Listeners integrate speech, gesture, and discourse structure to interpret the temporal structure of complex events

Andie Nishimi[1], **Esther Walker**[2] , **Benjamin K. Bergen**[2]
{anishimi, e1walker, bkbergen}@ucsd.edu
[1]Psychology and [2]Cognitive Science, University of California, San Diego

Tyler Marghetis
tmarghet@indiana.edu
Psychological and Brain Sciences, Indiana University, Bloomington

Abstract

Human communication has a remarkable capacity to describe events that occurred elsewhere and at other times. In particular, when describing complex narratives, speakers must communicate temporal structure using a mixture of words (e.g., "after"), gestures (e.g., pointing rightward for a later event), and discourse structure (e.g., mentioning earlier events first). How do listeners integrate these sources of temporal information to make sense of complex narratives? In two experiments, we systematically manipulated gesture, speech, and order-of-mention to investigate their respective impacts on comprehension of temporal structure. Gesture had a significant effect on interpretations of temporal order. This influence of gesture, however, was weaker than the influence of both speech and order-of-mention. Indeed, in some cases, order-of-mention trumped explicit descriptions in speech; for instance, if 'earlier' events were mentioned second, they were sometimes thought to have occurred second. Listeners integrate multiple sources of information to interpret what happened when.

Keywords: time; gesture; iconicity; multimodal communication; memory.

Introduction

Human communication stands out among naturally occurring communication systems in its ability to convey information about events occurring in other places and at other times, a feature known as displacement (Hockett, 1960). This includes the concrete details of displaced events—who did what to whom—but also when things occurred. If you observed a woman receiving the winning lottery ticket and also getting her purse stolen, then you would want to be clear about which event occurred first. While temporal order is an abstract feature of a complex event, it is often critical for communicative success.

To communicate about temporal order (and to communicate in general), speakers have several strategies to deploy. The first and most obvious is in their choice of words, like "before" or "after," "earlier" or "later." Second, speakers also communicate about temporal order using visible and systematic motion of their bodies (Cooperrider & Nunez, 2009; Casasanto & Jasmin, 2012). Spontaneous co-speech gestures produced by North-American native English speakers often indicate relative temporal order by locating events along an imagined spatial timeline, with earlier events placed to the left and later events placed to the right. Finally, speakers encode temporal order in the structure of their larger discourse. Earlier events are typically expressed earlier in an utterance, while later events are expressed later ("order-of-mention," a.k.a. temporal iconicity, Jakobson, 1971). For instance, if somebody *first* went to the gym and *then* stopped for coffee, it would be most natural for them to say, "I went to the gym and stopped for coffee," rather than the reverse; the order in which the events are mentioned can stand in for the order in which they occurred.

During real-world communication, all three of these strategies can be deployed at the same time, complementing each other. For instance, if a speaker were to describe a series of events that occurred on a recent vacation, they might use expressions like "first," "and then," and "finally" to express explicitly, using lexical resources, the temporal order of events. In coordination with these expressions, they might point along the left-to-right spatial axis to convey the temporal order of the events. And, at the same time, they might choose to describe the events in the same order in which they actually occurred.

While we know that speakers *do* this, less is known about whether listeners actually care. Temporal terms are notoriously hard for children to acquire (Tilman et al., 2017; Shatz et al., 2010); the words "before" and "after," for instance, continue to be confused by most children until they are five years old (Clark, 1971). Listeners are also known to rely on order-of-mention to infer the order in which events occurred (Jakobson, 1971), although past work has focused primarily on contexts where temporal order is ambiguous in speech and gesture.

It's also currently unknown whether listeners rely on temporal gestures to make inferences about the abstract concept of temporal order. By contrast, listeners are demonstrably sensitive to concrete information expressed in gesture. For example, concrete, iconic gestures boost comprehension (Thompson, Driscoll & Markson 1998) and can even add information not otherwise present in speech (Church et al. 2007; Singer & Goldin-Meadow 2005). Less is known, however, about the communicative impact of gesture on the interpretation of temporal structure. Speakers use gesture to express a range of information about time, including duration and sequential order (Cooperrider & Núñez, 2009). There is mixed evidence that, when speech is ambiguous (e.g., 'the meeting was moved forward,' which can mean earlier or later), observers use gesture to determine how the speaker's metaphorical conceptualization of time (Jamalian & Tversky, 2012), although perhaps only when communication is co-present and not computer-mediated (Lewis & Stickles, 2016). As far as we know, no previous research has investigated whether gestures about *temporal order* are actually communicative.

