COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13407

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13407

Hand Gestures Have Predictive Potential During Conversation: An Investigation of the Timing of Gestures in Relation to Speech

Marlijn ter Bekke,[a,b] Linda Drijvers,[a,b] Judith Holler[a,b]

[a]*Donders Institute for Brain, Cognition and Behaviour, Radboud University*
[b]*Max Planck Institute for Psycholinguistics*

Received 23 December 2022; received in revised form 9 July 2023; accepted 10 January 2024

Abstract

During face-to-face conversation, transitions between speaker turns are incredibly fast. These fast turn exchanges seem to involve next speakers predicting upcoming semantic information, such that next turn planning can begin before a current turn is complete. Given that face-to-face conversation also involves the use of communicative bodily signals, an important question is how bodily signals such as co-speech hand gestures play into these processes of prediction and fast responding. In this corpus study, we found that hand gestures that depict or refer to semantic information started before the corresponding information in speech, which held both for the onset of the gesture as a whole, as well as the onset of the stroke (the most meaningful part of the gesture). This early timing potentially allows listeners to use the gestural information to predict the corresponding semantic information to be conveyed in speech. Moreover, we provided further evidence that questions with gestures got faster responses than questions without gestures. However, we found no evidence for the idea that how much a gesture precedes its lexical affiliate (i.e., its predictive potential) relates to how fast responses were given. The findings presented here highlight the importance of the temporal relation between speech and gesture and help to illuminate the potential mechanisms underpinning multimodal language processing during face-to-face conversation.

Keywords: Multimodal communication; Language; Prediction; Gesture; Turn-taking; Conversation

Correspondence should be sent to Marlijn ter Bekke, Donders Institute for Brain, Cognition and Behaviour, Thomas van Aquinostraat 4, 6525 GD, Nijmegen, the Netherlands. E-mail: Marlijn.terBekke@donders.ru.nl

1. Introduction

Face-to-face conversation involves rapid turn-taking, a human interactional capacity that has sparked much curiosity as to the factors that make this possible. Part of the mechanism facilitating fast turn exchanges seems to involve speakers predicting upcoming semantic information of the incoming turn, such that turn planning can begin before a current turn is complete (e.g., Bögels, Magyari, & Levinson, 2015; Corps, Crossley, Gambi, & Pickering, 2018; Levinson & Torreira, 2015). It has been proposed that recipients base such predictions of upcoming semantic information not only on speech, but also on speakers' communicative movements (Holler & Levinson, 2019; Skipper, 2014). Co-speech hand gestures, for example, can carry a significant amount of semantic information (Holler & Beattie, 2003; McNeill, 1992), making them a strong contender for playing a role in facilitating prediction and fast responding in conversation. We define prediction as context changing the state of the language processing system before new input becomes available, thereby facilitating the processing of this new input (Kuperberg & Jaeger, 2016). Therefore, for gestures to be part of the context used to predict upcoming semantic information, one crucial prerequisite is that the gesturally depicted information temporally precedes corresponding semantic information in speech. If speakers typically first say a word (e.g., "typing") and then produce a corresponding gesture (e.g., a typing gesture), gestures would be too late for recipients to use them to predict this word (e.g., "typing"). Thus, in this study, we present an analysis of the timing relation between gesture and speech, to test whether co-speech gestures could potentially be used for prediction of corresponding semantic information in speech. Moreover, we investigate the role of gestures in facilitating fast responses by looking at how their presence and timing relate to response times in conversation.

1.1. Prediction, early planning, and fast responding

Most language use occurs in face-to-face conversation, where the modal gap between questions and responding turns is 0–200 ms across languages (Heldner & Edlund, 2010; S. G. Roberts, Torreira, & Levinson, 2015; Stivers et al., 2009). This speed is astonishing, especially considering that language production takes time: preparing to produce a single word takes at least 600 ms (Bates et al., 2003; Indefrey, 2011; Indefrey & Levelt, 2004), while planning a short utterance takes around 1600 ms (Griffin & Bock, 2000). Therefore, response planning must already start before the incoming turn is complete.

