The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech

Marlijn ter Bekke[1,2], *Linda Drijvers*[1,2], *Judith Holler*[1,2]

[1]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
[2]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
Marlijn.terBekke@mpi.nl, Linda.Drijvers@mpi.nl, Judith.Holler@mpi.nl

Abstract

In face-to-face conversation, recipients might use the bodily movements of the speaker (e.g. gestures) to facilitate language processing. It has been suggested that one way through which this facilitation may happen is prediction. However, for this to be possible, gestures would need to *precede* speech, and it is unclear whether this is true during natural conversation.

In a corpus of Dutch conversations, we annotated hand gestures that represent semantic information and occurred during questions, and the word(s) which corresponded most closely to the gesturally depicted meaning. Thus, we tested whether representational gestures temporally precede their lexical affiliates. Further, to see whether preceding gestures may indeed facilitate language processing, we asked whether the gesture-speech asynchrony predicts the response time to the question the gesture is part of.

Gestures and their strokes (most meaningful movement component) indeed preceded the corresponding lexical information, thus demonstrating their predictive potential. However, while questions with gestures got faster responses than questions without, there was no evidence that questions with larger gesture-speech asynchronies get faster responses. These results suggest that gestures indeed have the potential to facilitate predictive language processing, but further analyses on larger datasets are needed to test for links between asynchrony and processing advantages.

Index Terms: Multimodal communication, language prediction, co-speech gesture

1. Introduction

Most language use occurs in face-to-face conversation, which involves rapid turn-taking, with modal gaps between turns being around 0-200 ms across languages [1]. This is very quick, considering that language production takes time: preparing to produce a single word takes at least 600 ms [2]. This means that response planning must already start before the incoming turn is finished.

How do recipients do this? One possibility is that they predict the content (and end) of the incoming turn, such that they can begin to process the message and plan their response turn as soon as possible [3]–[6]. As a result, responses can then occur with minimal gaps between turns. Indeed, there is evidence that people predict various aspects of the incoming turn while listening, including the turn's speech act [7], the upcoming words [6], [8] and the turn end [9].

It has been proposed that, in face-to-face conversation, recipients might not only use speech but also the communicative bodily movements of the speaker to facilitate predictive language processing [10]. One hint that co-speech gestures may play a role for language prediction comes from a corpus study, in which it was found that questions with gestures get faster responses [11]. This may be the result of prediction, but as the authors state, potential other explanations cannot be ruled out.

For co-speech gestures to play a role in semantic predictive processing, it is crucial that the gestures (a) contain semantic information that is related to speech; and (b) precede the corresponding information in speech (often called the lexical affiliate, cf. [12]). Concerning the first criterion, people often use manual co-speech gestures that represent semantic information [13], [14]. These so-called representational hand gestures include gestures that depict information about objects or actions (iconic gestures) or about abstract concepts (metaphoric gestures), and hand gestures that point to concrete or abstract locations or objects (deictic gestures) [15]. Crucially, the semantic information the gestures depict is tightly related to the semantic information in speech [14], [16]. Regarding the second criterion, the question remains whether such representational hand gestures precede their lexical affiliates in turn-taking contexts.

Based on the existing literature, it has generally been accepted that representational gestures slightly precede their lexical affiliates. However, this conclusion is partly based on qualitative studies describing individual cases of preceding gestures observed in conversation [12], [17], partly on studies that involved task-elicited descriptions rather than conversation [18]–[24] and partly based on studies that mixed conversation and monologue data [25]. Concerning the question of what the gesture-speech timing is like in natural, interactive conversation, there are two relevant prior studies. For spontaneous French dialogues, it was found that 95% of iconic hand gestures (as a whole) started before lexical affiliate onset, on average preceding the lexical affiliate by 0.82 seconds [26]. Regarding the stroke phase, it was found that 72% of gesture strokes started before lexical affiliate onset, on average preceding the lexical affiliate by 0.45 s. The results from a study on Chinese multiparty conversations are less conclusive: it was found that 60% of iconic gesture strokes synchronized with the lexical affiliate, 36% preceded it and 4% followed it [27]. Unfortunately, it was unclear whether synchronization meant that the stroke onset occurred during the lexical affiliate, or that the stroke started before but overlapped with the lexical affiliate. These possibilities have differing implications for the question at hand, with only the latter speaking to the predictive potential of gestures.

