Title
Timing relationships between representational gestures and speech: A corpus based investigation

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors
Donnellan, Ed
Özder, Levent Emir
Man, Hillarie
et al.

Publication Date
2022

Peer reviewed

Timing relationships between representational gestures and speech: A corpus based investigation

Ed Donnellan (ed.donnellan@ucl.ac.uk)
Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom

Levent Emir Özder (lozder16@ku.edu.tr)
Department of Psychology, Koç University, Rumelifeneri Yolu, Sarıyer, 34450, İstanbul, Turkey

Hillarie Man (hillarieman@gmail.com)
Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom

Beata Grzyb (b.grzyb@ucl.ac.uk)
Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom

Yan Gu (yan.gu@ucl.ac.uk)
Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom
Department of Psychology, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK

Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)
Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom

Abstract

Theories suggest that representational gestures depicting properties of referents in accompanying speech could facilitate language production and comprehension. In order to shed light on how gesture and speech are coordinated during production, we investigate whether representational gestures are time-locked to the onset of utterances (hence planned when full events are encoded) or Lexical Affiliates (LAs; words most closely aligned with the gesture meaning; hence planned when individual concepts are encoded) in a large corpus of naturalistic conversation (n = 1803 gestures from n = 24 speakers). Our data shows that representational gestures are more tightly tied to LA onsets than utterance onsets, which is consistent with theories of multimodal communication in which gestures aid conceptual packaging or retrieval of individual concepts rather than events. We also demonstrate that in naturalistic speech, representational gestures tend to precede their LAs by around 370ms, which means that they could plausibly allow for an addressee to predict upcoming words (ter Bekke, Drijvers & Holler, 2021; Ferré, 2010; Habets et al., 2011).

Keywords: multimodal communication; gesture; representational gestures; iconicity; lexical affiliates

Introduction

In face-to-face conversation, speakers produce a range of co-speech gestures, e.g., pointing to, or producing representational gestures that imagistically depict properties of referents. We focus here on representational gestures that depict properties that are also expressed in co-occurring speech. It is argued that these gestures may support both language comprehension and production processes. Some theories focus on the role of these gestures in aiding comprehension as interlocutors can predict upcoming lexical items from seeing a representational gesture (Yap et al., 2011; ter Bekke, Drijvers & Holler, 2021). Other theories focus on the role of representational gestures in aiding production, e.g., the Gesture-for-Conceptualization Hypothesis (Kita, Alibaba & Chu, 2017; in particular the Information Packaging Hypothesis, Kita, 2000; Mol & Kita, 2012) and the Lexical Retrieval Hypothesis (Krauss, 1998; Rauscher, Krauss & Chen, 1996). These theories suggest that producing representational gestures can support the packaging of conceptual information for production, or can enhance lexical activation, facilitating word retrieval.

Previous work has tentatively established that representational gestures tend to precede Lexical Affiliates (LAs), e.g., words in speech which are closely aligned with the gesture meaning (Schegloff, 1984). In particular studies have demonstrated that this is true for the onset of key phases of a gesture, i.e., gesture preparation (where a speaker's hands raise from a resting position to perform the gesture), and the gesture stroke (the part of the gesture that conveys its meaning, e.g., movements that are clearly depictive of something, see Kendon, 1980; McNeill, 1992). Studies using elicitation paradigms have demonstrated this timing relationship in speech produced by people who are told to describe something to an experimenter, confederate or to camera in narrative fashion (Bergmann, Aksu & Kopp, 2011; Church, Kelly & Holcombe, 2014; Graziano, Nicoladis & Marentette, 2020; Morell-Samuels & Krauss, 1992). Moreover, there have been rare attempts to determine if this effect is present in 'the wild', in naturalistic conversations. Early qualitative research provided descriptive accounts of naturalistic interactions in which representational gestures preceded their LA(s) (Schegloff, 1984). More recently there has been data from conversations between familiar

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

individuals demonstrating that a majority of representational gestures precede their LAs in close temporal proximity when produced in naturalistic conversation (ter Bekke et al., 2021; Ferré, 2010, though see Chui, 2005; Urbanik & Svennevig 2021).

However, these studies use either a small sample of speakers (Chiu, 2005; Ferré, 2010, Urbanik & Svennevig 2021) or restrict the context in which representational gestures are used, e.g., focusing on only gestures accompanying questions (ter Bekke et al., 2021). Therefore, we still require large-sample verification using representational gestures produced in naturalistic conversation across multiple speech contexts.

