Event Chronography in Multi-modal Data: the BME Method for Quantitative Analyses

Anaïs Claire Murat, Maria Koutsombogera and Carl Vogel
Trinity Centre for Computing and Language Studies
Trinity College Dublin, the University of Dublin
{murata, koutsomm, vogel}@tcd.ie

Abstract

Methods for investigating multi-modality in human interactions remain open to refinement. Although the annotation process has been facilitated by tools like Elan, synchronising exported cross-tier data for further quantitative analyses remains challenging. We present the BME method: a new approach to data alignment. The idea is straightforward: instead of comparing exact times of onsets, durations, etc., the BME method focuses on their organisation. First, the method describes every annotation by at least two events: its beginning (B) and end (E). Then, it aligns them in chronological order. Middles (M) are precipitated to track events from other tiers which might occur between Bs and Es. We explore three cases in which such an arrangement of multi-modal data can benefit the scientific community: first, in getting insights about the dynamics and dependencies between tiers, second, in contemplating event-based duration rather than time-based ones, and, third, in contributing cross-annotator agreement assessment methods.

Keywords: Multimodality, Annotation, Alignment Methods, Segmentation Agreement, Multimodal Corpus, Multimodal Analysis

1. Introduction

Investigating multi-modality in human interactions relies on annotation software. With its user-friendly interface, Elan (Wittenburg et al., 2006) allows the rigorous synchronisation of multiple tiers based on a frame-by-frame decomposition of any video input and has become a default software for many researchers within the community. However, things get complicated when it comes to exporting the data in a meaningful manner and interpreting them when trying to understand relations among tiers. Indeed, distinct channels (e.g., vocal linguistic content, laughter, gesture, gaze, etc.) do not perfectly align in start or end times, which makes it difficult to count events in cross-modal comparisons.

Our BME method supplements the exact onset and offset times of the data with primary focus on their overall organisation. It describes every single annotation by its beginning (B) and its end (E) and, then only, aligns them in chronological order. More broadly, we call these beginnings and ends "events". Whenever the exploration of another modality – or tier – leads to the creation of an event in between already existing Bs and Es, it precipitates a middle (M) annotation (see Figure 2 and section 3 for an illustrated example). In this paper, after explaining in detail what this arrangement looks like, we explore three cases in which such an arrangement of multi-modal data can benefit the scientific community: first, in getting insights about the dynamics and dependencies between the tiers; second, in contemplating event-based durations rather than time-based ones; and, third, in re-exploring cross-annotator agreement method, notably for segmentation tasks. For many analyses we can ignore the exact start times and instead focus on orderings and count-based assessments of interactions between annotation categories.

2. Background

2.1. Annotation Tools and Data Processing

In the study of human-human interactions (HHI), working on multiple modalities is ubiquitous. For example, from videos can be retrieved gestures and gaze behaviour which can inform turn-taking, semantics, focus, and much more. However, to reach such fine-grained results, fine-grained annotations are also required, from time-aligned speech transcriptions to time-aligned gaze and gesture annotations.

In 2006, Wittenburg (Wittenburg et al., 2006) introduced Elan, a tool which allowed the annotation and the synchronisation of a multi-modal data set based on timestamps. Since then, its frame-by-frame video management has been widely used by the HHI community. Yet, as acknowledged by Parisse et al. (2022) in the making of the DINLANG corpus, extracting data for comparison purposes on Elan can be quite laborious. Moreover, once the query tool is mastered, the limited range of possible outcomes can become frustrating. Above all, two aspects can be unsatisfying. On one hand, the duplication of information and, on the other hand, the loss of information: for instance, an annotation that spans over two others might be counted twice in

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9217–9225
11–16 May 2026.

the exact same manner, and all information about where it started (either before or inside the first annotation) is lost. These limitations were pointed out in Murat et al. (2022) as a possible cause of unsatisfying results on very short annotations such as mutual gazes and turns. To overcome it, a few methodological decisions can be made, such as making use of temporal windows and split the annotations that cross the boundaries (Kousidis et al., 2009), considering longer intervals (such as dialogue sections (Gnjatović et al., 2018)), or discontinuous ones (e.g. in-between turns, cf. Nakano et al., 2003), along with some kinds of binary categorical data (presence or absence). In this paper, we propose another solution, one where we abstract from time to keep from annotations their chronological ordering rather than their strict timestamps. By doing so, we allow the characterisation of a segmentation by the context of in which happens its Beginning, its End in relation to other events of relevance (see section 3).

