A Tool for Determining Distances and Overlaps between Multimodal Annotations

Camila Antônio Barros[1]
Jorge Francisco Ciprián-Sánchez[2]
Saulo Mendes Santos[3]

[1]Freie Universität Berlin, Institute for Romance Philology,
Habelschwerdter Allee 45, 14195 Berlin, Germany
c.antonio.barros@fu-berlin.de

[2]University of Potsdam, Digital Engineering Faculty, Hasso Plattner Institute
Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany
Jorge.CiprianSanchez@hpi.de

[3]Federal University of Minas Gerais / Paris-Saclay University
Interdisciplinary Laboratory of Digital Sciences, bât. 507, Rue du Belvédère, 91400 Orsay, France
saulo.mendes-santos@universite-paris-saclay.fr

Abstract

Comparing annotations is a constant and necessary step in corpus analysis. Although the nature of these annotations is normally research-specific, the tools used for this purpose do not have to be. Here, we present a tool for extracting and comparing annotations from ELAN, despite their idiosyncrasies. The intention behind this tool is to provide a handy way to analyze ELAN annotated files by comparing tiers to a reference unit. Using the presented tool, it is possible to see how tiers overlap (even if they are of symbolic type), to which ratio, and the displacement regarding a reference unit. We present an example of multimodal corpus analysis, regarding the coordination between speech and gesture units based on a pragmatic reference. We argue that looking into overlap ratios can be more informative of the association between speech and gestures and that considering a time buffer between speech and gestural events can be misleading.

Keywords: Corpus Linguistics, Gesture, Prosody

1. Introduction

Over the last few decades, scholars have been studying the relationship between gesture and speech, which seem to play intertwined roles in communication. For instance, gesture and speech have been shown to exhibit fairly similar onsets and offsets and convergence in meaning (Kendon, 2004). Prosody is now deemed a "half-tamed" (Bolinger, 1978) linguistic component; it can change –not always alone, cf. (Tomasello et al., 2022)– the pragmatic and semantic meaning, and it plays a crucial role in conveying the segmentation of speech (Izre'el et al., 2020). Slowly but steadily, gestures have also been shown to play a significant role in conveying meaning and structuring speech with varied degrees of conventionalization. This relationship is grounded on the so-called *Synchronicity Rule* put forth by McNeill (McNeill, [1992] 1995).

The synchronicity rule spans three levels: phonological, semantic, and pragmatic. They are not hierarchically sorted but rather complementary to each other. The phonological synchronicity states that "the stroke phase of the gesture is integrated into the phonology of the utterance" (McNeill, [1992] 1995, p. 26), without defining the exact nature of this integration. The semantic synchronicity rule posits that gestures and speech co-occur because they cover the same idea unit (Chafe, 1980). The pragmatic synchronicity rule postulates that gesture and speech perform the same pragmatic function. Concretely, aspects of the communicative interaction (Streeck, 2006), such as the co-expressiveness of gestures and speech, can be used to aid the information patterning, i.e., the way a certain content is presented for a communicative purpose. The way the synchronicity rule is formulated indicates that a temporal overlap of gesture and speech content drives the synchronicity. However, it remains unclear in the literature to which extent the three levels of the synchronicity rule are entangled and whether they are the *only* factors in play. For instance, rhythmicity could also come into play in positions where semantic and pragmatic meanings are weaker.[1]

In this work, we are particularly interested in the synchronicity postulated for the pragmatic level, i.e., how the structures of gestures and information patterns are aligned. Since this analysis has been conducted in different ways in the literature, we

[1]We thank the first reviewer for pointing this out.

LREC-COLING 2024, pages 1705–1714

provide an overview of different perspectives to ground our annotation scheme. Then, we present a case study to argue that (i) the amount of overlap is crucial for the synchronicity, and (ii) the pragmatic synchronicity drives the phonological synchronicity. To conduct this analysis, we propose a tool that is fit not only for this purpose but also for other kinds of analysis.

2. Theoretical Background

This section is organized as follows: first, we present how the synchronicity rules were initially formulated and tested. For this purpose, we follow a historical order with different definitions used and their analytical implications. We put a special focus on the fact that prosody is a key feature in understanding the coupling of gestures to speech.

Kendon (1972) suggested that gesture phrases can be related to prosodic phrases. A *gesture phrase* is a movement of hands and arms that is salient enough to be perceived as one gesture. It necessarily contains a *stroke*, in which the effort and shape of the gesture are at the clearest. *Prosodic phrases* are the smallest syllabic group over which a completed intonation tune occurs, and the *prosodic phrase* has a certain kind of independence with respect to the *intonation group*, a higher-order grouping made up of multiple phrases. Kendon did not fully compromise with this definition pointing out that "[s]omewhat different definitions of it [the prosodic phrases] have been offered by other writers, but there seems to be wide agreement that it represents a basic unit of speech, the basic move, as it were, of the speaking process" (Kendon, 1972, p. 184). A generous reading of this entails that provided a *basic unit of speech*, gestures use this unit as a reference unit to synchronize phonological, semantic, and pragmatic meaning.[2] *Reference units*, also called *landmarks* by some authors, are the speech phenomena (words, prosodic phrases, terminated units, tone groups, pitch accents, etc.) used as a reference for comparison with gestures, both to its temporal alignment or to its semantic and pragmatic meaning. This classification motivated McNeill's synchronicity rule, opening a door for different theoretical perspectives on the prosodic side to take place while the division of gesture units and phrases remained.[3] McClave (1994) worked under the same paradigm, pointing out that beat gestures, flicks of the hand, could be patterned with stressed and unstressed syllables. The takeout of this is that the prosodic phrase ensembles a meaning through its coordination with gesture phrase.