More fundamentally however, previous work has been limited to assessing the timing relationship between gestures and LAs. Here we argue that this is only one part of the picture, and that it is important to further assess whether there are also dependency relationships in the timing of gestures and larger linguistic units such as utterances.

Are Representational Gestures Time-Locked to Utterances or Words during Production?

Answering this yet unexplored question can shed light on how these two modalities are coordinated during production processes. In turn, this can therefore constrain our interpretations of how gestures can be used for prediction in language comprehension (ter Bekke et al., 2021; Zhang et al., 2021).

There are at least three different types of timing relationships between gestures and speech that are of theoretical relevance. Firstly, as proposed by McNeill (1992; 2014) gestures and speech could be planned together at a conceptual (message) level and then each modality could be encoded separately. More specifically, this theory suggests that there is a hypothetical point (e.g., the growth point) at which a speaker formulates the idea of what is to be communicated, which is then put into action by vocal and gestural systems. Assuming that growth points occur at event boundaries, this proposal would predict that the beginning of an utterance (defined as a speech unit describing an event, Berman & Slobin, 1994) and the beginning of a gesture (preparation phase) are tightly linked. The timing of LAs would be independent of this relationship. This would also suggest that the fact that gestures tend to precede the LA is simply an artefact of the fact that the hands may arrive at the depiction of the LA before the LA can be produced. As McNeill (1985, p. 361) put it, "There exist anticipations where the concept revealed in the gesture becomes available before the sentence can grammatically make use of the linguistic item that signifies the concept."

However, gesture and speech could be linked in a more fine-grained way. Assuming incrementality in production processes (e.g., Ferreira & Dell, 2000; Zhao & Wang, 2016), the units of joint planning across the two modalities could be smaller than an event, corresponding to specific concepts encoded into words or phrases. Thus, we should find that gestures are more tightly linked to LA onset. This would mean that gesture preparation and stroke are better predicted by LA onset than by utterance onset.

There is a third possibility, namely that while speech and gesture are planned at the conceptual level on events, as proposed by McNeill (1992), gesture deployment is delayed until the corresponding lexical item is upcoming in speech (de Ruiter, 2010). If this is the case, gesture preparation would be more tightly linked to utterance onset, whereas gesture strokes would be tied to LA onset.[1] This suggests that gesture and speech may be co-constructed in a shared computational stage, but that the link between LA onset and stroke onset is not an artefact of other constraints.

Timing Relationships and Language Comprehension

Crucially, the type of timing dependencies present in production will constrain whether and how an addressee can use gestures during comprehension. For an addressee, gestures can be used to predict upcoming words in an utterance if the gesture stroke comes before the onset of the LA. Moreover, gestures should remain within tight temporal proximity. In naturalistic speech streams, it is not sufficient simply to show the gesture comes first: the gesture also needs to precede the LA by a specific time. It is unclear what the optimal time is, however, in a priming study investigating the N400 effect, Habets et al., (2011) found that 360ms was enough time for someone to generate an expectation of an upcoming word from seeing the stroke of a representational gesture. Their data suggested that no expectation of a lexical item had been generated from the representational gesture for a latency of less than 360ms (e.g., 160ms or simultaneous presentation of the stroke and LA). Given the possible relationships between units in speech and phases of gestures described above, this timing appears to be compatible with the second and third scenarios in which gesture stroke and LA are tightly linked together.

The Current Study

In the current study, we use data from naturalistic conversations between familiar individuals to evaluate the plausibility of different theoretical proposals about gesture and speech production processes. We do so by considering the timing relationship between relevant units in the speech (e.g., utterance and LA onset) and relevant phases of the gesture (e.g., preparation and stroke).

[1] If an LA comes early in an utterance (or indeed LA and utterance onset are the same, i.e., if the LA is the first word of an utterance), then gesture preparation and stroke may also be in close proximity to both utterance onset and LA onset. But in naturalistic speech, LA onset can come significantly after utterance onset, not only in cases where there are other constraints, but also in cases of word finding difficulties. It is this variance that allows us to determine the relationship more precisely.