2.2. Towards Event-based Methods

Across fields, looking at sequences of events is of frequent interest. In NLP, for example in name entity recognition (or annotation of phrasal expressions, more generally), it is common practice in text annotation to use BIO (for 'Begin', 'inside' and 'outside') to mark (B) the tokens that begin a named entity, that follow and are part of it (I), or that are not involved into a named entity at all (O) (Moreau et al., 2018; Xu et al., 2022). In formal semantics, Fernando (2019) highlighted the continuing relevance of the Aristotelian dictum "no time without change" by encoding the time of narration as a string of stutterless states, thus ignoring the actual duration of an event but considering one single time unit for two identical and juxtaposed states. Lastly, although the BME method was not inspired by it, T-Patterns (Magnusson, 2020) are a tool focusing on onsets and offsets which is vastly used in the behavioural study field to find (hidden) patterns within the data. They argue that their probabilistic approach to patterns allows for flexibility, notably when some components are missing, and that focusing on starts and ends enables investigation of the anteriority of events and hence helps with causality. Hunyadi (2019) showed the relevance of T-patterns to reveal patterns in the multi-modal HuComTech corpus. Such works leave us hopeful for investigating further event-based methods.

2.3. Cross-Annotator Agreement in Multi-Modal

In cross-annotator agreement, continuous data is a challenge. Very often, there is an attempt to bring back the challenge to a simple categorical task by already providing annotation boundaries (see the multi-modal corpora Recola and Sewa in Han et al. (2021) for which time and value-continuous dimensional annotations are provided following a constant frame rate of several milliseconds), or by considering it as a whole, either making it somewhat dynamic (e.g. Metallinou and Narayanan (2013) where emotions are considered as a continuous and gradual metric and the time series thus created can then be compared), or calculating the overall duration of convergence and divergence of categories (Gilmartin, 2021). In Carletta (1996) these observations were made, and they suggested adapting the Kappa metrics as a way to measure the amount of the annotations which agrees while taking into account what could have simply been due to chance. In this process, Holle and Rein (2015) provides a method to assist with the linking problem (matching two instances annotated by two different annotators). However, beyond the chase for the right metrics, Artstein (2017) also insisted on the need for further investigation into how the data looks, rather than only trusting a single number. We hope that the method we describe here and for which we provide Elan-compatible annotations can support exactly such examinations.

3. The BME Method

As stated above, the BME method organises multimodal data based on the relative onsets and offsets of their annotations. Our algorithm arranges the data from two different tiers accordingly; our implementation is available as open source code.[1] From exported tiers from Elan, it retrieves the onsets and offsets of each annotation and organises them relatively to each other. Middles are then added and enumerated, so that the final number of Ms for a given annotation can be displayed next to its E line.

For instance, consider Fig. 1 as an example:

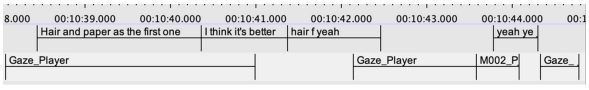

Figure 1: Example of two synchronised modalities: turns (above) and mutual gaze (below). Annotation extracted from the Multisimo Corpus, aligned on Elan.

The first mutual gaze annotation overlaps the two first turns. If we were to simply describe the data by turns then they would both repeat the exact same mutual gaze content and, at the same time, lose every information about where "Gaze_Player" started, and where it ended. If mutual gazes were to occur completely outside of a turn, they would

[1] GitHub: https://github.com/anaismu/BME_Method/

also be completely ignored and not represented in the synthesised data. Now, consider the same example with the BME arrangement (Figure 2).

TURNS	MUTUAL GAZE
0	B – Gaze_Player
B – Hair and paper...	M1 - Gaze_Player
E – Hair and paper...	M2 - Gaze_Player
B – I think it's better	M3 - Gaze_Player
M1 – I think it's better	E – Gaze_Player
E – I think it's better	0
B – hair f yeah	0
M1 – hair f yeah	B – Gaze_Player
E – hair f yeah	M1 - Gaze_Player
0	E – Gaze_Player
0	B – M002_Player1
B – Yeah ye	M1 – M002_Player1
M1 - Yeah ye	E – M002_Player1
E - Yeah ye	0

Figure 2: BME ordering based on the example of Figure 1

Now, the data frame reads: "mutual gaze 1 started outside a turn (0), it spanned through two turn beginnings (B), and one turn end (E), before ending in the middle (M) of the second turn"; hence preserving information about both the categories and the time arrangement of the data. Furthermore, since every "M" (middle) is indexed and corresponds to an event on the other tier, there is also no more exact repetition of content. We also argue that counting Ms can account for a relative, event-based definition of duration: thus, the first mutual gaze has a duration of 3, the two following ones, a duration of 1, and the first turn, a duration of 0.