How do interlocutors do this? One possibility is that they predict aspects of the content (and end) of the incoming turn, such that they can begin to process the message and plan their response before the incoming turn is complete (Garrod & Pickering, 2015; Levinson, 2016; Levinson & Torreira, 2015; Sacks, Schegloff, & Jefferson, 1974). Indeed, there is evidence that people make predictions about various aspects of linguistic messages while listening, such as the speech act (e.g., question, greeting, complaint; Gisladottir, Chwilla, & Levinson, 2015) and turn end (de Ruiter, Mitterer, & Enfield, 2006; Magyari & de Ruiter, 2012). Listeners can also predict upcoming words (Altmann & Kamide, 1999; DeLong, Urbach, & Kutas, 2005; Mantegna, Hintz, Ostarek, Alday, & Huettig, 2019; Nieuwland et al., 2020; Van

Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; but see Nieuwland et al., 2018). Moreover, studies show that in interactive settings, listeners begin to plan their responses as soon as possible (Barthel, Sauppe, Levinson, & Meyer, 2016, 2017; Barthel & Levinson, 2020; Bögels et al., 2015, 2018; Corps et al., 2018; Magyari, De Ruiter, & Levinson, 2017; Meyer, Alday, Decuyper, & Knudsen, 2018; Torreira, Bögels, & Levinson, 2015; but see Sjerps, Decuyper, & Meyer, 2020; Sjerps & Meyer, 2015). In this paper, we are interested in the role of hand gestures in these processes.

1.2. Are hand gestures part of the preceding context that could facilitate prediction of upcoming semantic information?

In face-to-face conversation, people also produce communicative signals with their bodies while talking, for example, by gesturing with their hands (Bavelas, 2022; Enfield, 2009; Kendon, 2004; McNeill, 1992). These co-speech hand gestures can represent semantic information (Kendon, 2004; McNeill, 1992), for example, by bringing a hand, shaped as if holding a glass, to the mouth to represent drinking, or may point to concrete or abstract locations or objects (deictic gestures) (Alibali, Heath, & Myers, 2001). These gestures, called representational gestures, form a tightly integrated system with speech both in production (Kita & Özyürek, 2003) and comprehension (Kelly, Özyürek, & Maris, 2010). Crucially, the semantic information that gestures convey is strongly related to the semantic information in speech (McNeill, 1985, 1992).

It has been proposed that recipients not only use speech but also co-speech visual communicative signals to facilitate predictive language processing (Holler & Levinson, 2019; Skipper, 2014). In this paper, we focus on whether representational hand gestures in particular may be part of the context that listeners use to predict upcoming semantic information. More specifically, gestures (e.g., a typing gesture) may be used to predict the corresponding semantic information to be conveyed in speech (typically called the lexical affiliate [Schegloff, 1984], e.g., “typing”). Representational gestures are a strong contender for playing a role in such prediction, because these gestures also depict semantic information, unlike, for example, rhythmic beat gestures.

Although gestures may play a role in the prediction of other aspects of speech, such as the turn end, in this paper, we specifically focus on the lexical affiliate, as that is the information in speech that is semantically most closely related to the information depicted by the gesture. We currently do not know at which level of linguistic representation such putative predictions based on hand gestures may occur: for example, it could be that hand gestures facilitate prediction at the level of concepts, but it could also be that specific lexical items are activated. Hereafter, we will refer to prediction as “prediction of the lexical affiliate,” but under the disclaimer that this does not mean we necessarily expect predictions to occur at the lexical level (they could also occur at the conceptual level, for example).

For hand gestures to play a role in the prediction of lexical affiliates during face-to-face conversation, it is crucial that gestures precede their lexical affiliates. Most previous research into gesture-speech timing either did not focus on conversation (Bergmann, Aksu, & Kopp, 2011; Bernardis & Gentilucci, 2006; Church, Kelly, & Holcombe, 2014; de Kok, Hough, Schlangen, & Kopp, 2016; Feyereisen, 1997; Graziano, Nicoladis, & Marentette, 2020; Levelt, Richardson, & La Heij, 1985; Liu & Kavakli, 2011; Morrel-Samuels & Krauss, 1992) or only described individual cases of preceding gestures (Ferré, 2019; Kendon, 1980; Schegloff, 1984; Zellers, Gorisch, House, & Peters, 2019). However, a few studies have looked at gesture-speech timing during conversations in a quantitative manner. In spontaneous French dialogues, it was found that 95% of iconic hand gestures started before lexical affiliate onset, on average preceding the lexical affiliate by 820 ms (Ferré, 2010). This finding aligns very well with data from English conversations about objects, where gesture preparations (i.e., the phase where speakers move the hands from the resting position to the gesture space) preceded their lexical affiliates by on average 820 ms (Donnellan et al., 2022).