Method

Participants

The sample consisted of N = 24 adults from the ECOLANG corpus (F = 14, M = 10, Age [years] M = 24.46, SD = 5.55).

ECOLANG Corpus

The ECOLANG corpus (Vigliocco et al., unpublished) is a new multimodal corpus of semi-naturalistic dyadic interactions between familiar adults (one designated the speaker, and the other the addressee for the whole interaction)[2]. Dyads were sat at a table in a lab at 90 degrees from each other. Speakers were asked to talk about 6 objects from 4 sets of topics (animals, tools, foods and musical instruments) in a natural manner to their addressee for 4-5 minutes per topic. For each set, speakers were asked to talk about three generally known objects (e.g., elephant, compass, mango, accordion) and three generally unknown objects (e.g., axolotl, strigil, cherimoya, xun). Speakers were taught about the unknown objects prior to the experiment to facilitate their conversation with the addressee. Replicas of the objects (e.g., models of the animals or food, or the tools and musical instruments) were both present or absent during the interaction (counterbalanced).[3] In total, the recording session took around 40 minutes. The interaction was video- and audio-recorded.

The speakers' speech has been transcribed at the utterance (Berman & Slobin, 1994) and word level and marked in ELAN (Sloetjes & Wittenburg, 2008). Utterances were defined as a unit that expresses a single situation (an activity, event or state). Additionally, representational gestures (where the speaker produced meaningful hand actions depicting an object or event, e.g., moving their hands in a circular motion to depict a wheel turning) have also been coded. Note that our definition of representational gestures includes iconic gestures (see McNeill 1985; 1992) and emblems (but not points).

Coding

Further coding for our project was conducted in ELAN, adding to the coded elements of the corpus.

Gesture Phases. For each representational gesture, we marked the onset of two phases: the preparation and the stroke (Seyfeddinipur, 2006; Kita, van Gijn & van der Hulst, 1998).

The onset of the preparation phase was marked when the hand(s) began moving to form the shape of the gesture. This could be marked at the point where the hand(s)/arm(s) began raising from a resting position (which could be in the lap, on the table, or mid-air). Alternatively, if the speaker was gesturing just prior to the representational gesture, the onset of the preparation phase was coded as the point that the hand shape of the previous gesture relaxed, and the hand(s) moved to make the representational gesture. Note this was only the case if the hands moved towards the representational gesture in one movement, i.e., in cases where the hand(s) retracted from the previous gesture to a resting position before preparing for the representational gesture, then the onset was marked at the point where the hands left the resting position.

The onset of the stroke phase marked the point where the hand(s) begin to display the meaning of the gesture, with the hands showing a well-defined configuration (shape) and well-articulated movement that clearly depicted some property. For gestures with multiple strokes (depicting the same meaning), we simply marked the onset time of the first stroke (note that if two strokes appeared to convey different meanings, then these were treated as separate gestures).

Lexical Affiliates. We considered words in close proximity to the gesture (i.e., in the second before it started and ended) as potential LAs for that gesture. For each gesture, the word(s) that corresponded most closely to a gesture in meaning were marked as LAs for that gesture, and the onset time taken. We constrained the LAs to the minimum amount of word(s) to convey the meaning (i.e., omitting definite articles). If there were multiple affiliates, we took the onset time from the earliest LA associated with a gesture.

Utterance onset. We also recorded the onset time of the utterance in which the LA was produced (as marked in the corpus).

Reliability

Ten percent of each speaker's representational gestures were double coded (resulting in n = 300 gestures for reliability calculations). Both coders coded the preparation and stroke phase of the gesture and identified LAs.

LAs. Each coder established if each potential LA (words occurring within 1000ms of the gesture begin and end), was or was not a LA of the gesture (resulting in n = 3878 potential LAs). Coders agreed on 95.00% of these words (Cohen's κ = .70 [95% CI = .65 to .74], indicating substantial agreement).

First LA. There was agreement for 72% of gestures on the first affiliate/that there was no LA (for 57% of gestures coders agreed on first LA, for 15% they agreed that there was no LA associated with the gesture). For half of disagreements between coders (accounting for 28% of gestures considered), one coder selected an LA when the other coded none. For the other half, coders did not agree on the first LA (having both selected a LA for the gesture). However, it is worth noting that in the majority (32/42) of these cases, the first LA

[2] Note that the ECOLANG corpus also includes adult-child dyads, not considered in the current study. Note also that the full sample of ECOLANG adult dyads is N=33. Our project is currently ongoing.

