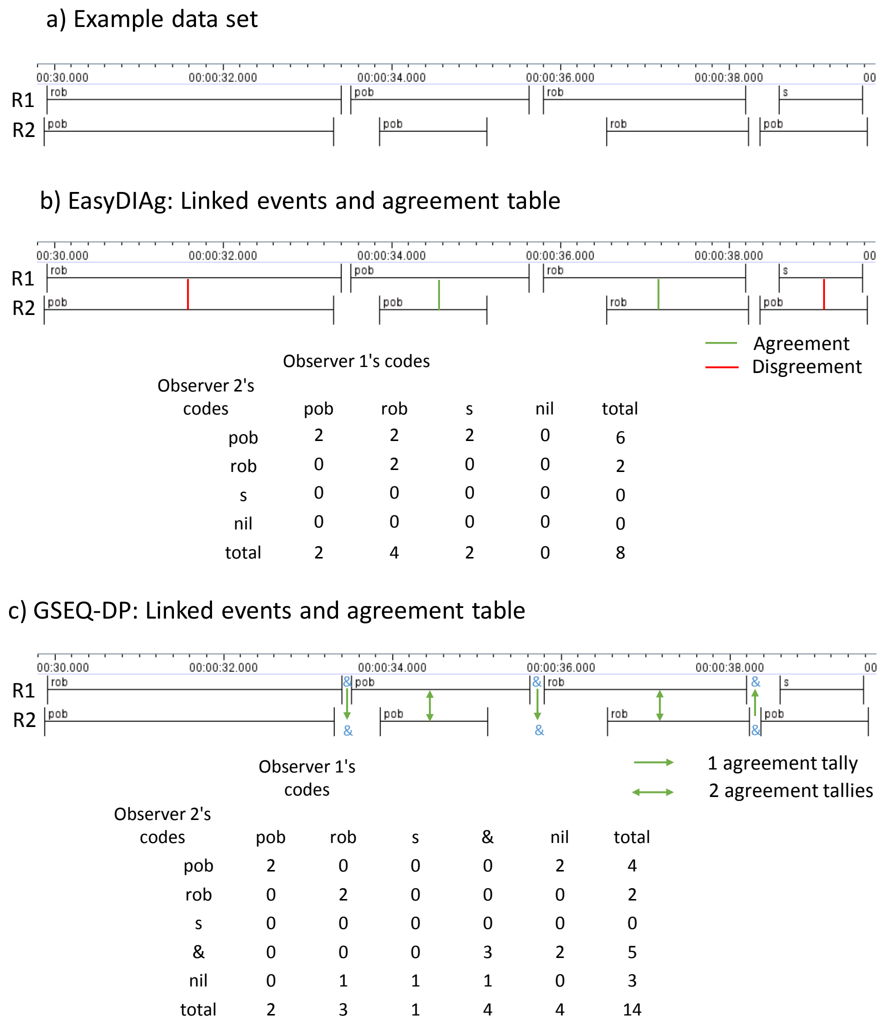
	iob	**pis**	**pob**	**ris**	**rob**	**s**	**nil**	**& (all others)**	**total**	**kappa**
iob	165	0	0	0	2	0	9	1	177	0.85
pis	1	696	0	1	0	0	93	4	795	0.86
pob	2	1	179	0	0	0	89	0	271	0.74
ris	0	0	0	231	0	0	26	0	257	0.86
rob	0	0	0	2	87	0	22	0	111	0.76
s	0	0	0	0	0	60	18	0	78	0.75
nil	37	69	24	40	26	21	0	224	441	
& (all others)	4	4	1	2	0	0	373	703	1087	0.57
totals	209	770	204	276	115	81	630	932	3217	

Overall agreement: Kappa = 0.59

One difference is that GSEQ-DP includes an additional category in the agreement table ('&, all others'). This is because the algorithm automatically fills in all gaps between a rater's annotations representing the absence of any of the defined categories (to be further explored below).