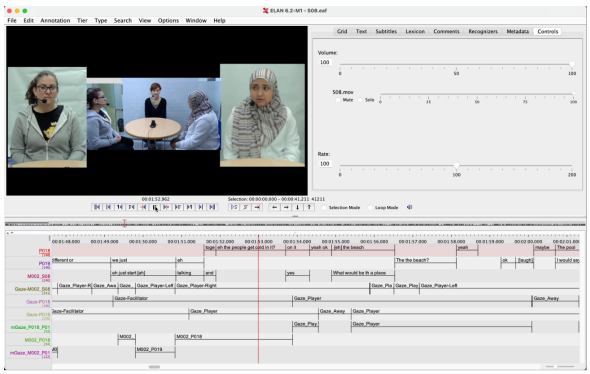

Figure 3: Excerpt of the Multisimo Corpus in Elan

	Mutual Gaze		Turns	
Turns	**B**	**E**	**B**	**E**
0	20	23	429	428
M	232	229	256	257

Table 1: Contingency Table of Turns' and Mutual Gazes' BMEs

4. Case Study #1: Tier Dynamics

4.1. Method

Our two first applications are based on the Multisimo corpus (Koutsombogera and Vogel, 2018). This multi-modal corpus investigates collaboration in three-party conversations. It is made of 18 dialogues in English of about 2,000 - 3,000 words each. Each dialogue includes two players and one facilitator. They all play a game in which the two participants have to guess and rank the most popular answers to three questions together. This section focuses on two modalities: the spoken and visual ones through turns and mutual gaze annotations. These were manually transcribed and annotated by experts, and thereafter, aligned on Elan. They were then extracted with their timestamps and rearranged according to the BME method with the Python code described in section 3.

For this first application, we try to see how these two tiers organise vis-à-vis each other. Hence, since we are not looking for a general truth but a practical example, only one conversation of the corpus (ID = "S07") was looked at. It spans a total of 252 mutual gazes and 685 turns which were then organised into a contingency table to reveal where these two tiers' Beginnings and Ends fell outside or inside the other tier's annotations.

4.2. Results & Discussion

Table 1 summarises by simply stating for both tiers whether they started and ended inside (M) or outside (0) an annotation of the other tier. Since each annotation necessarily precipitates two events: a beginning and an end, two separate χ^2-tests were run for Bs and Es, although, for this purpose, only one of the two is needed. Results are shown in Table 2. It confirms the low probability of such an arrangement being random ($p \leq 2.22e^{-16}$) in both cases. It notably shows that turns are more likely to start and end outside a mutual gaze (standard residuals > 14) and, on the other hand, mutual gaze are more likely to come inside turns (standard residuals > 14).

Contingency table analysis of the BME data frame of one conversation enabled the quantification of beginnings and endings happening either outside or inside the other tier's annotations. Re-

	MG Bs	Turns Bs
0	-14.86	14.86
M	-14.86	14.86

(a)

MG Es	Turns Es
-14.49	14.49
14.49	14.49

(b)

Table 2: χ^2 Standard Residuals for Mutual Gazes' (MG) and Turns' Beginnings (a) and Ends (b) to fall either Outside (0) or in the Middle (M) of an Annotation on the other Tier.

sults suggest a priority of the turn channel: while most turns start and end outside mutual gazes, mutual gazes tend to mostly start and end inside turns. Such a method can serve as a first step in a research design to quickly indicate how the data are arranged. While this abstracts over durations, the next case study shows that the method supports both elapsed time and event count duration measures.

5. Case Study #2: Event-based Duration

5.1. Method

This application contemplates a way of thinking about duration enabled by the BME method. Instead of considering solely time-based duration (e.g. a mutual gaze's duration is usually measured in milliseconds), we look into event-based duration which is determined here by the number of Ms available inside a given mutual gaze.

This case study is inspired by Murat et al. (2022), which investigated mutual gaze behaviour in relation to lexical repetition as a proxy for alignment (Pickering and Garrod, 2004). Yet, they could not find a relationship between mutual gaze duration and lexical repetition and hypothesised that this was due to the repetition and information loss mentioned in sections 2.1 and 3, rather than an actual absence of effect. We re-used a similar methodology: from the corpus (the entire Multisimo dataset (Koutsombogera and Vogel, 2018), i.e. 3825 mutual gaze & 9282 turns) to their method (repetition count as a Jaccard index between a given turn and the last turn of each participant (collected in a "register")). Turns in the register were encoded as types (set of words), and the turn under observation as tokens (bag of words). For each mutual gaze of the corpus, were retrieved: 1. the amount of repetition at its start, 2. the amount of repetition at its end, and 3. the average amount of repetition across its Middles. These repetitions are divided in two distinct categories: self-repetitions (SR) when the participant repeats themselves, and other-repetitions (OR) when the participant repeats something someone else has said before. Mutual gaze duration was defined twice: once as time-based (raw duration in milliseconds), and once as event-based (number of Ms).