[3] For the current study, we do not distinguish between communication about known/unknown or present/absent entities for the purpose of analysis.

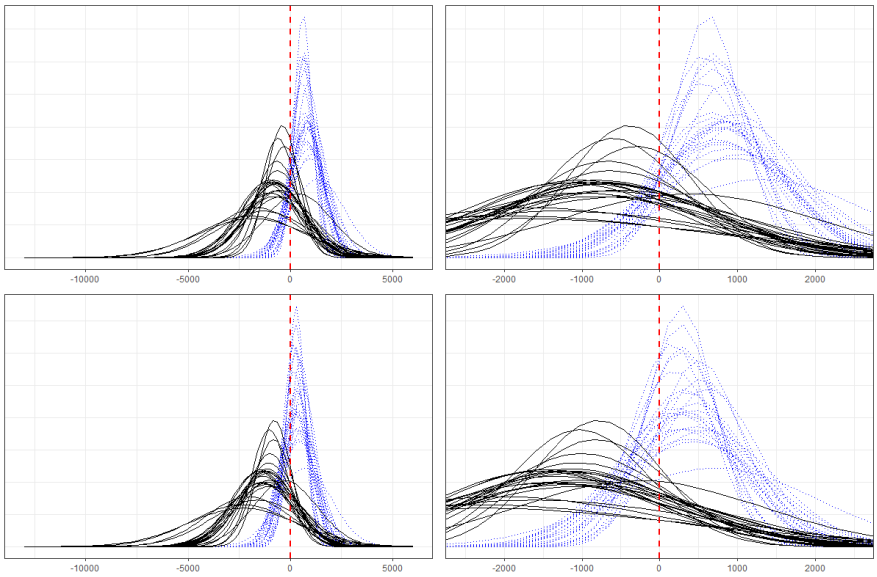

Figure 1: Latency between gesture preparation onset (top: 0ms indicated by the dashed red line) or stroke onset (bottom: 0ms indicated by the dashed red line), utterance onset (black solid lines) and LA onset (blue dashed lines). Normal distributions plotted for each speaker using *M* and *SD* across all gestures produced by that speaker (QQ plots indicated all normal distributions). Left: entire distribution; right: zoomed-in plot.

selected by one coder was included as a LA by the other (just not as the first LA).

Latencies. For gestures where coders agreed on the first LA, and that the gesture had a stroke and preparation phase (n = 171), reliabilities on the latencies considered in our analysis indicated high levels of agreement: (1) gesture preparation to LA onset, r = .95, (2) gesture stroke to LA onset, r = .94, (3) gesture preparation to utterance onset, r = .99, (4) gesture stroke to utterance onset, r = .99. Even when including cases where there was not agreement on the first LA (for n=213 gestures), correlations were high between coders for all latencies: (1) r = .77, (2) r = .77, (3) r = .94, (4) r = .93.

Analysis

For analysis, we only included representational gestures for which we could identify both the preparation phase and stroke phase, and that had at least one LA (n = 1803 representational gestures). All analyses were done in R 3.6.2 (R Core Team, 2019), with mixed effects models run using lme4 (Bates et al., 2015), and model summaries generated using lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017).

Links to LA and Utterance Onsets. To analyze whether gesture preparation and stroke onset are more related to LA or utterance onset we constructed mixed effects models with gesture preparation onset or stroke onset time predicted by either utterance or LA onset. Speaker ID was included as a random effect on the slopes, and all variables were mean-centered and scaled (M = 0, SD = 1) to allow for meaningful comparisons. Model comparison using Akaike's Information Criterion (AIC), specifically AIC difference (Δ_i: where for a model$_i$, Δ_i, = AIC$_i$ - AIC$_{min}$ over candidate models) allowed us to determine whether utterance or LA onset was a better predictor for either preparation or stroke onset. Δ_i = 0 indicates the best fitting model, larger values of Δ_i indicate worse fit, with $\Delta_i > 2$ indicating that a model is substantially less plausible compared to the best fitting model (Burnham & Anderson, 2002). Finally, models with both LA and utterance onset predicting stroke or preparation onset were constructed to allow for comparison of their relative contributions.