What about when these sources of information are not aligned but contradictory? Sometimes, we mention a later event first, perhaps because of the event's salience. When this occurs, the conflict is reflected in the listener's neural response to the sentence as they resolve the conflict (Münte et al, 1998). Previous research has assumed that, in cases of conflict, speakers will default to the information expressed lexically, overriding the temporal order suggested by order-of-mention. However, discourse comprehension involves probabilistic judgments about how best to integrate potentially contradictory information (e.g., Gibson et al, 2014). Under some circumstances, therefore, speakers are likely to rely on order-of-mention, overriding or ignoring the temporal order conveyed explicitly in speech.

To investigate the communication of complex temporal structure during multimodal discourse, we conducted two studies in which we systematically manipulated how information about temporal order was expressed in speech, gesture, and order-of-mention. Participants viewed brief videos in which a speaker described a complex series of events. Descriptions varied in the use of explicit temporal terms (e.g. "earlier" or "later") and temporal gesture (e.g., a leftward pointing gesture to indicate an earlier event) to order the events in the sequence. Within these descriptions, moreover, pairs of events were sometimes mentioned in the same order as they occurred, so that order-of-mention was a helpful guide to temporal order, but other times the order-of-mention did not align with their actual temporal order. All three sources of information—temporal terms, temporal gesture, and order-of-mention—were thus fully crossed within subjects.

We foresaw a number of possible outcomes. On the one hand, temporal terms are so explicit and unambiguous that they may overwhelm information from any other source, including gesture and order-of-mention. On the other hand, a complex situation can involve multiple interrelated events, outstripping the relatively simple binary distinctions that are most common in speech (before/after, earlier/later). Under these circumstances, temporal gestures may be especially beneficial, as they allow a listener to track the relative ordering of multiple events. A series of three gestures, for instance, can use relative spatial location to order events without any of the ambiguity that can plague speech. Lastly, order-of-mention may sometimes trump both speech and gesture. First, it uses time (of mention) to represent time (of occurrence)—a direct mapping that may be difficult for a listener to ignore when constructing their discourse model. Second, we know that memory for specific words isn't great as a delay period increases (Sachs, 1967)—when listeners are trying to reconstruct the order of events after the fact, the may be more likely to recall the order in which they were mentioned than the specific words used to describe their order. Thus, there are good reasons to predict that all three sources of information may dominate listener's interpretations of multimodal communication about temporal order.

Experiment 1

The main purpose of Experiment 1 was to investigate how participants use the temporal information available in multiple communicative resources to construct a temporal narrative of events. We were especially interested in how individuals reconcile situations in which different sources provide conflicting temporal information. And finally, because of past work suggesting that gesture effects on comprehension are amplified over a delay (Church et al. 2007), we added a Memory Condition (*Immediate* or *Delayed*) to investigate whether the temporal resources participants use to order a sequence of events changes over a delay.

Methods

Participants: Forty undergraduate students (N = 31 female) participated in this study in exchange for course credit. Sample size was determined in advance on the basis of similar studies of gesture (e.g., $N = 45$ in Church et al. 2007).

Materials: We filmed 16 vignettes in which a woman narrated a brief story consisting of four events. The events in a given sequence had all already occurred or all were going to occur. That is, half of the vignettes discussed future events (i.e. planning a hiking trip,

preparing to go camping) while the other half of the videos discussed a sequence of events that had already occurred (i.e. recalling a vacation, recapping a day at work).