When comparing the two overall agreement tables, it becomes apparent that GSEQ-DP identifies very few disagreements between raters. If two units cannot be linked as agreement, GSEQ-DP tends to tally them individually as commission/omission errors to the respective *nil* row or column. Only 9 instances are reported where raters disagreed about a given category label. On the other hand, EasyDIAg has more disagreements in the off-diagonal cells. For instance, according to Table 1, category 'pob' was a significant source of disagreement between raters. Many instances of category 'pob', as seen by Rater 1, were categorized as either 's', 'rob' or 'iob' by Rater 2. No such disagreements were observed when we analyzed the same data with GSEQ-DP (see Table 3).

We decided to further investigate what might cause such different outcomes, by exploring the two algorithms' behavior in more detail for one example video file. As can be seen in Figure 6, each rater provided 4 annotations. Based on the overlap criterion, EasyDIAg links these 8 annotations into four pairs, two of which are tallied as agreements and two as disagreements. Figure 6c shows how the same example file is evaluated by GSEQ-DP. In a first step, the algorithm fills in the gaps between annotations with the ('&', all other) category. These gap events are inserted even if there is only a very small gap between two consecutive ratings (e.g., in between the first two annotations given by Rater 1). The algorithm then considers pairs of codes in turn, deciding on the basis of a cost matrix whether a code from one rater is linked with a code from the other rater and tallied accordingly as either an agreement or disagreement. Unlinked events are tallied as a commission/omission error in the respective *nil* column or row. Since GSEQ-DP aligns the sequences twice (R2 to R1 and R1 to R2), individual units can either become linked during both iterations (in which case two scores are tallied to the agreement table) or only during one iteration, causing only one agreement score being tallied (indicated by the single vs. double-headed arrows in Figure 6c). For instance, the first gap event ('&') is only linked when Rater 1's sequence is aligned to Rater 2 but not during the reverse iteration when Rater 2's sequence is aligned to Rater 1. In contrast, the second pair of events are classified as agreement for category 'pob' during both iterations (see Figure 6c). It is a strength of the GSEQ-DP algorithm that it achieves optimal alignment between two sequences, in the sense that it provides an alignment that requires the minimum number of possible transformations and yields both the most matches and the lowest distance (Bakeman et al., 2009). However, due to its incremental and dynamic nature, it is sometimes difficult to predict a-priori whether two events that visually look like a match will be linked during both iterations or only during a single iteration.



b) EasyDIAg: Linked events and agreement table

Observer 2's codes	Observer 1's codes: pob	rob	s	nil	total
pob	2	2	2	0	6
rob	0	2	0	0	2
s	0	0	0	0	0
nil	0	0	0	0	0
total	2	4	2	0	8

c) GSEQ-DP: Linked events and agreement table

Observer 2's codes	Observer 1's codes: pob	rob	s	&	nil	total
pob	2	0	0	0	2	4
rob	0	2	0	0	0	2
s	0	0	0	0	0	0
&	0	0	0	3	2	5
nil	0	1	1	1	0	3
total	2	3	1	4	4	14

Figure 6 Illustration of how EasyDIAg and GSEQ-DP link units and determine agreement for an example data set. A) Example data set. The top line indicates time in seconds. On- and offsets of events identified by the two raters are shown below. Category abbreviations have been defined in Figure 3. B) Illustration of how EasyDIAg links the units of the example file. Green line connecting indicates linked events classified as agreement, red lines linked events classified as disagreement. C) Illustration of GSEQ-DP's processing of this file. The blue '&' indicates inserted gap events. Since GSEQ-DP aligns the sequences twice (R2 to R1 and R1 to R2), individual units can either become linked during both iterations, in which case two

scores are tallied to the agreement (indicated by double-headed arrows), or only during one iteration, causing only one agreement score being tallied (single-headed arrows).

Unlike EasyDIAG, GSEQ-DP does not classify the first and the last pair of events as disagreement (see Figure 6c). Instead, they are classified as commission/omission errors. This is less preferable, because knowledge about which categories are confusable is important for researchers, in order to identify training needs and sharpen diagnostic criteria to distinguish between categories.