5.2. Results & Discussion

Because of the three-party nature of the dataset, we make a distinction between mutual gazes which include the current speaker and those which do not. Correlations between mutual gazes and their amount of repetition were calculated using Spearman correlation tests. Results of all these 48 tests [2] can be appreciated in Tables 3, 4 & 5.

	Time Duration (S)		Event Duration (M)	
	P-Value	**Rho**	**P-Value**	**Rho**
OR overall†	***	**0.07**	***	**0.22**
SR overall†	***	**0.13**	***	**0.17**
OR at Bs	*	**-0.04**	0.34	0.016
SR at Bs	0.38	0.01	0.05	-0.033
OR in Ms†	0.20	0.02	***	**0.18**
SR in Ms†	***	**0.12**	***	**0.19**
OR at Es	*	**-0.04**	*	**0.037**
SR at Es	*	**0.04**	0.11	-0.03

Table 3: Repetitions Jaccard Index & Duration during any Mutual Gaze. P-Value coding: * < 0.05, ** < 0.005, ***; †: averaged values.

	Time Duration (S)		Event Duration (M)	
	P-Value	**Rho**	**P-Value**	**Rho**
OR overall†	***	**0.08**	***	**0.23**
SR overall†	***	**0.13**	***	**0.17**
OR at Bs	0.10	-0.03	0.37	0.02
SR at Bs	0.52	0.01	0.06	-0.04
OR in Ms†	0.67	0.01	***	**0.14**
SR in Ms†	***	**0.09**	***	**0.15**
OR at Es	*	**-0.04**	***	**0.04**
SR at Es	*	**0.04**	***	**0.21**

Table 4: Repetitions Jaccard Index & Duration during Mutual Gaze *including* the Speaker P-Value coding: * < 0.05, ** < 0.005, ***; †: averaged values.

Unlike the original paper (Murat et al., 2022), we were able to conclude on a relation between mutual gaze duration and linguistic repetition, for both time-based and event-based duration. Such an achievement might be that, indeed, as pointed out in their paper, duplicates of information when it comes to such short annotations was ubiquitous and prevented precise pairing of mutual gaze and repeti-

[2] Repetition type (OR/SR) × BME section (Overall average, at Bs, at Es or average during Ms) × Duration type (time-based or event-based) × Mutual gaze type (Any, including the speaker, excluding the speaker)

	Time Duration (S)		Event Duration (M)	
	P-Value	**Rho**	**P-Value**	**Rho**
OR overall†	***	0.01	***	**0.19**
SR overall†	***	**0.14**	***	**0.16**
OR at Bs	*	**-0.09**	0.77	0.01
SR at Bs	0.57	0.02	0.61	-0.02
OR in Ms†	0.40	0.04	***	**0.25**
SR in Ms†	***	**0.19**	***	**0.25**
OR at Es	0.07	-0.10	0.95	**
SR at Es	***	**-0.07**	*	-0.11

Table 5: Repetitions Jaccard Index & Duration during Mutual Gaze *excluding* the Speaker P-Value coding: * < 0.05, ** < 0.005, ***; †: averaged values.

tion index, whereas the BME method allowed the creation of a precise summary table containing the amount of repetition of the turn in which the mutual gaze started or ended (if there was any), as well as the average amount of repetition across the mutual gaze. Finding a difference between these two duration types was not a given. Indeed, the longer the annotation, the more annotations from another tier it might span through (a Spearman correlation test comparing our time-based durations and our event-based durations showed: $p < 2.2e-16$, rho $= 0.62$). Yet, when comparing the duration of mutual gazes and the amount of linguistic repetition occurring at one turn's Beginning, Middle or End (Tables 4, 5, 6), it appears that, even if both time-based duration and event-based duration lead to similar p-value patterns (Spearman correlation tests), event-based duration might be more closely correlated with the linguistic repetition than time-based duration (mean rho = 0.1 vs 0.03). However, further investigation is required to determine the contexts in which one is more relevant than the other.

From this case study, the conclusion is twofold. First, through the BME ordering, and its ease to identify context at the Beginnings, Ends and in the Middle of annotations, we offer a precise way of exploring co-occurring behaviour; second, we suggest to think duration not only as time but also as a quantification of events which might have started or ended in the span of a current window.