Videos were filmed from the neck downwards, ending at the top of the narrator's legs, with the arms clearly visible. Because we manipulated whether or not temporal information is delivered through explicit temporal terms in speech, we cut the narrator's head and neck out of the frame to avoid giving participants visual clues (i.e. voice box movements) from which to draw temporal information. All of the video stimuli were generated originally and contained arbitrarily related events to ensure that participants could not determine the temporal order of events by relying on causal or canonical relationships (i.e. hiking up a cliff generally precedes jumping down a waterfall).

Procedure: Participants watched short video clips that described a four-event vignette. For example, a participant could hear, about an upcoming climb up Mount Everest, that "I should probably replace my old hiking boots and then pick up some snow gear for encountering snowy conditions. I also cannot forget to first get a hiking permit and then purchase an airline ticket to Nepal." In this case, most of the clauses contain explicit temporal terms that disambiguate the actual temporal order. Within each sentence, order of mention also indicates the correct temporal order (e.g., the narrator intends to get a hiking permit before purchasing an airline ticket to Nepal); by contrast, the two sentences' order of mention conflict with the order in which the events occurred. There were sixteen different vignettes in total. The videos of each vignette were randomly presented, and participants saw each of the unique vignettes one time. The video of each vignette, however, was played twice, back to back, to the participant.

Participants also completed a set of seven comprehension questions after viewing a video stimulus. For half of the vignettes, participants received the comprehension task immediately following the presentation of that particular vignette video (our *Immediate* memory condition). For the other half of the vignettes, participants received the corresponding set of comprehension questions following a 10 minute delay (our *Delayed* memory condition). During this 10 minute delay, participants completed multiplication and long division problems.

Each comprehension question was presented in 2AFC format (i.e. "Do I need to buy more winter gear before or after getting an airline ticket?") with a 10 second response window. Four of these seven questions tested the temporal relationship of events in the story (*target questions*). The remaining three questions in the comprehension set were unrelated to temporal content and probed the basic content of each video (*filler questions*). Question order was randomized for each video and for each participant. At the end of the experiment, participants filled out a debrief questionnaire.

Analysis

Before analyzing the data, we removed filler questions, trials with a response time faster than 200 ms, and trials that were two and a half standard deviations faster or slower than each participant's mean response time on each vignette. We excluded participants whose accuracy on the comprehension task was below 50% (chance) when considering trials where temporal term, gesture information, or both were present, as these individuals were below chance performance even when explicit ordering information was available to them. We also excluded responses for participants who failed our debrief point-of-view item. In this question participants were asked, "Which of the following gestures would the narrator use to accompany the word 'earlier'?" They were given two short video clips, one with the narrator making a rightward (from her point of view) gesture stroke, and one with her making a leftward gesture stroke from which to respond. Participants who chose the rightward gesture stroke–which appears as a leftward stroke from their mirrored perspective–as accompanying the word "earlier," were considered to have failed the debrief and were excluded from this analysis since we wanted to ensure they were interpreting the gestures in the videos the way we intended.

Our primary dependent measure was participants' response (before vs. after), as a function of the information expressed in order-of-mention, speech, or gesture (before, nothing, or after). When neither speech nor gesture include explicit temporal information, there is no 'ground truth' about the order of events. Each resource was dummy coded for its temporal information ("before" = -1, no info = 0, "after" = 1). All analyses used generalized mixed-effects models with a logistic link function, with centered predictors and the maximal converging effects structure justified by the design (Barr et al, 2013).

Results

Effects of language, gesture, and order-of-mention

We first examined how comprehension of temporal sequences is affected by temporal terms in speech, temporal gesture, and order-of-mention. All three sources of information had a significant effect. Temporal terms reliably influenced participants' interpretation of temporal order ($b = 0.95 \pm 0.17$ SEM, $z = 5.45$, $p < 0.001$), with more *before* interpretations after the use of the word "before" ($M = 77\%$) but more *after* interpretations after the use of the word "after" ($M = 72\%$). Similarly, order-of-mention had a significant, if smaller, impact ($b = 0.72 \pm 0.15$ SEM, $z = 4.94$, $p <$

0.001), with more *before* interpretations when the event was mentioned first, but more *after* interpretation when the event was mentioned second. And temporal gestures, too, had a significant impact (b = 0.48 ± 0.12 SEM, z = 3.94, p < 0.001), with more *before* responses after leftward "past" gestures (*M* = 63%), and more *after* responses are rightward "after" gestures (*M* = 67%). Participants thus were sensitive to all three of the semiotic resources available during multimodal comprehension, with larger influences of temporal terms and order-of-mention, and a smaller but significant influence of gesture.