Thus, there is a clear need for further systematic, quantitative investigations of gesture-speech synchrony in datasets of non-task-elicited, naturalistic conversational interactions. The present study aimed to contribute new research to address this gap. This study extends the literature by looking not only at iconic gestures, but meaningful (i.e. representational) gestures more broadly. In everyday conversations other types of representational gestures (e.g. pointing) are also frequently used, and therefore it is essential to study these non-iconic gestures too. Moreover, the current study investigated gesture-speech timing in Dutch for the first time, an important step to see whether the results found in French [26] generalize to different languages as well.

We tested whether representational hand gestures temporally precede corresponding information in speech, such that they could potentially be used to facilitate predictive language processing. In a Dutch corpus of unscripted dyadic conversations, we annotated representational hand gestures. For each gesture, we coded which word(s) in speech corresponded most closely to the meaning depicted by the gesture (i.e. the lexical affiliate), and compared the timing of the word(s) to the timing of the gesture.

Moreover, to gain a first insight into whether any temporally preceding gestural semantic information may facilitate recipient's language processing, we focused on recipients' response speed. To this end, we focused the present analysis on conversational turns encoding questions, since by (Western) conversational norm interactants are typically expected to provide a responding turn (although not always, e.g. rhetorical questions). We thus analyzed gap durations for all question-response sequences contained in the corpus. This allowed us to test whether the more a gesture in a question precedes the corresponding information in speech (i.e. greater predictive potential), the faster that question gets a response. Of course, other mechanisms than prediction may explain such a pattern (e.g. gesture retractions may also occur earlier for those gestures beginning earlier, potentially acting as an early 'go ahead' signal, see [11]), but it would be a further piece of the puzzle in line with a prediction-based explanation.

2. Methods

2.1. Corpus

The present analyses are based on the Communication in Action (CoAct) corpus, which consists of 34 dyads of acquainted native Dutch speakers who each conversed for 60 minutes, 20 minutes of those freely, representing the basis for the present analyses. Interactants were paid for their participation and the corpus creation was approved by the Social Sciences Faculty Ethics Committee, Radboud University Nijmegen.

2.2. Apparatus

The conversations took place in a soundproof room, in which the participants sat opposite to each other. One camera recorded the body from head to toe, and one recorded the hands from a bird's eye perspective (both: Canon XF205, 25 fps). High quality audio was recorded using two separate microphones that stood on the floor close to the participants (Sennheiser ME64), and audio and video were synchronized in Adobe Premiere Pro CS6. Moreover, audio from the two microphones was combined into one recording that contained the audio of both participants at comparable volume, which was used for the present analyses. (Participants were filmed with three further cameras, one capturing the interaction as a whole, and two providing close-up views of each of the participants' faces, but these recordings were not used for the present analyses).

2.3. Coding

The present analysis focused on the timing relations between representational gestures and their lexical affiliates in the context of question-response (QR) sequences. To this end, we first identify the QR sequences, followed by the gestures, their corresponding lexical affiliates, and finally the gesture phases. All annotations were made in ELAN (version 5.5; [28]).

2.3.1. Question-response sequences

We initially identified the QR sequences based on the coding manual by Stivers and Enfield [29], and complemented this with additional coding rules on an inductive basis, to account for the breadth of data we observed in our corpus. Overall, we took a holistic approach, taking into account the phrasing, intonation, visual cues, context, pausing and addressee behaviour. Verbal responses to the questions were identified as anything that was said in response to the question. This included conventionalized sounds such as "ehhh" but did not include non-verbal sounds such as laughter, sighs and lip smacks. For the present study, we excluded questions that were not designed to get a response from the other person, i.e. self-directed questions (e.g. "What was it called again?") and questions in reported speech (e.g. "And then I said: "Why not?"").

Reliability for question and response identification was calculated based on 11.8% of the data. For questions, we observed 74.5% agreement, and a modified Cohen's Kappa of 0.74, indicating substantial agreement [30], [31]. For responses, we observed 72.7% raw agreement, and a modified Cohen's Kappa of 0.73, also indicating substantial agreement.