Timing between Gesture Phase (Stroke/Preparation) and LA. To determine exact latencies between gesture phase onsets and LAs (e.g., the latency between gesture stroke and LA), we constructed mixed effects models with gesture phase (stroke or preparation) onset predicted by LA onset. Speaker ID was included as a random effect on the slope, and gesture ID included as a random intercept. Note, variables were not mean-centered and scaled, so as to obtain an estimate of the exact latencies (in order to establish their plausibility as primes).

Results

Links to LA and Utterance Onset

Figure 1 shows the relationship between gesture preparation and stroke onset and LA and utterance onset. As LA onsets are more densely distributed in proximity to both preparation and stroke onset than utterances, this shows that gesture

preparation and stroke onset is more tightly linked to LA onset. Model comparison confirmed that LA onset was a stronger predictor than utterance onset of both gesture preparation onset ($\Delta_{Utterance\ onset}$ = 2562.588) and stroke onset ($\Delta_{Utterance\ onset}$ = 2749.114). Tables 1 and 2 show that the relative contribution of LA onset to preparation and stroke onsets is higher than for utterance onset.

Table 1: Gesture preparation onset predicted by LA and utterance (Utt) onset

	β	*SE*	*df*	*t*	*p*
(Int)	0.000	0.000	24.153	-0.020	.984
LA	0.920	0.019	1718.171	48.526	< .001
Utt	0.080	0.019	1632.928	4.241	< .001

Table 2: Gesture stroke onset predicted by LA and utterance (Utt) onset

	β	*SE*	*df*	*t*	*p*
(Int)	0.000	0.000	23.054	0.025	.980
LA	0.931	0.014	1437.968	67.364	< .001
Utt	0.069	0.014	1483.578	4.978	< .001

Timing Between Gesture and LA

Gesture stroke onsets tended to precede LA onset for all speakers (see Figure 1 for distributions for each speaker). The model revealed that the estimated latency between stroke and LA is around 370ms (B = -370.87, SE = 30.57, t(23.11) = -12.13, p < .001) when accounting for random slopes across speakers. By comparison, gesture preparation phases preceded the LAs by around 814ms (B = -814.32, SE = 37.41, t(24.08) = -21.77, p < .001).

Discussion

We have found that in naturalistic production gesture preparation and stroke phases are more closely linked to the onset of the corresponding LA than to the beginning of the corresponding utterance. These findings do not support hypotheses that the shared planning unit of speech and gesture is as large as a whole event (operationalized here in terms of utterance), as the growth point theory by McNeill (1992; 2014) suggests. They also do not support hypotheses that speakers routinely prepare their gesture at the beginning of the utterance and delay the stroke phase until the LA onset (as suggested by de Ruiter, 2010). This indicates that the fact that representational gestures tend to precede their LA is not simply an artefact of formal constraints affecting speech (McNeill, 1985).

Constraints on Theories of Multimodal Production

The timing relationships we observed are consistent with theories of multimodal production that assume that the unit of shared planning of the speech and the gesture at the conceptual level is not as large as a whole event, but smaller corresponding to sub-units (e.g., words or phrases), in line with incrementality in language production. This proposal aligns with theoretical views according to which crucially, the gestural system is not recruited to the service of encoding operations by the language production system, but at a more general conceptual level (see the Information Packaging Hypothesis, e.g., Mol & Kita (2012), as part of the Gesture-for-Conceptualization Hypothesis (Kita et al., 2017)).

Our results are also compatible with alternate views according to which the gestural system is engaged only after conceptual encoding, during lexical retrieval (e.g., Butterworth & Hadar 1987; Hadar & Krauss, 1999; Krauss, 1998; Rauscher, Krauss & Chen, 1996). A speaker formulates an idea of what is to be communicated in speech, with gestural systems recruited after. Under this theory, gestures are tightly linked to their LAs, because they are deployed within an utterance as pre-planned lexical items are reached to aid in their retrieval. Such theories assume that recruitment of the gestural system would be driven by the speech production system.