The only other significant effect we observed was an interaction between temporal terms and temporal gesture (*b* = -0.38 ± 0.16 SEM, *z* = -2.21, *p* = 0.0271). This was driven by a large effect of gesture when temporal terms were absent entirely from speech (*b* = 0.54 ± 0.17 SEM, *z* = 3.19, *p* = 0.0014), but a much smaller effect of gesture when accompanied by explicit temporal terms (b = 0.72 ± 0.21 SEM, z = 3.47, p = 0.00541).

Effects of recall

We were also interested in the effect a delay period would have on participants' comprehension of temporal events. Specifically, we wondered whether we would see the effects of particular resources (i.e. temporal gesture) strengthen over time, as previous research with iconic gesture has found (Church et al., 2007). Our results indicated that participants actually performed the same on the comprehension task regardless of whether it was completed immediately following the video vignette or after a 10 minute delay period. We did not find evidence of interactions between any of the resources and our memory condition factor.

Discussion

Our study aimed to investigate how people draw on and integrate multiple sources of temporal information during comprehension of complex temporal sequences. We found that participants are independently influenced by the information available through temporal terms, gesture, and order-of-mention.

The presence of temporal terms and temporal gesture each influences participants to respond according to the order presented through these resources perhaps in explicitly conveying temporal order. Order-of-mention is also largely influential as a listener builds a temporal narrative, perhaps because of the salience of the iconicity (i.e. letting the order events are uttered in speech stand in for the order events occur in time).

Interestingly, we found an interaction between temporal terms and gesture, mediated by whether or not information from temporal speech is present—when ordering information from temporal terms is already present, we see less of an impact of temporal gesture than when it is absent.

We were additionally surprised to not detect an effect of memory condition given our predictions that the effects of gesture are strengthened over time. Perhaps our delay was not long enough to elicit a difficult recall situation, in which temporal ordering information would decay over time. Creating that kind of recall situation is important to reveal any effects that our temporal resources may selectively provide over time.

Experiment 2

Experiment 2 was designed to replicate and extend slightly the results of Experiment 1. The slight extension was to address our unexpected finding that the relative impact of gesture did not differ after a delay. This appeared to contradict previous evidence that the impact of gesture increases with the passage of time (Church et al, 2007). Based on this, we predicted that, as more time passed after observing an utterance, the relative contributions of explicit terms, gesture, and order-of-mention should change—with, in particular, an increased reliance on gesture.

However, participants' recall was not severely impacted after the delay, suggesting that this delay may not have been sufficiently long enough to observe a shift in importance between temporal terms, temporal gesture, and order-of-mention. We thus increased this delay from 10 minutes to 30 minutes.

Methods

Participants: Adults (*N* = 50, 33 women) participated in exchange for partial course credit.
Materials: The same as in Experiment 1.
Procedure: The same as in Experiment 1, except that we extended the delay period from 10 to 30 minutes.

Analysis

All exclusionary criteria and data cleaning procedures used in Experiment 1 were also applied for Experiment 2. Analyses again used logistic mixed-effects models, with centered predictors and the maximal converging effects structure justified by the design (Barr et al, 2013).

Results

Effects of language, gesture, and order-of-mention

Experiment 2 replicated the main findings of Experiment 1. Participants reliably drew on information presented through temporal terms (b = 0.86 ± 015 SEM, z = 5.68, p < 0.001), with more *before* interpretations after the use of the word "before" (*M* = 72%) but more *after* interpretations after the use of the word "after" (*M* = 71%). They also use the information available via order-of-mention (b = 0.62 ± 0.12 SEM, z = 5.00, p < 0.001), with more *before* interpretations when the event was

mentioned first, but more *after* interpretation when the event was mentioned second. And similarly, participants also rely on the information present in temporal gesture (b = 0.28 ± 0.14, z = 2.00, p < 0.05), with more *before* responses after leftward "past" gestures (*M* = 62%) but more *after* responses are rightward "after" gestures (*M* = 59%). These findings replicate the results of our previous experiment: participants are sensitive to the information from temporal terms and order-of-mention in particular, and less reliably applying the information gleaned from gesture (Fig. 1).