2.3.2. Representational gestures

Next, we coded representational gestures, which are gestures that "depict semantic information by virtue of handshape, placement, or motion" [15, pp. 173]. This included iconic gestures, metaphoric gestures, and deictic gestures (concrete or abstract), and excluded pragmatic or interactive gestures, beats, or emblems.

We looked through the videos in their entirety and for each representational gesture we checked whether it related in meaning to a nearby question or response. This way, we did not decide on an a priori time window about how far away from questions/responses we might still expect to see related representational gestures. In total, this resulted in 281 annotated gestures. Of these, 139 occurred during questions, 131 during responses, and 11 during utterances that were simultaneously a question and a response (e.g. "Were you invited to the party?" "Which party?").

Reliability was calculated based on 22.3% of the data. These randomly chosen segments contained 21.4% of the relevant representational gestures identified by the first coder (*n* = 60), and were also used for lexical affiliate (section 2.3.3) reliability. Agreement was 80.3%. Cohen's kappa could not be calculated because there was only one gesture category.

2.3.3. *Lexical affiliates*

For each gesture the lexical affiliate was determined, which was defined as the word(s) deemed to correspond most closely to a gesture in meaning [12]. The main strategy was to first see what information the gesture depicted, and to then choose the corresponding lexical affiliate. This meant for example that if gestures depicted an action, we chose the corresponding action verb. To illustrate, when the handshape depicted holding a glass, and the hand was then brought to the mouth, the verb phrase "slokje nemen" (to take [a] sip) was chosen as the lexical affiliate.

Following [18], we excluded articles from lexical affiliate selection, and when possible, we also omitted prepositions and the amount of entities. When an entity was described using both a demonstrative and a noun (e.g. "But these are natives"), we chose the noun as the lexical affiliate, because it contains most semantic information. Similarly, demonstratives before nouns were also excluded, e.g. the lexical affiliate would be "bridge" instead of "that bridge". For 23 gestures (8.1%), the lexical affiliate could not be determined. This was either because the information in the gesture was complementary to and thus not present in the speech (20 gestures, 7.1%) or because the gesture was too ambiguous (3 gestures, 1.1%). Lexical affiliate onset/offset was defined as the moment at which the lexical affiliate started/stopped to be vocalized, as identified in Praat (version 5.1; [32]).

For reliability, the coders first indicated for each gesture whether the lexical affiliate was present, absent or ambiguous. We observed 98.3% agreement, and a modified Cohen's Kappa of 0.90, indicating almost perfect agreement on whether a gesture has a lexical affiliate [30]. Second, the coders indicated which word(s) they interpreted as the lexical affiliate. Here, we observed 95.4% agreement, and a modified Cohen's Kappa of 0.95, indicating almost perfect agreement [30].

2.3.4. *Gesture phases*

For gesture phase coding, first the gestures were segmented into dynamic and static gesture phases using the frame-by-frame method described in [33]. Next, the segmented phases were identified as preparation, pre-stroke hold, stroke or stroke hold, post-stroke hold, or retraction [34].

Overall, the first frame of a gesture was typically the first blurry frame of the preparation. The first frame of a stroke or stroke hold was the first frame in which the meaning of the gesture was expressed. The last frame of a gesture was the first frame in which the hands were still in their rest position.

2.4. Analysis

First, we asked whether gesture onsets and gesture strokes precede their lexical affiliate onsets. For these gesture-speech asynchrony analyses, the difference was calculated between gesture/stroke onset time and lexical affiliate onset time for each gesture-affiliate pair.

Next, we ask whether questions with representational hand gestures get faster responses than questions without such gestures. Finally, we ask whether the response speed to questions with gestures depends on the degree of asynchrony between a gesture and its lexical affiliate, testing the hypothesis that the more gestures precede their lexical affiliates (i.e. higher predictive potential) the faster the response to the question.