6. Case Study #3: Cross-Annotator Agreement

6.1. Method

As mentioned in section 2.3, assessing agreement over continuous channels can be challenging. In this case study, we reuse the annotations made available by Sasmita and Swallow (2022). In their paper, Sasmita and Swallow asked a large number of annotators to segment videos by pressing the spacebar every time they deemed that a new event occurred. They then compared the resulting segmentations using four methodologies for four different levels: co-presence of boundaries (peakiness) to assess agreement within a group of annotators, distance of these peaks of co-present boundaries (peak-to-peak distance) in-between two groups of annotators, or likelihood of a boundary to fall in the same bin as a peak (agreement and surprise index) for the individual agreement to a group.

Their results demonstrate the reliability of human annotations, with an above-chance possibility to identify meaningful annotations from both online and in-lab annotators. However, to be able to detect such annotations from one study, they need at least a medium group size ($n \geq 6$). We investigate whether the BME ordering allows for comments on the pairwise similarity of two movie annotations. The entirety of the dataset accounts for 6 videos: 2 which are a combination of movie excerpts (total length of about 9 min each) and 4 videos of everyday activities that are between 4.55 and 6.03 min long each. Each of the two movie-videos were annotated by 64 participants in a lab set-up, and the everyday ones by 72 online participants via Mechanical Turk using CloudResearch.

For the sake of brevity, we only focus this case study on the subset of coarse-grained segmentations of the everyday activity "wdishes" (washing dishes) videos (6.03 min long, 35 different annotators). From the raw annotation file recording the timestamp of every click that an annotator made to mark the boundary of a new event, we paired up all annotators ($35 \times 34 = 595$ pairwise comparisons available) and produced a BME ordering of the events. For each pairing-up of the annotators with each other, we arbitrarily assigned them as "Annotator A" or "Annotator B", these categories are important for the rest of the analysis, as they are crucial to the investigation of the directionality of the difference (Δ_M) later explained in eq. 1 , but it is important to understand that any Annotator ID can be characterised as Annotator A or B depending on the given pair. One specificity of this corpus is that annotators were asked to segment a video rather than to annotate it; as such, a Beginning (B) is also an End (E), and there is no possible gap in-between an E and the next B. We then computed two tests: (1) a χ^2-test that relates the number of Es of each annotator to their number of Ms in the total of the video; and (2) the difference in the median number of Ms per segment for the given annotator. χ^2-tests allow to identify significant differences in the ratio of Ms compared to the number of annotated events (in our case encapsulated through the number of Es). We do not account for the number of Bs as these are not independent from Es and, in

fact would, by definition, be exactly the same values as the Es. A χ^2-test whose p-value returns as $p < 0.05$ would indicate that the two annotators significantly differed in their annotations Beginnings, Ends, and lengths. However, it tells very little about the consistency of these differences. It could be possible, for instance, that a video was annotated with very different patterns throughout, or simply that, despite very similar annotations overall, one segment that encompasses several annotations carries most of the difference. With this in mind, we decided to add an asymmetrical post-hoc test which looks at the difference (Δ_M) of numbers of Middles between the annotators A and B, following equation 1.

$$\Delta_M = \tilde{M}_{S_A} - \tilde{M}_{S_B} \quad (1)$$

where $\tilde{M}_{S_x}$ is the median of the number of Middles (M) per Segment (S) for annotator x.

Observing Δ_M permits identifying where the bigger annotations tend to be: a null difference represents either (1) the ideal case where annotators agree exactly on the same millisecond for boundary creation, (2) the more likely case where two annotators agree on the segmentation but end up having a constant mismatch in their onsets and offsets (for instance, a constant motor-induced delay), or (3) the case where most of the segmentations agrees, even though a significant difference might lie in punctual mismatches. On the other hand, if Δ_M is greater or less than zero, the asymmetry of the difference allows to determine which of the two annotators tends to have longer or shorter segments.

6.2. Results & Discussion

Figure 4 synthesises (1) the results of the χ^2-tests that state whether Annotator A's and Annotator B's segmentations significantly differed or not, and (2) their difference in the median number of Ms of each of the segment they created. Looking at the results of the individual χ^2-tests, one can notice amongst the valid (no warning) tests ($n = 509$) a fair ratio between the number of significant ($n = 288$) and non-significant tests ($n = 221$). 86 tests returned invalid, likely due to too few segments from at least one of the annotators (indeed, the average duration for segments by annotator 218 and 229 is respectively 178s, 14 event long and 178s, 21 event long while the average over all the annotators is 29s and 2.6 events).