The follow-up came from the Autosegmental Metrical Theory (Pierrehumbert, 1980). The *basic unit of speech* under scrutiny was narrowed down to *intonational phrases*, a speech span delimited by a *boundary tone*, and carrying *pitch accents*, "tonal movements associated with stressed syllables" (Loehr, 2004, p. 57). In contrast to the broad and intuitive definition of the *prosodic phrase* by Kendon (1972) and McClave (1994), the *intonational phrase* is determined by the placement of a high or low tone. This means that, under this paradigm, gestures are specifically applicable to a narrower range of speech.

In gesture studies, this was taken on by Loehr (2004) to ascertain how gesture apexes, the kinetic goal of a gesture, were distributed in relation to pitch accents. To understand this better, it is convenient to define the gesture's inner parts more closely. The first possible segmentation is into *gesture units*, delimited by rest positions. The gesture unit is the full excursion that hands and arms undertake before resting. Varying from speaker to speaker, rest positions are configurations in which there is little to no tension in hands and arms. Most importantly, it is not possible to recognize any communicative value in a rest position. Within this unit, there is the *gesture phrase* in which it is possible to recognize that some parts are more and others less salient. The more salient parts that can be distinguished are the *strokes*, the central part of a gesture phrase in which postures and handshapes "are better defined than elsewhere in the excursion" (Kendon, 2004, p. 112). The stroke can (but must not) be framed by a preparation, hold, and retraction. The *preparation* comprises the positioning and tensioning of arms and hands that will lead to the stroke. The *hold* is when a speaker sustains their hands in the same position and handshape as the stroke. When the speaker relaxes the hands, it is called *retraction*. In the stroke, there is a specific point that can be distinguished as having the "kinetic goal of the stroke" (Loehr, 2004, p. 89), an *apex*.

Loehr (2004) expected the kinetic goal of the stroke to be associated with tonal movements linked to accented syllables.The author used words as a reference to analyze whether apexes and pitch accents were associated (i.e., given a word, how apexes and pitch accents aligned in time) and whether gesture phrases and intonational phrases were associated regarding their on- and offsets. The results pointed to an association of apexes and pitch accents, given a *time buffer*, a distance of 275 ms in which apexes and pitch accents could be apart but still considered associated – a rarely

[2]The definitions and terminologies presented here were taken from their readings in gesture literature to be as faithful to their use as possible. We do not intend here to put all those definitions under scrutiny but rather to understand the terminology used in gesture studies.

[3]The reader is referred to Müller (2018) for a review on the terminology.

challenged value. Under the synchronicity rule, this can be seen as a semantically driven phonological synchronicity instead of raw temporal coordination with the prosodic phrase, as words were used as reference units for synchronicity.

The analysis of how gesture and speech correlate based on a semantic reference served as the basis for many gesture-speech synchronicity studies, which tested phenomena associated with pitch accents, such as the relation with focused words (Butterworth and Beattie, 1978; Roustan and Dohen, 2010; Dohen and Roustan, 2017) and lexically stressed syllables (Rochet-Capellan et al., 2008), or using a pragmatic reference to analyze intonational apexes (Nobe, 1996; de Ruiter, 1998), prominences (Esteve-Gibert and Prieto, 2013; Rohrer et al., 2022), and speech boundaries (Lelandais and Thiberge, 2023).

As per the literature, it seems that, if the synchronicity of gesture and speech is pragmatically or semantically driven, so should the phonological synchronicity derive from this temporal coordination: a reference unit in the pragmatic or semantic domain drives the phenomena that appear at the phonological level.We agree with this point: speech is not organized aimlessly; its organization rather serves a communicative purpose. It is this communicative purpose that determines the placement of prosodic phenomena in time and drives the placement of gesture apexes. This makes the domain in which these phenomena occur crucial for the analysis. The definition of the domain, or the *basic unit of speech*, on which this is anchored, is not as agreed upon as many of the authors claim (Izre'el et al., 2020).