We fitted linear mixed effects models using the lme4 package (version 1.1-21; [35]) in R (version 3.5.3; [36]), with p-values calculated using the package lmerTest (version 3.1-1; [37]). When multiple tests were run on the same data (or subsets of it), we corrected for multiple comparisons using the False Discovery rate (FDR). For all models, we ideally wanted to use the maximal random effects structures [38]. When these maximal models did not converge, the number of iterations was increased, and subsequently the random effects structure was simplified by removing the terms with lowest variance first.

3. Results

3.1. Do gestures precede speech?

The overwhelming majority of gestures (96%) started before their lexical affiliate, around 672 ms on average (Figure 1, 2). An intercept-only model with random intercept for dyad revealed that overall, gesture onsets significantly preceded lexical affiliate onsets (β = -667.50, *SE* = 47.60, *t* = -14.02, *p* < 0.001).

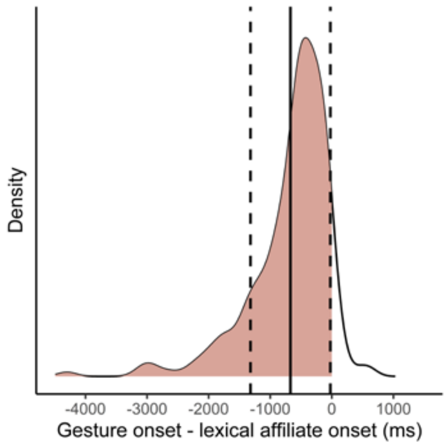

Figure 1: *The overwhelming majority of gestures preceded their lexical affiliate. Density plot of the difference between gesture onset and lexical affiliate onset in ms, where negative values (pink) indicate gestures that preceded their lexical affiliate. Solid line indicates mean asynchrony, dashed lines indicate 1 standard deviation around the mean.*

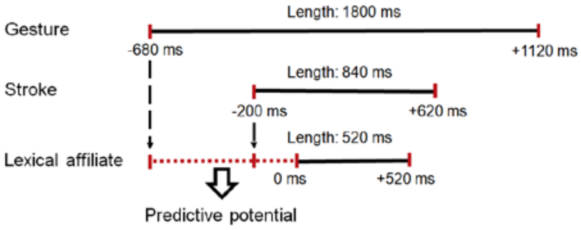

Figure 2*: Mean timing relations between representational gestures, their strokes and their lexical affiliates. Values were rounded off to match video frame precision (40 ms).*

3.2. Do strokes precede speech?

The majority of strokes (62%) started before their lexical affiliate, around 215 ms on average (Figure 2, 3). An intercept-only model with random intercept for dyad revealed that stroke onset significantly precedes lexical affiliate onset (β = -207.40, *SE* = 46.63, *t* = -4.45, *p* < 0.001). Thus, not only gesture onsets as a whole, but also gesture strokes typically started before their corresponding information in speech (Figure 2).

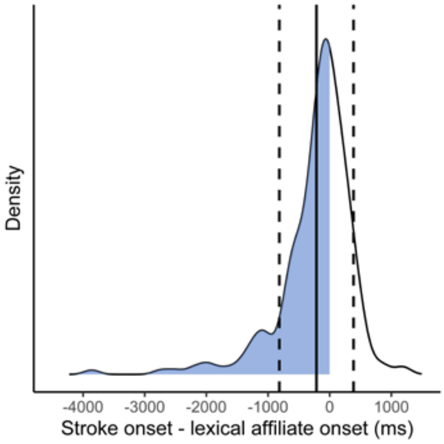

Figure 3: *The majority of strokes preceded their lexical affiliate. Density plot of the difference between stroke onset and lexical affiliate onset in ms, where negative values (blue) indicate strokes that preceded their lexical affiliate. Solid line indicates mean asynchrony, dashed lines indicate 1 standard deviation around the mean.*

3.3. Does gesture-speech asynchrony predict response times to questions?

To investigate whether the temporally preceding semantic information in gestures facilitates language processing, we analyzed response times for the individual question-response sequences as a proxy. In total, the 34 conversations contained 2186 questions, of which 1869 got a verbal response. The questions that did not get a verbal response were for example questions that got a non-verbal responses (e.g. nodding), or rhetorical questions. Of the questions that got a response, 94 were produced with one or more representational gestures (i.e. 5.0%).