To what extent gesture strokes (i.e., the most meaningful part of the gesture) start before the lexical affiliate shows a less consistent pattern. In spontaneous French dialogues, 72% of gesture strokes started before lexical affiliate onset, on average preceding the lexical affiliate by 454 ms (Ferré, 2010). During the English conversations about objects, gesture strokes on average preceded the lexical affiliate by 370 ms (Donnellan et al., 2022). During naturally occurring multilingual interactions at a Norwegian construction site, 51% of the iconic gesture strokes started before lexical affiliate onset, while the other strokes either started during or after lexical affiliate onset (Urbanik & Svennevig, 2021). By contrast, a study of Chinese multiparty conversations found that 60% of iconic gesture strokes synchronized with the lexical affiliate onset, while only 36% preceded it and 4% followed it (Chui, 2005). Thus, it is unclear to what extent gesture strokes start before their lexical affiliates, calling for further studies.

Moreover, while all these studies looked at conversation, most of them used a small number of speakers (Chui, 2005; Ferré, 2010; Urbanik & Svennevig, 2021), making it unclear how generalizable these results are to the wider population. In addition, while the conversations in the study by Donnellan et al. (2022) were unscripted as such, participants had to talk about specific objects, as elicited by the researchers. This is useful for studying how people communicate about objects, but in natural, face-to-face conversations people talk about a much wider range of topics, very often not focused on physical objects, and people have more varying communicative intentions than only conveying referent information. Moreover, as a result of their task, in the study by Donnellan and colleagues, it is likely that most gestures referred to physical objects. However, during free conversations, many gestures depict different meanings beyond physical objects, including actions, motion, spatial relations, locations, as well as many types of abstract concepts.

Our casual conversation corpus captures a wide range of gesturally depicted meanings common in everyday conversation, communicative intentions, and conversation topics, thus considerably expanding on previous studies. This significantly enhances the generalizability of conclusions drawn about gesture-speech timing, a very important step in advancing our understanding about whether prediction is a possible mechanism underlying multimodal language processing during conversation. If gestures and their most meaningful parts start before their lexical affiliates, they could potentially be used by recipients to predict the lexical affiliates before they are heard (i.e., gestures have predictive potential).

1.3. Could hand gestures facilitate fast responding?

If speakers' hand gestures can be used to predict lexical affiliates, these improved predictions might result in earlier response planning and faster articulation of these responses. Indeed, there is some evidence that perceiving multimodal input results in faster responses. For example, considering basic multisensory integration, participants are faster to detect low-level multimodal stimuli such as a beep and flash compared to their unimodal equivalents (e.g., Miller, 1986; Molholm, Ritter, Javitt, & Foxe, 2004; Romei, Murray, Merabet, & Thut, 2007; Senkowski, Molholm, Gomez-Ramirez, & Foxe, 2006; Suied & Viaud-Delmon, 2009). When it comes to speech and gesture, in some experimental studies, participants were faster to respond to stimuli containing speech and iconic gestures compared to speech only (Holle, Gunter, Rüschemeyer, Hennenlotter, & Iacoboni, 2008; Kelly et al., 2010; Krason, Fenton, Varley, & Vigliocco, 2021; Nagels, Kircher, Steines, & Straube, 2015; Wu & Coulson, 2005), but in some, they were not (Bernardis, Salillas, & Caramelli, 2008; Drijvers, Özyürek, & Jensen, 2018, 2019; He et al., 2015).

Moving more toward natural, face-to-face conversation, two studies deserve highlighting. In the first, participants shadowed speech from face-to-face conversations in an audiovisual context, an audiovisual context without visual speech (i.e., lip/mouth movements), and an audio-only context (Drijvers & Holler, 2023). Participants responded faster and made fewer errors when seeing audiovisual messages compared to audio-only. The other relevant study looked at response times to questions with and without gestures in a corpus of face-to-face, triadic conversations (Holler, Kendrick, & Levinson, 2018). Questions with hand gestures got faster responses than questions without, suggesting that hand gestures are indeed associated with fast responding.

However, it is still unclear whether questions with representational gestures, which could play a role in the prediction of upcoming semantic information, also get faster responses in conversation, as Holler et al. (2018) only looked at the combined effect of pragmatic, deictic, and representational gestures performed with the head and hands. In addition, little is known about whether conversational turns performing social actions other than questions also get faster responses when they are produced with a gesture. For such turns, the pressure to respond fast is not as strong as it is for questions (where response speed carries pragmatic meaning, see 1.4) and this may affect the role of gestures. Most importantly, it is unknown whether there might be especially fast responses for gestures that precede their lexical affiliate more (i.e., gestures that have greater predictive potential). When a gesture precedes its lexical affiliate more, the recipient may have more opportunity to use gestural information to predict this lexical affiliate and may thus respond faster. Of course, other mechanisms than prediction may explain such a pattern (e.g., gesture retractions may also occur earlier for those gestures that begin earlier, potentially acting as an early "go ahead" signal; Holler et al., 2018), but it would be a further piece of the puzzle in line with a prediction-based explanation.