A potentially more adequate framework that can be considered for addressing the matter is the Language into Act Theory (L-AcT) (Cresti, 2000), a corpus-driven theory that expands on Austin's Speech Act Theory (Austin, 1962) to explain the organization of speech. This theory defines the prosodic units based both on their pragmatic as well as prosodic autonomy, making its basic units more suited to establishing a domain for the pragmatic association with gestures. The L-AcT has two main reference units. The *terminated unit* (TU) is the smallest speech chunk that has pragmatic and prosodic autonomy. By pragmatic autonomy, the theory means that this unit performs a speech act; by prosodic autonomy, that the unit is enclosed by a perceived terminal boundary. This unit can be internally further parsed through non-terminal boundaries, thus creating other *prosodic units*. This ensemble of prosodic units within the TU is called a prosodic *pattern*. The pattern can be simple (when it has only one prosodic unit) or compound, when it has more than one. The L-AcT put forth that the prosodic patterning exhibits a tendential isomorphism with the informational patterning (at the pragmatic level). This means that each prosodic unit of the pattern will accomplish an information function at the pragmatic level. The L-AcT proposed a closed framework of *Information Units* (IUs). Some are devoted to building up the text of the TU, such as the Topic, the Comment, the Parenthesis, the Locutive Introducer, and Appendices of Topic and Comment; others are devoted to regulating the interaction – functions typically known as Discourse Markers. [4] Each IU is characterized by an information function, a prosodic form, and distributional constraints. The most important IU is the Comment, which is defined as the unit carrying the speech act. This is thus the sole IU needed to form a TU. The TU can be of two kinds: the *utterance* and the *stanza*. The utterance is formed by a unique pattern, i.e. one illocutionary IU (the Comment) and other optional IUs. But some kinds of interactions may present TUs in which more than one pattern is juxtaposed by non-terminal boundaries. In this case, we have a stanza, which is characterized by having more than one – sometimes many – illocutionary units to which other optional IUs may be attached. Stanzas typically happen when the interaction exhibits a weaker activation. For instance, while monologues tend to produce longer stanzas that follow the speaker's flow of thought, more active interactions (a football match or receiving instructions in a gym) tend to exhibit much more utterances built upon a simple prosodic pattern.

Differently than the Autossegmental Metrical Theory definition of the basic unit of speech, the *intonational phrase* being limited by a *boundary tone*, the L-AcT uses the interplay of prosodic and pragmatical features to parse spontaneous speech into TUs and *prosodic units*.

For this work, we will use the IU of *Parenthesis* (PAR). PAR provides metalinguistic information aimed at offsetting imprecision or lack of information that can make the interpretation of the TU difficult. It was described as having a f0 (fundamental frequency) level that contrasts with its neighboring units, higher speech rate, and lower intensity (Firenzuoli and Tucci, 2003) PAR can occur in any position of a pattern, except for the absolute beginning.

We presented different theoretical frameworks that use similar procedures of setting a reference for a certain domain in which gesture and speech are instantiated. The three possible basic units of speech presented, the *prosodic phrase*, the *intonational phrase*, and the TU are neither exactly the same nor are necessarily exclusive; their main difference lies in the theoretical underpinnings that guide the segmentation of speech and functional

[4]The reader is referred to Moneglia and Raso (2014) for the whole framework.

assignment. The tool that will be presented in the next section has the main goal of enabling researchers to go through annotations regardless of their theoretical affiliation and to look for how different reference units can be coordinated.

3. Building a Tool

The kick-off for studying a given corpus is to go through its annotation, which can be rather particular to one's purposes. The annotation itself entails, in many cases, long and laborious manual work requiring tailored analyses. Specific guidelines that serve a handful of studies hinder replicability, as they make it difficult to compare results with other frameworks. This problem is complicated within the framework of multi-modal corpora analysis since many corpora cannot be openly shared due to data protection. Current attempts, such as the RedHen Lab (redhenlab.org/) or the EnvisionBOX (envisionbox.org/) try to achieve replicability by providing pipelines that can be used on different datasets providing comparable results. In line with such initiatives, we present a tool for extracting annotations that allows comparisons of their temporal alignment, based on a reference unit.

This tool was built in Python to automatically extract time marks and annotation values from ELAN files (Wittenburg et al., 2006), using at least two tiers: a reference tier and one or more comparison tiers.[5] The *reference tier* is used to underpin the comparisons, by assigning a particular value to search for overlaps (*searched annotation value*). The tiers used for comparison (*compared tiers*) are then analyzed in relation to the time marks of the reference of this particular value. This tool has the advantage of working with tiers whose types are symbolic (Lubbers and Torreira (2013-2021) does not allow this combination), and it can be executed from multi-platform command lines. A time buffer can be used to fix limits within which overlaps should be deemed synchronous. As indicated in the literature, gestures and speech units may present some degree of displacement and still be considered synchronous cf. (Loehr, 2004; Pouw et al., 2020). The tool outputs a CSV file containing the following columns for all possible co-occurrences:

1. ReferenceTier_value_start_ms: onset of the searched annotation value on the reference tier in milliseconds;
2. ReferenceTier_value_end_ms: offset of the searched annotation value on the reference tier in milliseconds;
3. ReferenceTier_value_duration: duration of the searched annotation value on the reference tier in milliseconds. It is calculated as follows:

$$ref\_tier\_duration = (ref\_tier\_end + buffer) - (ref\_tier\_start - buffer); \quad (1)$$

4. Buffer_ms: assigned buffer time (defaults to zero);
5. Tier: compared tier, ordered first by relation to reference tier then by time;
6. Value: annotation values for each unit in the compared tiers;
7. Begin_ms: onset of the compared tier for each annotation value (listed in item 6);
8. End_ms: offset of the compared tier for each annotation value (listed in item 6);
9. Duration: duration of the compared tier for each annotation value (listed in item 6). It is obtained as follows:

$$comparison\_tier\_duration = comparison\_tier\_end - comparison\_tier\_start; \quad (2)$$