We did not, however, replicate the interaction of temporal terms and temporal gesture (*b* = -0.10 ± 0.14, z = -0.77, *p* = 0.44).

Effects of recall

We next turned our attention to the effect of the delay period, and how the available temporal resources would be deployed over time. Our results did not reveal an effect in line with our prediction that the impact of gesture would increase over time (Gesture x Memory Condition, b = 0.27 ± 0.17 SEM, z = 1.47, p = 0.14), even with a longer delay period.

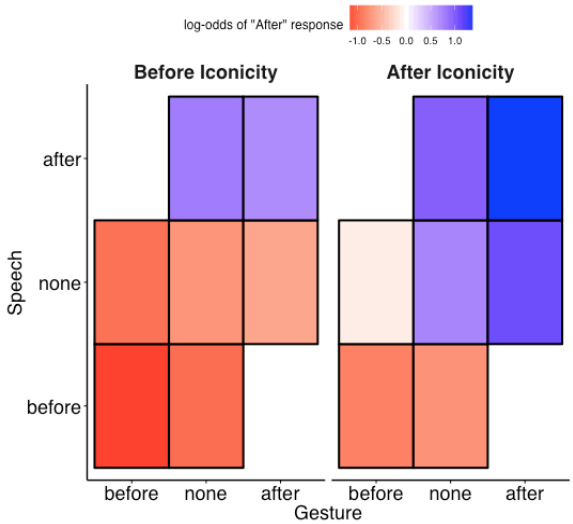

Figure 1. Impact of each resource on participants' response. Color indicates the log-odds of interpreting an event to have occurred after (vs. before). All three resources had an impact. As gesture went from expressing 'before' to 'after' (i.e., moving rightward), 'after' responses increased (i.e., shift from red to blue). Similarly, as speech went from 'before' to 'after' (ie., bottom to the top), 'after' responses increased. And when order-of-mention suggested that the event occurred after (i.e., right panel), there was a higher proportion of 'after' responses (i.e., shift toward blue).

Discussion

Experiment 2 replicated the main findings of Experiment 1. We found that in multimodal communication, participants reliably glean information from explicit temporal terms in the utterance, order-of-mention, and to a lesser extent temporal gesture. The only effect that did not replicate was the interaction between gesture and temporal terms, which suggests that this effect is either small and fickle or potentially a false-positive from Experiment 1.

Even with a 30-minute delay, the benefits of these metaphorical temporal gestures did not increase with the passage of time, unlike past findings for concrete representational gestures (Church et al., 2007). One explanation is that 30 minutes is insufficient to elicit the selective benefits of gesture.

General Discussion

We set out to investigate how we communicate about the temporal structure of complex events. Multimodal communication offers a range of resources for expressing temporal order: words, gestures, discourse structure. Across both studies, we found that listeners made use of all three of these sources of information, integrating them to make sense of the temporal structure of complex narratives.

The ephemeral and spatial nature of gesture

A central finding of these studies is that gestures that encode temporal order are genuinely communicative. Gestures are ephemeral, disappearing as soon as they are produced, and are only intermittently interjected into the the speech stream. Despite this, listeners made reliable use of gesture to interpret complex narratives.

Perhaps temporal gesture is especially useful in that it can help create a schematic "bird's eye" view of a complex event by laying out all of the events in their temporal order. A single temporal gesture can depict a pairwise relation between two events by placing them in space — but a sequence of gestures can construct a schematic representation of an entire narrative, including multiple subevents. By enacting a spatial timeline, temporal gestures supply an object for joint attention, available to both speaker and listener.