As for the difference in Ms, while it seems that the non-significant tests are close to $\Delta = 0$ –i.e. both annotators constantly span through a similar number of boundaries of the other annotators – the significantly different pairs can be marked by a possible – not necessary – difference in the median number of their Ms per annotation. While the intensity of the colour marks the absolute difference between the two annotators, a blue tint denotes that Annotator B had longer events, while a red tint supposes otherwise. Taking a more global look at the patterns created thanks to the organisation of tests per ID in fig. 4, one can hence observe horizontal and vertical lines of significantly different annotations. This allows to quickly identify annotators that differed from the rest. For instance, annotators 468 and 445 – which were arbitrarily constantly identified as annotator B – result in two vertical lines of significantly different tests and positive difference in number of Ms: they likely annotated overall shorter events than their Annotator A peers, especially when paired with 218 and 229 which, in turn, are marked by many invalid paired tests suggesting fewer – hence longer – events. The same reasoning about 187, 197 and 204 can apply: when considered as Annotator B, their paired tests are significantly different and the difference in number of Ms is positive, but negative when considered as Annotator A (fig. 4), which suggests that their segments tend to be shorter than the average.

In sum, the BME framework allowed explorations of annotators agreement at two different levels. At the individual level, two sets of annotations can be compared and classified as significantly – or not significantly – different; and looking into the median – as opposed to the means which is more sensitive to outliers – number of Ms allows to characterise how the two annotators differ in granularity. To our understanding, this one-to-one level (group of size 2) was not reliably performing above chance when it came to showing agreement in the original research (Sasmita and Swallow, 2022). For overall every-day activity videos, they reported a fairly poor agreement index (0.15); this quick visualisation of a subset of their data allows to see why: out of 595 pairs of annotators, only 221 were not significantly different; 374 were either significantly different from one another or the test could not be successful due to too little segments –hence very coarsly segmented videos. Furthermore, at the task level, the representation in figure 4 shows how consistent the results of the χ^2-tests based on the BME are. At one glance, one can visualise all the productions of the annotators and how they fit in relation to the others; they can easily for instance identify which annotators stand out – in our case, 187, 197, 204, 445 and 468 overall annotated shorter segments than others; 218 and 229 annotated fewer longer segments. In line with the recommendations of Artstein (2017), our analysis produces statistical evidence for (mis-)matching annotations, but also allows further understanding of what the annotations look like with respect to each other.