1.4. Current study

The current study aims to fill these gaps in the literature by testing whether representational hand gestures temporally precede corresponding information in speech (i.e., have predictive potential), whether they result in faster responses and whether this is especially so for gestures with more predictive potential. In doing so, this study helps to illuminate the potential mechanisms underpinning multimodal language processing during face-to-face conversation and provides a crucial empirical foundation for controlled experiments aimed at establishing causality.

In a large corpus of Dutch unscripted dyadic conversations between acquaintances, we annotated the precise timing of representational hand gestures, their strokes (the most meaningful part), their lexical affiliates, the speech turns they were produced with, and the corresponding response turns. In addition to analyzing when strokes started, we also analyzed when gestures as a whole started, as gesture preparations can already contain information that recipients can use to disambiguate the gesture's meaning (Holler, Bavelas, Woods, Geiger, & Simons, 2022; Obermeier, Holle, & Gunter, 2011). We analyzed question-response sequences separately from conversational turns performing other social actions, such as statements and storytellings (hereafter: nonquestion turns). Questions are different in that by (Western) conversational norms, recipients are generally expected to provide a response and the speed with which they do so carries pragmatic meaning (e.g., delayed responses are typically associated with dispreferred responses; Kendrick & Torreira, 2015; F. Roberts, Francis, & Morgan, 2006; Schegloff, 2007). Therefore, during question-response sequences, the responder may make more use of the (predictive) information in the gesture and, therefore, gestures may play a larger role in facilitating fast turn transitions for question-response sequences compared to other sequential environments. To date, only a very recent study has examined response times for nonquestions turns in English triadic conversations, and found faster transitions between turn-constructional units when there was a gesture (Kendrick, Holler, & Levinson, 2023). Here, we intend to find out whether this also holds at the level of full turns, whether it generalizes to Dutch conversation, and whether the effect is similar across questions and nonquestion turns when we directly compare data from the exact same conversations.

We hypothesized that gestures and their strokes would typically start before their lexical affiliates (thereby replicating, e.g., Donnellan et al., 2022; Ferré, 2010; Urbanik & Svennevig, 2021). Moreover, we expected that questions with gestures get faster responses than questions without gestures (thereby replicating Holler et al., 2018). For gestures that precede their lexical affiliates more, the recipient may have more opportunity to use gestural information to predict the lexical affiliate and may thus especially respond faster. Given the differences between questions and nonquestion turns outlined above, we expected that gestures may play a smaller role in facilitating fast turn transitions for nonquestion turns compared to questions.

Preliminary results of this study were published in ter Bekke and colleagues (2020). Compared to these preliminary results, the current study makes several important advances. First and foremost, the amount of question data was tripled to provide us with improved statistical power, thus allowing us to draw firmer conclusions regarding the patterns we find (or lack thereof). Critically, we also included data from nonquestion turns, allowing us to test for the first time whether timing patterns are specific to questions or whether they extend to other turn types. Moreover, the additional data come from different conversation topics: in addition to free, entirely unguided conversation, we now also include conversations where people discussed opinions and viewpoints, thus issuing agreements and disagreements on certain topics, as well as conversations in which participants engaged in collaborative activity, namely, planning a joint holiday together. With this wider range of conversation topics and communicative intentions, our results are more representative of how language is used everyday.

2. Methods

2.1. Corpus and participants

The present analyses are based on a corpus of face-to-face conversations, consisting of 34 dyads of acquainted participants ($M_{\text{age}} = 23.10$, $SD = 8$, 51 female, 17 male). Participants were fluent speakers of Dutch who learned the language while growing up. Participants self-reported normal or corrected-to-normal vision, and no motoric or language problems. They gave informed consent prior to and after the recordings were made and were paid for their participation. Participants knew the focus of the study was on “patterns in the behaviour of people in conversations (including speech and bodily movements),” but nothing specific was mentioned about hand gestures. The corpus creation was approved by the Social Sciences Faculty Ethics Committee, Radboud University Nijmegen.