The utility of gesture may depend on the listener's perspective on the speaker. Because time recruits lateral space, assigning meaning to the right (*future*) and left (*past*) sides of space, interlocutors who face each other are confronted with an added challenge: adopting the perspective of their partner. The fact that participants in our study were able to interpret the speaker's lateral temporal gestures — despite the fact that the speaker was both head-on and video recorded — is a testament to the centrality of gesture to human communication.

Integrating the complementary and contradictory

While we began by considering reasons that one source of information might dominate listeners' comprehension of temporal order, both studies found that all three sources of information make independent and reliable contributions. While the impact of gesture was less pronounced than that of speech or discourse structure, it was nonetheless robust across both studies. All three sources of information appear to make independent contributions to the interpretation of temporal structure.

One avenue for future research is whether there are individual differences in the reliance on these sources of information. Are some listeners especially sensitive to *when* an event is mentioned in the discourse, while others are more sensitive to how that event is *gestured* relative to others? If such individual differences exist, and we suspect they do, then these may lead to radically different interpretations when different sources of temporal information come into conflict — when the first event mentioned actually occurred last, or when the speaker gestures *left* but accidently says 'and then afterward….' Understanding these dynamics will help us understand how miscommunications occur — and are repaired.

Similarly, these three sources of information may have differential impacts in different communicative settings or under different task demands. For instance, Lewis and Stickles (2016) reported that gestures expressing metaphors for time were communicative—but only when the speaker was co-present with the listener, rather than appearing on video. Gestures expressing temporal order may also decrease in importance during video-mediated communication, with speech and order-of-mention weighted more heavily. This may account for gesture's relatively smaller effect size in the current studies.

Conclusion

We began by asking how listeners understand the temporal structure of complex narratives by integrating information from various sources: words, gesture, order-of-mention. We found that each of these resources made independent contributions to the comprehension of temporal order. In particular, these results demonstrate that metaphorical gestures can communicate complex temporal relations. The power of human communication may lie in its use of multiple strategies to communicate abstract information.

Acknowledgments

Thanks to R. Núñez and M. Kutas for helpful feedback.

References

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in psychology*, *4*, 328.

de Hevia, M. D., Izard, V., Coubart, A., Spelke, E. S., & Streri, A. (2014). Representations of space, time, and number in neonates. *PNAS*, 111, 4809-4813.

Casasanto, D., & Jasmin, K. (2012). The hands of time: Temporal gestures in English speakers.

Church, R., Garber, P., Rogawlski, K. (2007). The role of gesture in communication and social memory. *Gesture*, 7(2), 137-158.

Clark, E. V. (1971). On the acquisition of the meaning of before and after. *Journal of verbal learning and verbal behavior*, 10, 266-275.

Cooperrider, K., & Núñez, R. (2009). Across time, across the body: Transversal temporal gestures. *Gesture*, *9*(2), 181-206.

Gibson E., Bergen & Piantadosi (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *PNAS*, 110, 8051-8056.

Hockett, C. D. (1960). *The origin of speech*. Freeman.

Jakobson, R. (1971). Language in relation to other communication systems. *Selected writings*, *2*, 570-579.

Jamalian, A., & Tversky, B. (2012). Gestures alter thinking about time. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Lewis, T. N., & Stickles, E. Gestural modality and addressee perspective influence how we reason about time. *Cognitive Linguistics*.

Münte, T., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, 395, 71-73.

Sachs, J. S. (1967). Recogpition memory for syntactic and semantic aspects of connected discourse. Attention, Perception, & Psychophysics, 2(9), 437-442.

Shatz, M., Tare, M., Nguyen, S. P., & Young, T. (2010). Acquiring non-object terms: The case for time words. *Journal of Cognition and Development*, *11*(1), 16-36.

Singer, M. A., & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. *Psychological Science*, *16*(2), 85-89.

Tillman, K. Marghetis, T., Barner, D., & Srinivasan, M. (in press). Today is tomorrow's yesterday: Children's acquisition of deictic time words. *Cognitive Psychology.*

Thompson, L. A., Driscoll, D., & Markson, L. (1998). Memory for visual-spoken language in children and adults. *Journal of Nonverbal Behavior*, *22*(3), 167-187.