There were two special types of QR-sequences for which the response time could not easily be calculated, and which were therefore excluded: 1. multiple questions in a row that get a single response (289 cases); and 2. questions that get multiple responses (12 cases). This way, 13 questions with gestures were removed because of multiple questions and 1 because of multiple responses, leaving 80 questions with gestures for the analysis. Of these 80 questions with gestures, 60 included 1 gesture, 15 included 2 gestures and 5 included 3 gestures.

Overall, the average response time to questions was 330 ms (*SD* = 640 ms, range = [-4205, 5143 ms]). The average response time to questions with representational hand gestures was 54 ms (*SD* = 759 ms, range = [-2979, 2583 ms]). The average response time to questions without representational hand gestures was 345 ms (*SD* = 630 ms, range = [-4205, 5143 ms]). These values indicate that questions with gestures got faster responses than questions without gestures (Figure 4). This pattern was tested with the following model: response time as the dependent variable, gesture presence as a fixed effect, a random slope for gesture presence by participant nested within dyad, and random intercept for dyad. The model revealed that questions with gestures got significantly faster responses, around 297 ms faster (β = -296.65, *SE* = 80.20, *t* = -3.70, *p* < 0.001)

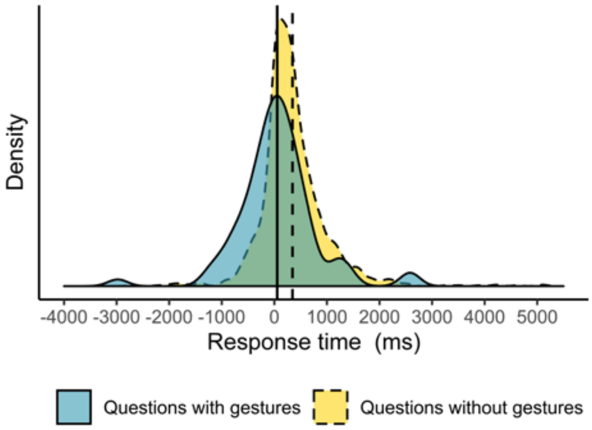

Figure 4: *Questions with gestures got faster responses. Distribution of the response times for questions with and without gestures in milliseconds. Negative values indicate overlap between question and response, positive values indicate gaps. The vertical lines represent the mean response time to questions with (solid line) and without gestures (dashed line).*

One reason why questions with gestures could get faster responses is because potentially the recipients use the information in the gesture to predict upcoming speech. If this is the case, then the response time to questions with gestures might be predicted by the degree to which the gestures precede their lexical affiliate.

Out of the 80 questions that were produced together with a gesture, 74 were produced with a gesture that had a lexical affiliate. We ran two linear models with response time as the dependent variable and gesture-speech asynchrony as the predictor. In the first model, asynchrony was calculated as stroke onset – lexical affiliate onset. The model revealed that stroke-speech asynchrony did not predict response times in our data (β = 0.24, *SE* = 0.15, *t* = 1.55, *p* = 0.29).

Because it is unknown whether only gesture strokes could play a role for prediction, or whether gesture preparations already contain some relevant semantic information, in a second model we calculated asynchrony as gesture onset (as a whole) – lexical affiliate onset. This model also revealed that gesture-speech asynchrony does not predict response times (β = 0.20, *SE* = 0.14, *t* = 1.37, *p* = 0.31) (Figure 5).

4. Discussion

During natural, face-to-face communication, gestures as a whole as well as their most meaningful parts, start before the corresponding information in speech. Therefore, representational manual gestures fulfil the two criteria necessary for language prediction based on gestures to be possible: 1. they share semantic information with speech; 2. they precede this shared semantic information in speech. Thus, co-speech representational gestures indeed appear to have predictive potential.