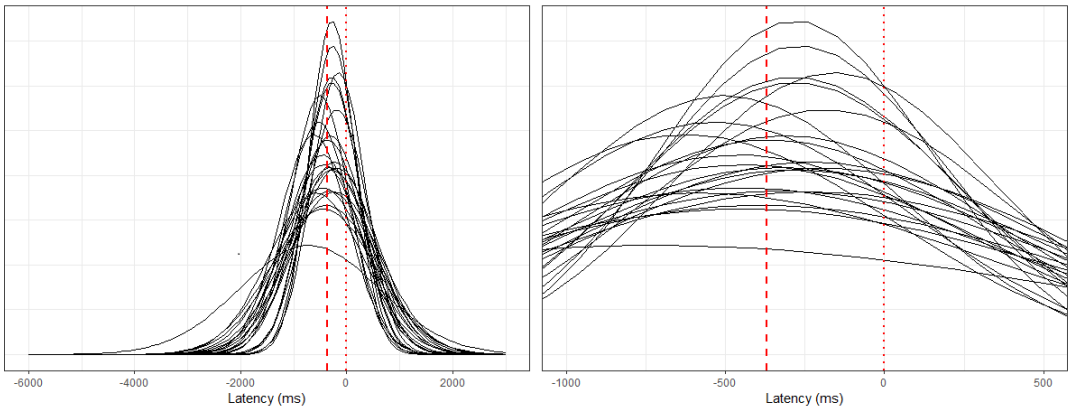

Figure 2: Latency between gesture stroke onset and LA onset (0ms indicated by the dotted red line). Normal distributions plotted for each speaker using M and SD across all gestures produced by that speaker (QQ plots indicated all normal distributions). Dashed red line indicates the coefficient estimate extracted from the mixed-effects model (370.87ms). Left: entire distribution; right: zoomed-in plot.

While we cannot decide between these two alternative views solely on the basis of the current results, there is evidence that argues against this second view (that gestural systems are driven by speech production systems, i.e., by the need for lexical retrieval). Recent evidence suggests this is unlikely, as restricting speaker's ability to gesture does not impair fluent speech production (Kısa, Goldin-Meadow & Casasanto 2021).

Gestures Can be Used to Predict Upcoming Words in Comprehension

Here we have demonstrated that in naturalistic conversation, the two necessary criteria for representational gestures to be able to predict an upcoming word in comprehension are met. Firstly, the stroke phase of representational gestures (representing their meaning) tends to be deployed just prior to LAs. Secondly, they remain within tight temporal proximity, tending to be produced around 370ms before. This is similar to previously reported latencies (e.g., ter Bekke et al., 2021; Ferré, 2010), and is strikingly close to the 360ms thought to be required for a representational gesture to act as a prime for a lexical item (Habets et al., 2011). This makes theories that listeners can use speakers' representational gestures to predict upcoming speech ecologically plausible.

Future Directions

While the general picture we have presented is valid, it is clear that there are some individual differences in the timing relationship between representational gestures and LAs, and variation whereby a small percentage of gestures (preparation and stroke) start well in advance or indeed after their LA. To some extent this could be due to properties of the LAs. Previous work (using n = 60 gestures) demonstrated that word familiarity predicted variation in the latency between LA and gesture preparation (Morrell-Samuels & Krauss, 1992). Further investigation (currently ongoing) should seek to determine what lexical properties of LAs can affect the timing relationship with gesture preparation and stroke phases, e.g., word frequency or lexical surprisal. Additionally, the corpus includes both fluent and disfluent speech (as the speech is naturalistic), and it is possible that this accounts for variation in timing (see Arslan & Goksun, 2022).

Future work should also seek to determine if the timing relationships observed here are similar in gestures produced to children. It is thought that children younger than 3 years old struggle to interpret the meaning of representational gestures, developing the means to interpret them through preschool (Tolar, Lederberg, Gokhale & Tomasello, 2008). If the timing of co-speech gestures is under a speaker's intentional control (i.e., are designed for audience comprehension), then we may see accommodations made in the timing of representational gestures for young children.

Acknowledgments

The work reported here was funded by a European Research Council Advanced Grant (ECOLANG, 743035) and Royal Society Wolfson Research Merit Award (WRM\R3\170016) to GV.

References

Arslan, B., & Göksun, T. (2022) Aging, gesture production, and disfluency in speech: A comparison of younger and older adults. *Cognitive Science*, *46*, e13098.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1-48.

Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures : Temporal synchrony follows semantic synchrony. *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction.*

Berman, R. A., & Slobin, D., (1994). *Relating events in narrative: A crosslinguistic developmental study.* Lawrence Erlbaum Associates.