10. Overlap_time: provides the time in milliseconds for all cases in which there is an overlap between the reference tier and the compared tier and is calculated as follows:

$$overlap = \min((ref\_tier\_end + buffer), comparison\_tier\_end) - \max((ref\_tier\_start - buffer), comparison\_tier\_start). \quad (3)$$

It shows all units that can be correlated;

11. Overlap_ratio: provides the proportion of the overlap time in relation to the length of the reference value and is obtained as follows:

$$overlap\_ratio = \frac{overlap\_time}{ref\_tier\_duration}. \quad (4)$$

The result is then rounded to three decimal places;

12. Diff_start: provides the starting time of the comparison tier in relation to the start of the reference value as follows:

$$diff\_start = (ref\_tier\_start - buffer) - comparison\_tier\_start; \quad (5)$$

[5] The code is available at: https://github.com/JorgeFCS/multimodal-annotation-distance

13. Diff_end: provides the ending time of the comparison tier in relation to the end of the reference value as follows:

$$diff\_end = (ref\_tier\_end + buffer) - comparison\_tier\_end. \quad (6)$$

The most important values for the span comparison are (9) to (12), as they indicate the relation between the compared tiers and the reference tier. A scheme for the calculations is provided in Figure 1. It is possible to search for an annotation value in the reference tier. In our case study, for instance, we searched for "PAR", the Parenthesis IU. The reference tier is the upper one and the searched annotation value is in blue. The comparison tiers are the bottom ones, the start difference $diff\_start$ being in orange and the end difference $diff\_end$ in red. The buffer time is in the blue selection. In the absence of a buffer time, the boundaries of the searched annotation value (in dotted line) will be taken. For $diff\_start$, a positive value means that the unit on the compared tier starts before the unit on the reference tier (considering the buffer time). A negative value means that the unit on the compared tier starts after the unit on the reference tier (considering the buffer time). For the $diff\_end$, a positive value means that the unit on the compared tier ends before the unit on the reference tier (considering the buffer time). A negative value means that the unit on the compared tier ends after the unit on the reference tier (considering the buffer time).[6]

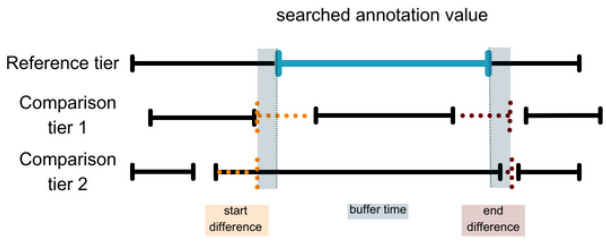

Figure 1: Illustration of the overlap analysis between a reference tier and two comparison ones

Two kinds of comparison can be carried out: span comparison and point comparison. The first is related to the comparison of timespans of each annotation. A *timespan comparison* takes the values of the boundaries of the reference tier and compares it to the onset and offset values of one or more compared tiers. Looking at these boundaries, it is possible to infer whether the comparison tiers start before or after a given tier, how much they overlap (ratio), and the duration of the corresponding tiers.

A *point comparison* searches for the offset of the reference tier and outputs the closest match in the compared tiers. This function is useful to analyze items that do not have a span in time, such as apexes or pitch accents (when considered points in time). When point comparison is set, the tool looks for the offset of the reference value – because most frequently apexes are at the right-most edge of the stroke – and in some annotation guidelines it is the reference used (Rohrer et al., 2022). As points cannot overlap with another unit, this type of comparison outputs only the values (1) to (9).

The advantage of this tool is that it takes protocols that were already used, a reference unit, such as prosodic phrase in Kendon (1972), words in Loehr (2004), focused words in Butterworth and Beattie (1978); Roustan and Dohen (2010); Dohen and Roustan (2017), intonational apexes in Nobe (1996); de Ruiter (1998), prominences in Esteve-Gibert and Prieto (2013); Rohrer et al. (2022), and speech boundaries in (Lelandais and Thiberge, 2023), and compares it with any tiers that are considered relevant (such as gesture phrases in Kendon (1972), pitch accents and apexes in Loehr (2004)). Hypothetically, data annotated with different annotation schemes could be tested thoroughly with only this tool.

4. Case Study

This case study aimed at surveying the relationship between gesture phrases and IUs so as to analyze how they overlap in terms of proportion, and how they are fitted together. To this, we used the BGEST corpus in Brazilian Portuguese (Barros, 2021; Barros and Mello, 2023).

4.1. BGEST Corpus

The BGEST corpus comprises Brazilian Portuguese spontaneous speech texts of ten different speakers, and its total duration is 24 minutes and 28 seconds. The speech annotation was based on the C-ORAL-BRASIL corpus (Raso and Mello, 2012). The first annotation tier includes the TUs, which were perceived as pragmatically and prosodically autonomous (Moneglia and Raso, 2014). The second tier is annotated with prosodic units internal to the TU, among which is the targeted IU (PAR). The third tier contains IU annotation, which was based on the definitions of the L-AcT (Cresti, 2000; Moneglia and Raso, 2014). The gesture annotation is a simplified scheme based on the Linguistic

[6] In Figure 1, the annotation in Comparison tier 1 (in black) is within the boundaries of the searched annotation value (in blue). Its $diff\_start$ (orange) will be negative and its $diff\_end$ (red) will be positive. The annotation in Comparison tier 2 will have a positive $diff\_start$ and a positive $diff\_end$.