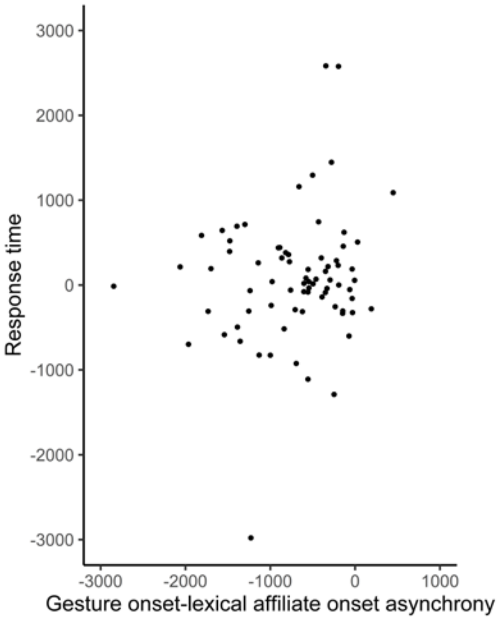

Figure 5: *Response times by gesture onset-lexical affiliate onset asynchrony. Negative asynchrony values indicate that the gesture starts before lexical affiliate onset, positive values indicate that the gesture starts after lexical affiliate onset. Negative response time values indicate overlap between question and response, positive values indicate gaps.*

The finding that gesture onsets and stroke onsets precede their lexical affiliates in Dutch converges with prior work on French conversation [26], and extends it from iconic gestures only to representational gestures more broadly. The result also aligns with previous gesture-speech timing studies that did not look at natural conversation [12], [17]–[25], thus extending the finding that gestures precede speech to natural, face-to-face conversation. Although the precise degree of temporal asynchrony differs across studies, possibly the result of using different languages, communicative contexts, gesture types, or lexical affiliate definitions, most studies do report that gestures start before their lexical affiliate.

To test whether the gestures that precede their lexical affiliates benefit language processing, we took response times during natural conversation as a rough measure of ease of processing. We found that questions with representational hand gestures got faster responses than questions without, replicating [11], which found the same pattern for English three-person conversations. The mechanisms underlying this pattern remain unclear, but it is for example possible that the gestures draw more attention to what is being said, or that there is additional semantic information in the gestures which facilitates message processing [11]. One other interpretation, which is especially interesting in the current context, is that the additional, earlier information in the gestures facilitates the prediction of upcoming words, and that recipients are therefore able to respond faster.

Along this line of thinking, we hypothesized that when a gesture precedes its lexical affiliate more, the recipient may have more time to process the gesture and use it to predict how likely possible upcoming words are. Therefore, we hypothesized that gestures with larger gesture onset-lexical affiliate onset asynchrony would be associated with faster conversational response times. The current sample does not support this hypothesis, but it is important to take into account that for this analysis only 74 gestures were available. For more conclusive results, it is necessary to gather more data, and we are currently in the process of doing this. Moreover, during conversation, several other factors also influence response times, including the durations of the question and response [39], syntactic complexity [39], or the question format [11]. Together, these and other factors may have masked an effect of gesture-speech asynchrony in these corpus data, which might be visible in more controlled experiments. More generally, further research is necessary to see whether, under which conditions and how recipients might use the visual bodily signals of the speaker to predict upcoming speech during face-to-face communication (e.g. [40]).

In sum, we found that during natural, face-to-face conversations, representational hand gestures have predictive potential, as gesture onsets and stroke onsets tended to start before the corresponding information in speech. Thus, it is possible that during conversation, recipients might not only use speech but also the communicative bodily movements of the speaker to facilitate predictive language processing [10]. This could help interlocutors to achieve the rapid turn-taking that is characteristic of human communication. To fully understand these processes, future research may explore the temporal relations and predictive functions of non-representational gestures (e.g. interactive and pragmatic gestures), which occur frequently in natural conversation [11]. Moreover, documenting the predictive potential of gestures in more languages is an important next step to see how generalizable the findings are across languages. Finally, it is important to investigate speech-gesture timing (and the effect on response times) in turn-taking contexts other than question-response sequences, since the coordination of turns may be characterised by different timing constraints depending on their sequential environment.

5. Acknowledgements

This work was supported by an ERC Consolidator grant (#773079, awarded to JH). We thank Mojenn Schubert and Mareike Geiger for assistance with coding.

6. References

[1] T. Stivers *et al.*, “Universals and cultural variation in turn-taking in conversation,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 26, pp. 10587–10592, Jun. 2009, doi: 10.1073/pnas.0903616106.