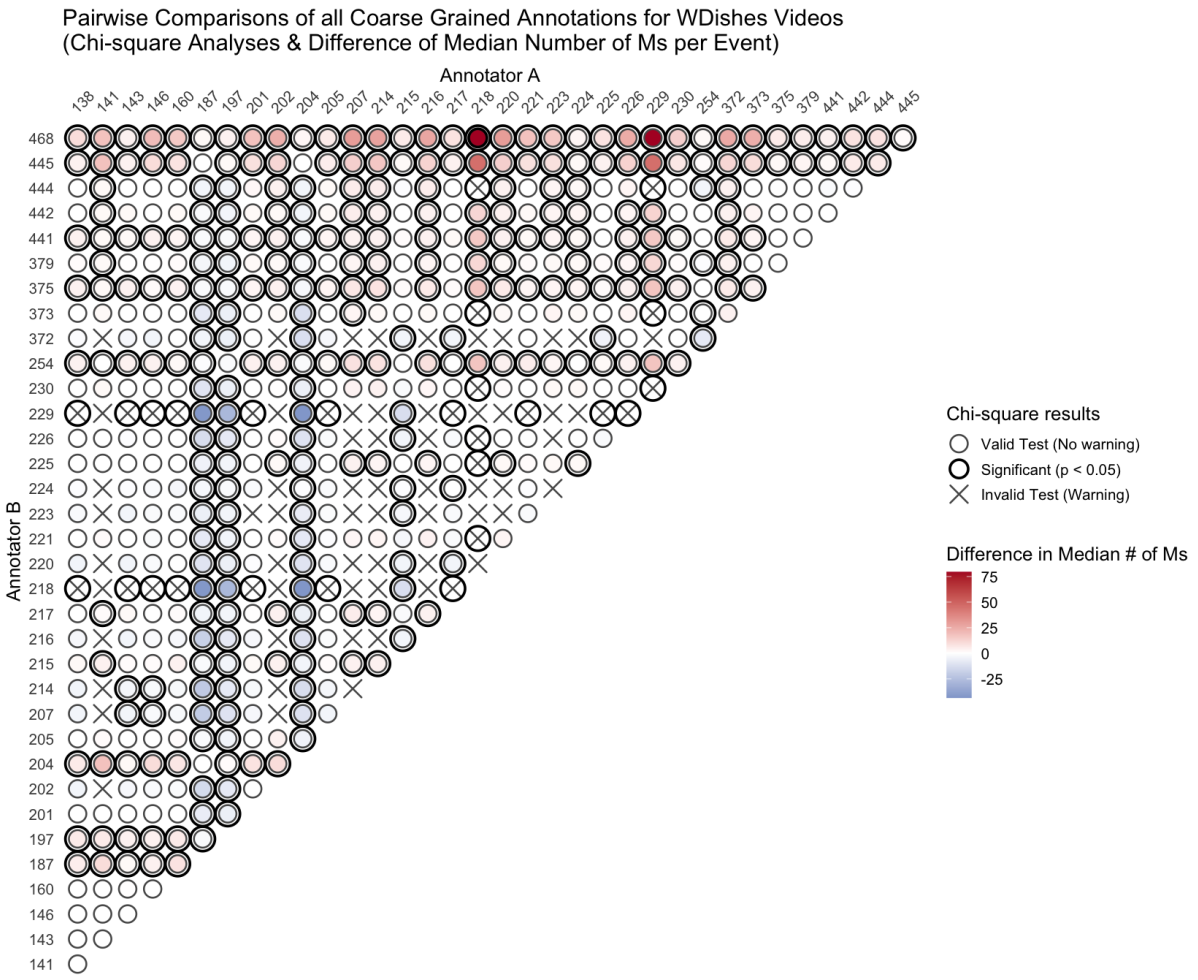

Figure 4: Illustration of Pairwise Comparison ($n = 595$) of all Coarse-Grained Annotations for WDishes Videos. Bold circles represent significant ($p < 0.05$) χ^2-tests (i.e. disagreement) while the colour indicates the difference in Median of the number of Ms per event in the pairs (i.e. scale of granularity misalignment). "Warning" ($\times$) suggests that the number of expected values for the χ^2-test were too few.

7. The BME method: Discussion

The BME method is a first step towards exploring multi-modal data alignment differently. As shown in case study 1 (section 4), the BME ordering can be a first step to better understand the data dynamics. It can allow visualisation of a range of relationships between the modalities: for instance, which one has longer duration? Which is more coarse-grained or fine-grained? Are they likely to co-occur or instead to be completely independent of each other? And so on. Investigating what happens precisely at the Beginning, End or in the Middle of a certain annotation, like in case study 2 (section 5) can, on top of providing further evidence of correlation, also be used to start think of, if not causality, chronological evidence of occurrence, in the sense that one can observe what came first and the context in which something is likelier to occur or terminate, similar to the claim made by the Magnusson (2020) about the T-patterns method which focuses on onsets and offsets to probabilistically find patterns in the data. As shown in section 5, on top of allowing to quantify Beginnings and Ends, the BME arrangement of data also permits thinking of duration not solely as a time-based measurement, but as an *event-based* measurement. That is, the length of a given annotation could not only been considered in terms of (milli-)seconds but in terms of how many events this annotation has spanned through. In a way, when studying gaze, for instance, this could provide a sense of the persistence of a look where one staring into one direction for 10 seconds is all the more noticeable if someone else is hailing at them during that gaze, trying in vain to gain their attention.

Eventually, our third case study (section 6) moved away from mere analysis of multi-modal data to offer a view into another challenge of corpus making: cross-annotator agreement, with a focus on segmentation. By arranging the annotations

(decisions on event boundaries) of two annotators in a BME fashion, we showed how comparing the number of Beginnings or Ends and the number of Middles of the two annotation sets can provide insights into whether they are significantly different from one another. Additionally, quantifying in detail the number of Ms per annotated event helps understand whether one of the annotators consistently accounted for bigger or smaller events. If many annotators are available, similar to our case study, one can, by computing all annotators' production in a pairwise manner, also find patterns in the outcome results and thusly retrieve annotators which stand out from the others.

8. Conclusion

The BME method is an event-based annotation method that we suggest using in the case of overlapping data such as multi-modal human interaction processing, but also any other field that has to deal with time-continuous data and possibly overlapping annotation tier boundaries. Indeed, its particularity lies in the treatment of such annotations: instead of focusing on their exact onset and offset times, the BME focuses on their Beginnings, their Ends, and what happens in their Middles (understand "in between their beginnings and ends").

This creates an alternative to fixed-duration sampling of the state in each modality (where there is a row in any data frame for each such snapshot). Instead, state changes in each modality are recorded (so that there is a row in any data frame for a change of state in any modality). This alternative typically involves more compact data frames than the fixed-duration method, while still recording the timing and durations of events. As it is event-based rather than fixed-duration sampling, the method enables count-based analytical comparisons while still supporting durational comparisons. We demonstrated the relevance of such arrangement in three distinct applications: to study tier dynamics, to think of duration not only in terms of time but also in terms of number of events co-occuring, and to assess and visualise cross-annotator agreement. The potential of the BME however is not restricted to these exact applications, and the authors of this paper welcome collaboration to implement such technique on various datasets and cases from data analysis to annotator agreement.