Burnham, K. P., & Anderson, D. R. (2002) *Model selection and multimodal inference: A practical and information-theoretic approach (Second Edition)*. New York, NY: Springer-Verlag.

Butterworth, B., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review*, *96*, 168–174.

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, 37, 871–887.

Church, R. B., Kelly, S., & Holcombe, D. (2014). Temporal synchrony between speech, action and gesture during language production. *Language, Cognition and Neuroscience*, *29*, 345–354.

de Ruiter, J. P. (2000). The production of gesture and speech. In McNeill, D. (Ed.), *Language and Gesture*. Cambridge University Press.

Ferré, G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous French. *LREC: Workshop on Multimodal Corpora*, 86–91.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*, 296–340.

Graziano, M., Nicoladis, E., & Marentette, P. (2020). How referential gestures align with speech: Evidence from monolingual and bilingual speakers. *Language Learning*, *70*, 266–304.

Habets, B., Kita, S., Shao, Z., Özyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*, 1845–1854.

Hadar, U., & Krauss, R. K. (1999). Iconic gestures: the grammatical categories of lexical affiliates. *Journal of Neurolinguistics*, *12*, 1–12.

Kendon, A. (1980). Gesticulation and Speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relation between verbal and nonverbal communication*. Mouton.

Kita, S. (2000). How representational gestures help speaking. In McNeill, D. (Ed.), *Language and Gesture: Language, Culture and Cognition*. Cambridge University Press.

Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review, 124*, 245–266.

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. *Gesture and Sign Language in Human-Computer Interaction, International Gesture Workshop Bielefeld, Germany, September 17-19, 1997, Proceedings. Lecture Notes in Artificial Intelligence, 1371*, 23–35.

Kısa, Y. D., Goldin-Meadow, S., & Casasanto, D. (2021). Do gestures really facilitate speech production? *Journal of Experimental Psychology: General*.

Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science, 7*, 54–60.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review, 92*, 350–371.

McNeill, D. (1992). *Hand and mind: what gestures reveal about thought.* The University of Chicago Press.

McNeill, D. (2014). Gesture–speech unity: Phylogenesis, ontogenesis, and microgenesis. *Language, Interaction and Acquisition, 5*, 137–184.

Mol, L., & Kita, S. (2012). Gesture structure affects syntactic structure in speech. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 761–766.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 615–622.

R Core Team. (2019). R: A Language and Environment for Statistical Computing. https://www.r-project.org/

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: the role of lexical movements in speech production. *Psychological Science, 7*, 226–231.

Schegloff, E. A. (1984). On some gestures' relation to talk. In Atkinson, I., & Maxwell, J., *Structures of Social Action*. Cambridge University Press.

Seyfeddinipur, M. (2006). *Disfluency: Interrupting Speech and Gesture.* Doctoral dissertation, Radboud Universiteit Nijmegen.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation.*

ter Bekke, M., Drijvers, L., & Holler, J. (2020). The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech. *Proceedings of the 7th GESPIN - Gesture and Speech in Interaction Conference. Stockholm: KTH Royal Institute of Technology,* 1–6.

Tolar, T. D., Lederberg, A. R., Gokhale, S., & Tomasello, M. (2008). The development of the ability to recognize the meaning of iconic signs. *Journal of Deaf Studies and Deaf Education*, 13, 225–240.

Urbanik, P., & Svennevig, J. (2021). Action-depicting gestures and morphosyntax: the function of gesture-speech alignment in the conversational turn. *Frontiers in Psychology, 12*, 1–24.

Vigliocco, G. et al. (unpublished) The ECOLANG corpus of dyadic interactions between caregivers and their 2-4 year-old child and between two adults.

Yap, D.-F., So, W.-C., Yap, J.-M. M., Tan, Y.-Q., & Teoh, R.-L. S. (2011). Iconic gestures prime words. *Cognitive Science, 35*, 171–183.

Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J. I., & Vigliocco, G. (2021). More than words: word predictability, prosody, gesture and mouth movements in natural language comprehension. *Proceedings of the Royal Society B: Biological Sciences, 288*, 20210500.

Zhao, L.-M., & Yang, Y.-F. (2016). Lexical planning in sentence production is highly incremental: Evidence from ERPs. *PLOS ONE, 11*, e0146359.