[2] P. Indefrey and W. J. M. Levelt, “The spatial and temporal signatures of word production components,” *Cognition*, vol. 92, no. 1–2, pp. 101–144, May 2004, doi: 10.1016/j.cognition.2002.06.001.

[3] S. Garrod and M. J. Pickering, “The use of content and timing to predict turn transitions,” *Front. Psychol.*, vol. 6, p. 751, Jun. 2015, doi: 10.3389/fpsyg.2015.00751.

[4] S. C. Levinson, “Turn-taking in human communication – Origins and implications for language processing,” *Trends Cogn. Sci.*, vol. 20, no. 1, pp. 6–14, Jan. 2016, doi: 10.1016/j.tics.2015.10.010.

[5] S. C. Levinson and F. Torreira, “Timing in turn-taking and its implications for processing models of language,” *Front. Psychol.*, vol. 6, p. 731, Jun. 2015, doi: 10.3389/fpsyg.2015.00731.

[6] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for

conversation," *Language*, vol. 50, no. 4, pp. 696–735, Dec. 1974, doi: 10.2307/412243.

[7] R. S. Gisladottir, D. J. Chwilla, and S. C. Levinson, "Conversation electrified: ERP correlates of speech act recognition in underspecified utterances," *PLoS ONE*, vol. 10, no. 3, p. e0120068, Mar. 2015, doi: 10.1371/journal.pone.0120068.

[8] L. Magyari and J. P. de Ruiter, "Prediction of turn-ends based on anticipation of upcoming words," *Front. Psychol.*, vol. 3, p. 376, Oct. 2012, doi: 10.3389/fpsyg.2012.00376.

[9] J. P. de Ruiter, Holger. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, vol. 82, no. 3, pp. 515–535, Sep. 2006, doi: 10.1353/lan.2006.0130.

[10] J. Holler and S. C. Levinson, "Multimodal language processing in human communication," *Trends Cogn. Sci.*, vol. 23, no. 8, pp. 639–652, Aug. 2019, doi: 10.1016/j.tics.2019.05.006.

[11] J. Holler, K. H. Kendrick, and S. C. Levinson, "Processing language in face-to-face conversation: Questions with gestures get faster responses," *Psychon. Bull. Rev.*, vol. 25, no. 5, pp. 1900–1908, Oct. 2018, doi: 10.3758/s13423-017-1363-z.

[12] E. A. Schegloff, "On some gestures' relation to talk," in *Structures of Social Action: Studies in Conversation Analysis*, J. M. Atkinson and J. Heritage, Eds. Cambridge, England: Cambridge University Press, 1984, pp. 266–296.

[13] A. Kendon, *Gesture: Visible action as utterance*. Cambridge, England: Cambridge University Press, 2004.

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

[15] M. W. Alibali, D. C. Heath, and H. J. Myers, "Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen," *J. Mem. Lang*, vol. 44, no. 2, pp. 169–188, Feb. 2001, doi: 10.1006/jmla.2000.2752.

[16] D. McNeill, "So you think gestures are nonverbal?," *Psychol. Rev.*, vol. 92, no. 3, pp. 350–371, 1985, doi: 10.1037/0033-295X.92.3.350.

[17] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, M. Key, Ed. The Hague, the Netherlands: Mouton, 1980, pp. 207–227.

[18] K. Bergmann, V. Aksu, and S. Kopp, "The relation of speech and gestures: Temporal synchrony follows semantic synchrony," in *Proc. 2nd Workshop Gesture and Speech in Interaction*, Bielefeld, Germany, 2011.

[19] P. Bernardis and M. Gentilucci, "Speech and gesture share the same communication system," *Neuropsychologia*, vol. 44, no. 2, pp. 178–190, Jan. 2006, doi: 10.1016/j.neuropsychologia.2005.05.007.

[20] R. B. Church, S. Kelly, and D. Holcombe, "Temporal synchrony between speech, action and gesture during language production," *Lang. Cogn. Neurosci.*, vol. 29, no. 3, pp. 345–354, Mar. 2014, doi: 10.1080/01690965.2013.857783.