Annotation System for Gestures (Bressem et al., 2013), and contains gesture units, gesture phrases, gesture phases, and in the stroke annotation the features of position, hand shape, orientation, and movement type for left and right hands. The corpus annotation files contain 14 tiers each: three of which are related to speech production and 11 to gesture production. In total, the corpus contains 450 strokes and 3984 words (tokens).

The data presented here is restricted to gestures that were coordinated with long PARs, a PAR unit filled with more than a phonological word. Out of the 74 long PARs found in the corpus, 17 were not gesturally mapped. The PAR could be mapped using two main strategies, a change in the pattern that was carried in the utterance (e.g., change in the handshape when starting the PAR) or a suspension of the pattern (e.g., no gesture was made during the PAR, but different gestures are made in the other IUs). The gestures taken into consideration for the quantitative analysis are the ones that show a change in patterning in relation to the neighboring gestures, a total of 54 gestures. It is also important to point out that some gestures spanned across more than one PAR (and they were not excluded from the analysis), which can be better followed in the documentation presented in (Barros, 2021).

The use of the BGEST corpus for our case study is very interesting for two reasons. Firstly, the corpus is annotated in gesture phrases and informational units, which have not yet been fully explored in multi-modal speech and gesture processing. Secondly, the annotation scheme makes it possible to compare gestural and speech events spanning over time, thus enabling the span comparison mode of the proposed tool.

As highlighted in Section 2, the synchronicity rule can be seen as: (i) a prosodic parameter that drives gesture synchronicity (semantic/pragmatic synchronicity derived from phonological synchronicity), e.g. prosodic boundaries that predict the gesture boundaries; or (ii) a semantic or pragmatic reference that grounds the synchronicity between speech and gesture (phonological synchronicity derived from semantic/pragmatic synchronicity), e.g. how IUs can derive the synchronicity of strokes and non-terminal units. We analyzed the latter case, looking first if there was a correlation between overlap ratio and gesture boundaries. This can hint at whether more overlap indicates stronger coordination of gesture and speech units and whether a particular information structure can impact gesture-speech coordination. For instance, if PAR only overlaps with 20% of a gesture, it makes more sense to consider it coordinated with another IU that spreads to PAR due to speech phenomena (hesitations, lengthening, etc.) or physiological phenomena (articulation time of a gesture). To our best knowledge, a scrutiny of the impact of overlap on the synchronicity was not yet made.

4.2. Phonological Synchronicity Derived from Semantic/Pragmatic Synchronicity

For this case study, we analyzed the synchronicity of PAR (in the reference tier) with gesture phrases (in comparison tier 1) and strokes (comparison tier 2) spans (mode of comparison). PAR typically exhibits a prosodic detachment with respect to its hosting TU, making it potentially subject to changes at the gestural level, too. We did not assign a buffer time. The output of the script turned out to be very informative of how gesture phrases were dispersed, centered on the initial and final boundaries of the prosodic units.

In an exploratory analysis, we analyzed the dispersion of on- and offsets for gesture phrases and phases using median and standard deviation. The reported values are listed in the Table 1. The table contains the median (M) and standard deviation (SD) for centered start times (CST) and centered end times (CET) for gesture phrases and the annotations therein – preparation, stroke, retraction, and hold.

Table 1 shows that the gesture phrase tends to start before and to end after the prosodic unit (respectively positive for CST_M and negative CET_M values), i.e., the gesture phrase seems to be longer than the target unit (PAR). The consolidated measures show that the phases (preparation, stroke, hold, and retraction), on the other hand, seem to have the opposite pattern (with negative and positive medians respectively). This indicates they tend to occur within the span of the target unit. However, considering the standard deviations (all above 640 ms), we see that these tendencies are not representative of the whole data. In many cases, as we show, the phases can begin before the target prosodic unit. To give a dimension of the relevance of these SD values, PAR's median duration is 1250 ms (SD = 547.984). This indicates that there is variation that can span up to half of the analyzed unit and three times the value of 200 ms normally assumed in the literature (Loehr, 2004; Pouw et al., 2020; Barros, 2021) as a degree of mismatch between a prosodic and a gestural event.

The next step was to check how the compared units were associated. We did this by looking at the overlapping ratios calculated by the tool. This association can be implied by a smaller dispersion of start differences and end differences: if PAR drives the synchronicity, one should expect that there is a positive correlation between a higher overlap ratio and smaller start differences. Small overlap ratios can have a broader dispersion regarding start dif-

	Occurrences	CST_M	CST_SD	CET_M	CET_SD
gesture phrase	131	306	1178.8	-266	1117.4
preparation	55	-190	731.2	151	655.0
stroke	85	-2	891.3	278	948.3
hold	14	-425	754.9	333	641.6
retraction	46	-386.5	665.3	157.5	667.9

Table 1: Dispersion table on gesture phrase and phases, without buffer

ferences. This may function as an indication that the gesture (phrase or stroke) is coordinated with IUs other than PAR. The overlap can also inform us in what order gestures and prosodic units occur. According to (Loehr, 2004), the start of gesture phrases works as a kick-off followed by the prosodic event. The expected order of events is thereafter the stroke, the end of the IU, and the end of the gesture phrase. This is in line with the values shown in Table 1: positive start difference and a negative end difference for gesture phrases, and the opposite for strokes. We must thus check whether this tendency is clearer for higher overlap ratios. Here, our hypothesis is that gesture phrases beginning before PAR will have higher overlap ratios.