9. Ethics Statement and Limitations

This work proposes a method for thinking and computing multi-modal annotations, as well as its potential for assessing annotator agreement in placing boundaries on continuous data. As such, we did not produce new corpora and worked from already ethics-approved and publicly accessible resources: the Multisimo Corpus (Koutsombogera and Vogel, 2018) and the annotations collected and used in Sasmita and Swallow (2022). Even if section 5 did observe significant correlation between gaze and repetition which previous literature had failed to observe in the same dataset, the purpose of this paper is not so much to directly contribute to the field of multi-modal interaction but rather to offer a new perspective on dealing with overlapping annotations. As such, our argumentation is very little – if not at all – impacted by possible biases that might have occurred during data collection. As for the limitation of the paper, we need to acknowledge that ordering data following the BME method can, at first, reveal interesting considerations and patterns, as developed through the three case-studies that we have explored but it does not do it all. It is mostly a first seed into a new way of thinking about time-aligned data. It will need to diversify as it faces different research questions and methods, and we are hopeful that these will lead to greater and more stable methods to investigate multi-modal interactions.

10. Statement on the use of Generative AI

No Generative AI tools have been used in the performance or documentation of this research: not in literature review, idea development, coding, data analytics, writing, etc. This work is entirely the product of humans.

11. Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223.

12. Bibliographical References

Ron Artstein. 2017. Inter-annotator Agreement. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 297–313. Springer Netherlands, Dordrecht.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254. Place: Cambridge, MA Publisher: MIT Press.

Tim Fernando. 2019. Pictorial Narratives and Temporal Refinement. *Semantics and Linguistic Theory*, 29(0):43–62. Number: 0.

Emer Gilmartin. 2021. *Composition and Dynamics of Multiparty Casual Conversation: A Corpus-based Analysis*. Ph.D. thesis.

Milan Gnjatović, Jovica Tasevski, Branislav Borovac, and Nemanja Maček. 2018. An Entropy-Based Approach to Automatic Detection of Critical Changes in Human-Machine Interaction. In *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000175–000178. ISSN: 2380-7350.

Jing Han, Zixing Zhang, Maja Pantic, and Björn Schuller. 2021. Internet of emotional people: Towards continual affective computing cross cultures via audiovisual signals. *Future Generation Computer Systems*, 114:294–306.

Henning Holle and Robert Rein. 2015. EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, 47(3):837–847.

Laszlo Hunyadi. 2019. Agreeing/Disagreeing in a Dialogue: Multimodal Patterns of Its Expression. *Frontiers in Psychology*, 10:1373.

Spyros Kousidis, David Dorran, Ciaran Mcdonnell, and Eugene Coyle. 2009. Convergence in human dialogues time series analysis of acoustic feature. *Convergence*, 2009:01–01.

Magnus S. Magnusson. 2020. T-Pattern Detection and Analysis (TPA) With THEMETM: A Mixed Methods Approach. *Frontiers in Psychology*, 10.

Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

Erwan Moreau, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel, and Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In *Multiword expressions at length and in depth*, pages 177–207. Language Science Press, Berlin.

Anais Murat, Maria Koutsombogera, and Carl Vogel. 2022. Mutual Gaze and Linguistic Repetition in a Multimodal Corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2771–2780, Marseille, France. European Language Resources Association (ELRA).

Yukiko Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a Model of Face-to-Face Grounding. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 553–561, Sapporo, Japan. Association for Computational Linguistics.

Christophe Parisse, Marion Blondel, Stéphanie Caët, Claire Danet, Coralie Vincent, and Aliyah Morgenstern. 2022. Multidimensional coding of multimodal languaging in multi-party settings. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2781–2787, Marseille, France. European Language Resources Association.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02).

Karen Sasmita and Khena M. Swallow. 2022. Measuring event segmentation: An investigation into the stability of event boundary agreement across groups. *Behavior Research Methods*, 55(1):428–447.

Hesheng Xu, Bin Hu, and Suneet Kumar Gupta. 2022. Legal text recognition using lstm-crf deep learning model. *Intell. Neuroscience*, 2022.

13. Language Resource References

Koutsombogera, Maria and Vogel, Carl. 2018. *Modeling Collaborative Multimodal Behavior in Group Dialogues: The MULTISIMO Corpus*. European Language Resources Association (ELRA).

Sasmita, Karen and Swallow, Khena M. 2022. *Measuring event segmentation: An investigation into the stability of event boundary agreement across groups*.

Wittenburg, Peter and Brugman, Hennie and Russel, Albert and Klassmann, Alex and Sloetjes, Han. 2006. *ELAN: a Professional Framework for Multimodality Research*. European Language Resources Association (ELRA).