[21] I. de Kok, J. Hough, D. Schlangen, and S. Kopp, "Deictic gestures in coaching interactions," in *Proc. Workshop Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, Tokyo, Japan, 2016, pp. 10–14, doi: 10.1145/3011263.3011267.

[22] M. Graziano, E. Nicoladis, and P. Marentette, "How referential gestures align with speech: Evidence from monolingual and bilingual speakers," *Lang. Learn.*, vol. 70, no. 1, pp. 266–304, Mar. 2020, doi: 10.1111/lang.12376.

[23] W. J. M. Levelt, G. Richardson, and W. La Heij, "Pointing and voicing in deictic expressions," *J. Mem. Lang*, vol. 24, no. 2, pp. 133–164, Apr. 1985, doi: 10.1016/0749-596X(85)90021-X.

[24] P. Morrel-Samuels and R. M. Krauss, "Word familiarity predicts temporal asynchrony of hand gestures and speech," *J. Exp. Psychol. Learn. Mem. Lang.*, vol. 18, no. 3, pp. 615–622, May 1992, doi: 10.1037/0278-7393.18.3.615.

[25] B. Butterworth and G. Beattie, "Gesture and Silence as Indicators of Planning in Speech," in *Recent Advances in the Psychology of Language*, R. N. Campbell and P. T. Smith, Eds. Boston, MA, USA: Springer US, 1978, pp. 347–360.

[26] G. Ferré, "Timing relationships between speech and co-verbal gestures in spontaneous French," in *Proc. 7th Int. Conf. Language Resources and Evaluation*, Valetta, Malta, 2010, pp. 86–91.

[27] K. Chui, "Temporal patterning of speech and iconic gestures in conversational discourse," *J. Pragma.*, vol. 37, no. 6, pp. 871–887, Jun. 2005, doi: 10.1016/j.pragma.2004.10.016.

[28] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a Professional Framework for Multimodality Research," in *Proc. 5th Int. Conf. Language Resources and Evaluation*, Genoa, Italy, 2006, pp. 1556–1559.

[29] T. Stivers and N. J. Enfield, "A coding scheme for question–response sequences in conversation," *J. Pragma.*, vol. 42, no. 10, pp. 2620–2626, Oct. 2010, doi: 10.1016/j.pragma.2010.04.002.

[30] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977, doi: 10.2307/2529310.

[31] H. Holle and R. Rein, "EasyDIAg: A tool for easy determination of interrater agreement," *Behav. Res. Meth.*, vol. 47, no. 3, pp. 837–847, Sep. 2015, doi: 10.3758/s13428-014-0506-7.

[32] P. Boersma, "Praat, a system for doing phonetics by computer.," *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, Nov. 2001.

[33] M. Seyfeddinipur, "Disfluency: Interrupting speech and gesture," Ph.D. dissertation, Radboud University, Nijmegen, 2006.

[34] S. Kita, I. van Gijn, and H. van der Hulst, "Movement phases in signs and co-speech gestures, and their transcription by human coders," in *Gesture and Sign Language in Human-Computer Interaction*, vol. 1371, I. Wachsmuth and M. Fröhlich, Eds. Berlin/Heidelberg, Germany: Springer Berlin Heidelberg, 1998, pp. 23–35.

[35] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Soft.*, vol. 67, no. 1, Oct. 2015, doi: 10.18637/jss.v067.i01.

[36] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Core Team, 2018.

[37] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in Linear mixed effects models," *J. Stat. Soft.*, vol. 82, no. 1, pp. 1–26, Dec. 2017, doi: 10.18637/jss.v082.i13.

[38] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *J. Mem. Lang*, vol. 68, no. 3, pp. 255–278, Apr. 2013, doi: 10.1016/j.jml.2012.11.001.

[39] S. G. Roberts, F. Torreira, and S. C. Levinson, "The effects of processing and sequence organization on the timing of turn taking: a corpus study," *Front. Psychol.*, vol. 6, p. 509, May 2015, doi: 10.3389/fpsyg.2015.00509.

[40] Y. Zhang, D. Frassinelli, J. Tuomainen, J. I. Skipper, and G. Vigliocco, "More than words: The online orchestration of word predictability, prosody, gesture, and mouth movements during natural language comprehension," 2020.