In Figure 2, the x-axis is the overlap ratio and the y-axis is the centered start difference. Positive values indicate that the gesture event (gesture phrase or stroke) starts before PAR. Negative values indicate that the gesture event starts after PAR's onset. Points on the black horizontal line (0 value) indicate that the starting points of the gesture phrase (upper graph) or stroke (lower graph) perfectly match the PAR's onset. The blue vertical line indicates when the overlap has a 50% ratio. The overlap ratio is color-coded by quartiles: purple represents the second quartile and the third quartiles; and grey stands for the first and fourth quartiles.

The trend we can observe in both graphs is the same: the earlier the gesture phrase or the stroke begins, the higher the overlap ratio (and not the contrary). Assuming that the amount of overlapping is a good predictor for the coordination between the gestural and the prosodic-informational events (as argued in Section 2), we can conclude that the trend observed in the literature holds true: the gesture begins earlier and signals a change in the pragmatic level, i.e. the introduction of PAR. This trend seems to hold for both gesture phrases and strokes in a complete overlap: both gesture phrase and stroke start before PAR when there is 100% overlap, as shown in the first quadrant of both graphs. Both graphs also indicate that there is dispersion regarding CSTs that is by far higher than Loehr (2004) buffer times, being the second quartile of gesture phrases within -563 ms and 1125 ms, and for strokes within -485 ms and 414 ms. Assigning a buffer time would mislead the interpretation. Here, overlap seems to be more revealing than buffer times.

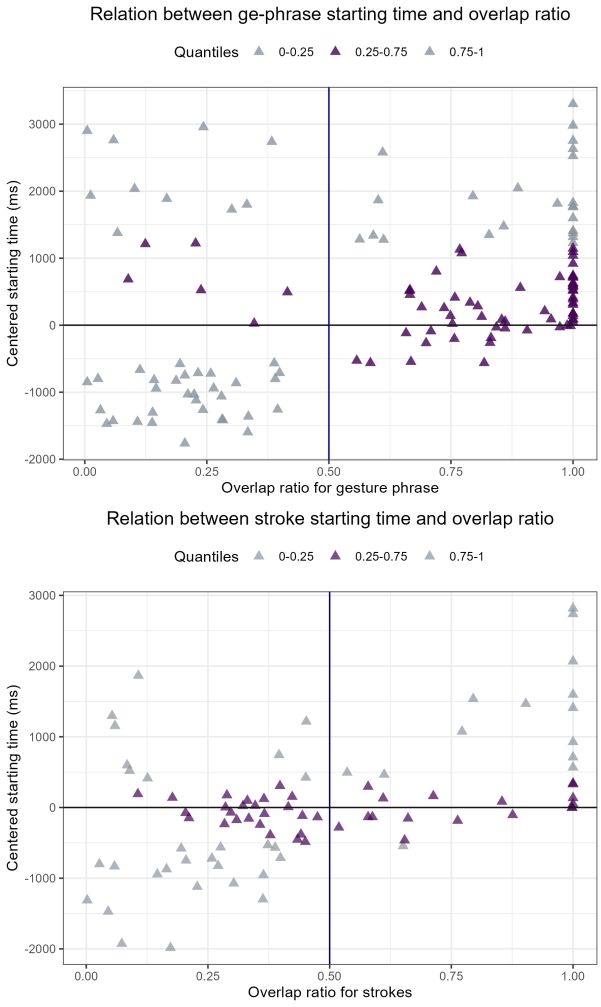

Figure 2: Dispersion of starting points in relation to overlap ratio

For strokes, a more in-depth qualitative analysis is needed as the amount and placement of the overlap can impact the pragmatic interpretation (Ebert et al., 2022). Quantitative analysis shows us that strokes do seem to be more evenly distributed around smaller starting times. None of the graphs indicates a strong correlation between the overlap ratio and start time difference. We want, thus, to check what happens to the onset alignment of

PAR and the gesture-phrase containing the stroke. It would be reasonable to think that if a gesture-phrase is more overlapped with PAR, so it would be with its respective stroke. Figure 3 shows this relation.

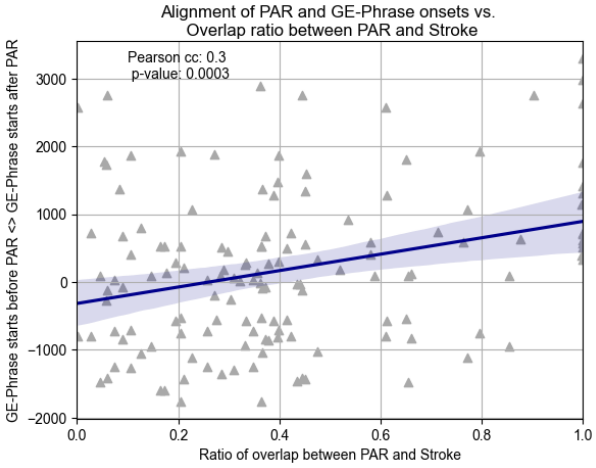

Figure 3: Alignment of PAR and GE-Phrase onsets vs. Overlap ratio between PAR and Stroke

As can be observed, there is still a weak correlation between the stroke overlap ratio (with respect to PAR) and the onset alignment of PAR and the gesture phrase. Although the direction of the correlation is in line with our expectations, its strength is not. As we have seen in Table 1, gesture phrases tend to begin before and end after the overlapping PAR unit. We can hypothesize that in many cases, the stroke will tend to align more with the PAR offset, crossing the boundary to the following IU. We can check whether this trend holds by looking at Figure 4.