Finally, although the overall agreement indices provided by the two algorithms are very comparable, GSEQ-DP tends to provide higher category-specific kappa values than EasyDIAg (compare the last column of Table 3 with Figure 4), with values being about 12% higher on average. EasyDIAg has a problem when one rater identifies one long segment but the other rater annotates the same interval as consisting of multiple short sequences of the same type. Whereas EasyDIAg fails to link events in such a case, because the overlap criterion tends not to be fulfilled, GSEQ-DP, due to its dynamic nature, is better suited for such situations (for an illustration, see Supplementary Online Figure 1).

General Discussion

We have presented a new approach for estimating inter-rater agreement for timed-event sequential data. Although our toolbox was developed with an audience of gesture researchers in mind, the toolbox should be useful for other observational approaches as well.

Existing algorithms for linking sequential events of variable duration (for an overview, see Bakeman et al., 2009) require the specification of at least two parameters, a tolerance parameter (i.e. how much temporal offset one is willing to accept) and an overlap parameter (i.e., how much temporal overlap there should be between events). In contrast, our algorithm requires only one user-specified parameter; the percentage of required temporal overlap. This has several advantages. Fewer parameters mean that the results of an inter-rater agreement analysis are less influenced by arbitrary settings made by the user. It also makes the algorithm easier to use and more user-friendly. A second advantage is that an algorithm that does not have a hard-coded temporal offset parameter (e.g., in seconds) is better suited to deal with coding systems where different event categories tend to have very different event lengths. For example, in the data shown above, annotations for category ‘iob’ are almost 8 times as long as annotations for category ‘s’ (see Figure 3), which would create problems for an algorithm with a temporal offset parameter.

The strong emphasis on temporal overlap as a pre-requisite for the linking of events between raters has several desirable consequences. First, it provides a straightforward method to link events in the first processing step. It is also a fair and unbiased method of linking events, because labels are not considered in this initial linking step. Another advantage, at least for the field of gesture, is that the absence of a temporal offset parameter avoids linking temporally distinct (i.e. non-overlapping) events. Co-speech gestures rapidly develop and illustrate the ongoing discourse (McNeill, 1992), and it would not make sense to try to link annotations that are far apart in time. Finally, since %overlap is calculated as the length of overlap between annotations divided by the length of the longer annotation, the algorithm is symmetric. In other words, it produces identical results when annotations are swapped between raters (i.e., declaring Rater 1's annotations as coming from Rater 2, and vice versa) so the results are less arbitrary.

In comparison with a previously published algorithm, the GSEQ-DP (Bakeman et al., 2009), our toolbox achieved comparable overall agreement indices, suggesting that it is a valid tool to determine inter-observer reliability for timed-event sequential data. Users that require optimal category-specific agreement indices could consider using GSEQ-DP, since it provides slightly higher category-specific kappa values. On the other hand, if users place more emphasis on intuitive rules for linking annotations, in combination with additional features designed to assist in rater training and coding system development (see below), then our toolbox might be useful. Another strength of EasyDIAg is that it not only provides composite agreement indices that jointly take segmentation and categorization into account, but that it can also be used by researchers who require separate indices for segmentation and categorization agreement, respectively (e.g., Bavelas et al., 2008). Users that disagree with our proposed method of linking events, but prefer to link annotations manually, can unfortunately not use our toolbox to calculate agreement indices. In this case, the researcher would have to manually assemble an agreement table (e.g., as shown in Table 1), and use one of the available tools to calculate raw as well as chance-corrected agreement indices (e.g., http://graphpad.com/quickcalcs/kappa1/).

We designed the toolbox in such a way that it will be particularly useful during an initial phase of observational research when a coding scheme is developed and/or raters are trained. The large agreement table quickly provides insights into which categories are confusable, and may require sharper diagnostic distinctions. Additional functions allow researchers to identity which videos and categories are the biggest source of disagreement between raters.