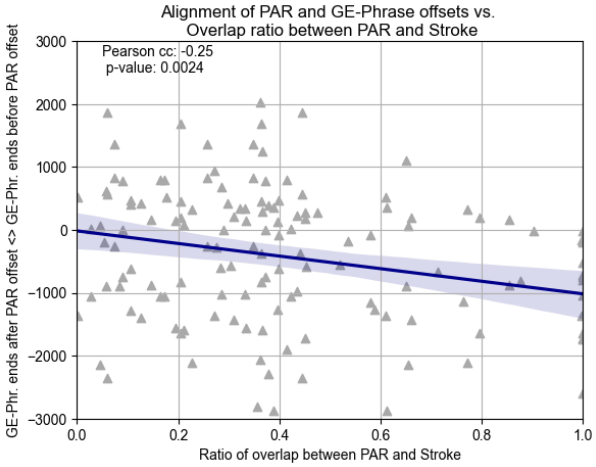

Figure 4: Alignment of PAR and GE-Phrase offset vs. Overlap ratio between PAR and Stroke

Here, gesture-phrases ending after PAR's offset will tend to have strokes more overlapped with the target IU. The dispersion is much more concentrated on the second quadrant, and all strokes that exhibit total overlapping are within gesture phrases that end after PAR. On the other hand, strokes with lower overlap ratios are dispersed within the third and fourth quadrants, confirming our hypothesis: the stroke is not perfectly aligned with PAR, but there is a trend for gesture phrases beginning before and ending after PAR to exhibit strokes more overlapping. Notice that this is not a necessary consequence. If gesture phrases began after the onset and ended before PAR's offset, the overlap ratio should necessarily be 100%. We deem it possible that the gesture-phrases that begin after PAR's onsets are actually not coordinated with PAR but with the following IU. To be sure of that, we would need to perform a qualitative analysis of the data (looking for pairings gesture-PAR) and check the overlap ratios with IUs preceding and following our target unit. This is, however, out of our scope.

A question remains: how much overlapping is enough to indicate a coordination between a gesture-phrase and the IU? For the gesture-phrase, the answer seems to be straightforward: 50% of the overlap ratio seems to separate almost all gesture-phrases that start before PAR's onset (1st quadrant of Figure 2). For the stroke, this pattern is not very clear. Here, it would be interesting to evaluate the spanning alignment of the stroke with a lower-level prosodic event. For instance, we could examine the alignment of strokes (or the central point thereof) with the prosodic nuclei of, say, illocutions or focuses in other kinds of IUs. This could also be useful to explain why apexes tend to be associated with pitch accents. In any case, this does not indicate a need for a time buffer to posit synchronicity.

As a final word, it is possible to say that PAR seems to motivate some coordination between boundaries of gesture and speech, which means that the assumption that phonological synchronicity is derived from semantic/pragmatic synchronicity might hold. Cases in which the synchrony is less relevant remain unclear, such as how the three synchrony rules are entangled. Another important consideration is that although the literature points towards the need for some sort of buffer time to analyze speech-to-gesture coordination, all the data presented so far did not have any buffer. We argue here that the dispersion of boundaries is better explained in relation to the overlap ratio rather than with a buffer time, which may bring in some arbitrariness in the analysis.

5. Conclusion

This paper provides a tool that gathers annotation from ELAN files automatically and enables an ex-

ploratory quantitative analysis, regardless of the annotation scheme. In a case study for the BGEST corpus within the framework of the L-AcT, we argue that the overlap ratio, in opposition to a buffer time, is a useful indicator of coordination between gesture and prosodic boundaries. It was also highlighted that strokes and non-terminal boundaries seem to have a more direct link in their coordination than the latter and gesture phrases.

The coordination of boundaries is not yet fully understood to posit buffer times that are truly informative of this relationship. As such, this paper emphasizes the need for a more fine-grained definition of the buffer and to which contexts it applies, rather than rigid time assumptions.

The main limitations that must be taken into consideration include that annotations should be seen critically (they are prone to biases) and that there is still no synchronicity measure that can be used for this scenario.

6. Acknowledgements

This work was partially funded by the Service-Oriented Systems Engineering Research School of the Hasso Plattner Institute. This work is supported by the "ADI 2020" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02, and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Programa de Excelência Acadêmica (PROEX) - Brasil.

7. Bibliographical References

J.L. Austin. 1962. *How to do things with words*. William James Lectures. Oxford University Press.

C. A. Barros. 2021. A relação entre unidades gestuais e quebras prosódicas: o caso da unidade informacional parentético. Unpublished MA Thesis at Universidade Federal de Minas Gerais.

C. A. Barros and H. Mello. 2023. The c-oral-brasil proposal for the treatment of multimodal corpora data: the bgest corpus pilot project. In Andrés Grajales Ramírez, Jorge Molina Mejía, and Pablo Valdivia Martin, editors, *Digital Humanities, Corpus and Language Technology: A look from diverse case studies*. University of Groningen Press.

D. Bolinger. 1978. Intonation across languages. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravsick, editors, *Universals of human language*, volume 2 of *Language science and national development series*, pages 471–524. Polity Press, Stanford.

J. Bressem, S.H. Ladewig, and C. Müller. 2013. 71. linguistic annotation system for gestures. In C. Müller, A. Cienki, E. Fricke, S.H. Ladewig, D. McNeill, and S. Tessendorf, editors, *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 38/1*, pages 1098–1124. DE GRUYTER.

B. Butterworth and G. Beattie. 1978. Gesture and silence as indicators of planning in speech. In Robin N. Campbell and Philip T. Smith, editors, *Recent Advances in the Psychology of Language*, pages 347–360. Springer US, Boston, MA.

W. Chafe. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Advances in Discourse Processes. Bloomsbury Academic.

E. Cresti. 2000. *Corpus del italiano parlato*. Accademia della Crusca, Firenze.

J.P. de Ruiter. 1998. *Gesture and speech production*. Ph.D. thesis, Radbound University Nijmegen.

M. Dohen and B. Roustan. 2017. Co-production of speech and pointing gestures in clear and perturbed interactive tasks: Multimodal designation strategies. In *Interspeech 2017*, pages 166–170, ISCA. ISCA.

C. Ebert, G. Pirillo, and S. Walter. 2022. The role of gesture-speech alignment for gesture interpretation. In *Proceedings of Linguistic Evidence 2020: Linguistic Theory Enriched by Experimental Data*.

N. Esteve-Gibert and P. Prieto. 2013. Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research*, 56(3):850–864.

V. Firenzuoli and I. Tucci. 2003. L'unità informativa di inciso: correlati intonativi. In *La coarticolazione*, Collana degli atti dell'Associazione italiana di acustica, pages 185–192, Pisa. ETS.

. Izre'el, H. Mello, A. Panunzi, and T. Raso, editors. 2020. *In Search of Basic Units of Spoken Language: A corpus-driven approach*. John Benjamins.

A. Kendon. 1972. Some relationships between body motion and speech. In Aron Wolfe Siegman and Benjamin Pope, editors, *Studies in dyadic communication*, Pergamon general psychology

series, pages 177–216. Pergamon Press, New York.

A. Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press, Cambridge and New York.

M. Lelandais and G. Thiberge. 2023. The role of prosody and hand gestures in the perception of boundaries in speech. *Speech Communication*, 150:41–65.

D.P. Loehr. 2004. *Gesture and intonation*. Ph.D. thesis, Georgetown University, Georgetown.

M. Lubbers and F. Torreira. 2013-2021. pympiling: a python module for processing elans eaf and praats textgrid annotation files. https://pypi.python.org/pypi/pympi-ling. Version 1.70.

E. McClave. 1994. Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, 23(1):45–66.

D. McNeill. [1992] 1995. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, Chicago.

M. Moneglia and T. Raso. 2014. Appendix: Notes on the language into act theory. In T. Raso and H. Mello, editors, *Studies in Corpus Linguistics*, volume 61, pages 468–495. John Benjamins Publishing Company, Amsterdam.

C. Müller. 2018. Gesture and sign: Cataclysmic break or dynamic relations? *Frontiers in Psychology*, 9.

S. Nobe. 1996. *Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network / threshold model of gesture production*. Unpublished doctoral dissertation, University of Chicago, Chicago.

J.B. Pierrehumbert. 1980. *The phonology and phonetics of English intonation*. Thesis (ph.d.), Massachusetts Institute of Technology.

W. Pouw, S. Harrison, N. Esteve-Gibert, and J. Dixon. 2020. Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *The Journal of the Acoustical Society of America*, 148(3):1231.

T. Raso and H. Mello, editors. 2012. *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Editora UFMG, Belo Horizonte.

A. Rochet-Capellan, R. Laboissière, A. Galván, and J.L. Schwartz. 2008. The speech focus position effect on jaw-finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, 51(6):1507–1521.

P.L. Rohrer, P. Prieto, and E. Delais-Roussarie. 2022. Le rythme prosodique guide le rythme gestuel. In *XXXIVe Journées d'Études sur la Parole – JEP 2022*, pages 588–596, ISCA. ISCA.

B. Roustan and M. Dohen. 2010. Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus. In *Speech Prosody 2010 - 5th International Conference on Speech Prosody*, pages 1–4. ISCA Archive.

J. Streeck. 2006. Pragmatic aspects of gesture. In E. Keith Brown, editor, *The encyclopedia of language & linguistics*, pages 71–76. Elsevier, Amsterdam and Boston.

R. Tomasello, L. Grisoni, I.P. Boux, D. Sammler, and F. Pulvermüller. 2022. Instantaneous neural processing of communicative functions conveyed by speech prosody. *Cerebral Cortex*.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of LREC 2006*, pages 1556–1559. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen. [Online; accessed 2020-03-07].