Finally, we hope that the toolbox will help to enable more reliable measurements in observational research. As mentioned in the introduction, many observational studies either do not address inter-observer reliability at all, or fail to provide chance-corrected agreement indices. With the available toolbox, researchers can calculate agreement indices that are chance-corrected and take into account all segmentation and categorization decisions.

Literature

Bakeman, R., & Quera, V. (1992). SDIS: A sequential data interchange standard. *Behavior Research Methods, Instruments, & Computers, 24*(4), 554-559. doi: 10.3758/BF03203604.

Bakeman, R., & Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*. New York: Cambridge University Press.

Bakeman, R., Quera, V., & Gnisci, A. (2009). Observer agreement for timed-event sequential data: a comparison of time-based and event-based algorithms. *Behav Res Methods, 41*(1), 137-147. doi: 10.3758/brm.41.1.137.

Bakeman, R., & Robinson, B. F. (1994). *Understanding log-linear analysis with ILOG: An interactive approach*. Hillsdale, NJ: Erlbaum.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58*(2), 495-520.

Bavelas, J. B., Kenwood, C., & Phillips, B. (2002). Discourse Analysis. In M. Knapp & M. Daly (Eds.), *Handbook of Interpersonal Communication* (3 ed., pp. 102 - 129). Thousand Oaks, CA: Sage.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol, 43*(6), 551-558. doi: 0895-4356(90)90159-M [pii].

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Deming, W. E., & Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics, 11*(4), 427-444. doi: 10.2307/2235722.

Dijkstra, W., & Taris, T. (1995). Measuring the Agreement between Sequences. *Sociological Methods & Research, 24*(2), 214-231. doi: 10.1177/0049124195024002004.

Haccou, P., & Meelis, E. (1994). *Statistical analysis of behavioural data: An approach based on time-structured models*. Oxford: Oxford University Press.

Holle, H., & Rein, R. (2013). The Modified Cohen's Kappa: Calculating Interrater Agreement for Sgementation and Annotation. In H. Lausberg (Ed.), *Understanding Body Movement : A Guide to Empirical Research on Nonverbal Behaviour: With an Introduction to the NEUROGES Coding System* (pp. 261 - 275). Frankfurt am Main: Peter Lang Verlag.

Jansen, R. G., Wiertz, L. F., Meyer, E. S., & Noldus, L. P. (2003). Reliability analysis of observational data: problems, solutions, and software implementation. *Behav Res Methods Instrum Comput, 35*(3), 391-399.

Kaufman, A. B., & Rosenthal, R. (2009). Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. *Animal Behaviour, 78*(6), 1487-1491. doi: http://dx.doi.org/10.1016/j.anbehav.2009.09.014.

Kaufmann, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc.

Lausberg, H. (2013). *Understanding Body Movement - A Guide to Empirical Research on Nonverbal Behaviour*

With an Introduction to the NEUROGES Coding System. Frankfurt am Main, Germany: Peter Lang Verlag.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES--ELAN system. *Behav Res Methods, 41*(3), 841-849. doi: 10.3758/brm.41.3.841.

Mackinnon, A. (2000). A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. *Comput Biol Med, 30*(3), 127-134. doi: S0010-4825(00)00006-8 [pii].

McNeill, D. (1992). *Hand and Mind - What Gestures Reveal about Thought*. Chicago: The University of Chicago Press.

Quera, V., Bakeman, R., & Gnisci, A. (2007). Observer agreement for event sequences: methods and software for sequence alignment and reliability estimates. *Behav Res Methods, 39*(1), 39-49.

Rein, R. (2013). Using 3d kinematics of hand segments for segmetnation of gestures: A pilot study. In H. Lausberg (Ed.), *Understanding body movement: A guide to empirical research on nonverbal behavior: with an introduction to th eNEUROGES coding system* (pp. 163 - 187). Frankfurt am Main: Peter Lang Verlag.