Each dyad conversed for 60 min. The first 20 min they held a free, entirely unguided conversation. In the second 20 min, participants discussed either privacy (e.g., how much privacy are you willing to sacrifice for convenience?), social media (e.g., do you ever consider deleting your accounts?) or Dutch versus English language in teaching (e.g., would you rather be taught in Dutch or in English?). These three discussion topics were chosen to spark discussion even among close friends. Participants were encouraged to discuss their opinions and explore their agreements and disagreements. In the third part, participants were asked to talk about the ideal, affordable holiday for them together. In this part, participants were first given 2 min to write their own ideas down, after which they discussed these ideas together for 20 min, aiming to find a compromise they both agreed on.

2.2. Apparatus

The conversations took place in a soundproof room, in which the participants sat opposite to each other (Fig. 1). Two cameras recorded each participant’s body from a 45 degree angle (Canon XF205 Camcorder), two cameras (Canon XF205 Camcorder) recorded each participant from a birds-eye view while mounted on a tripod, and finally, one camera (Canon Legria HF G10) recorded the scene view, displaying both participants at the same time. (Two further cameras recorded a close-up of each participant’s face, which were not used for the present analyses.) High-quality audio was recorded using two separate microphones (Sennheiser ME64) mounted on tripods close to the participants. Audio and video from each recording session were synchronized using Metus INGEST and Adobe Premiere Pro CS6. Moreover, audio from the two microphones was combined into one recording that contained the audio of both participants at a comparable volume, which was used for the present analyses. Videos were recorded at 25 frames per second, resulting in a time resolution of 40 ms (the duration of a single frame).

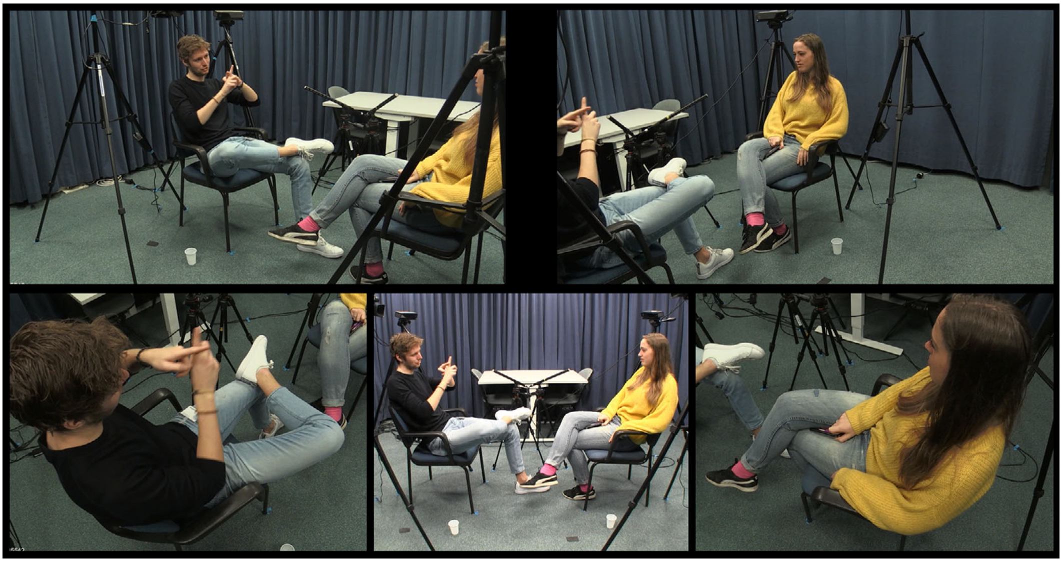

Fig. 1. Example of video material used for coding. Left participant gesturally depicts *crossed*. Top panels: body view, bottom left and right panels: bird view, bottom middle panel: scene view. Bottom panels were cropped in width for illustration purposes, but were actually wider.

2.3. Coding

The present analyses focused on the timing relations between representational gestures and their lexical affiliates, and on the response times during the conversations. The coding was done separately for gestures occurring during questions and other conversational turns. The sections below explain in detail how the coding was done.

All annotations were made in ELAN (version 5.5; Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006). For annotations involving speech, onset and offset were subsequently precisely annotated in Praat (version 5.1; Boersma, 2001). Example data and coding can be found in the Supplementary Materials. For all of the coding reliability reported below, we calculated raw agreement and modified Cohen's kappa between the main coder and the reliability coder using EasyDIAg (Holle & Rein, 2015). We used the EasyDIAg method because in the identification of, for example, a question, there are no segmented utterances for which coders need to decide whether it is a question or not. Rather, coders need to create these segmentations themselves based on a continuous stream of video data. EasyDIAg provides a solution to this problem by first linking annotations from coders together and then checking whether they agree on their judgment. Annotations from the coders were matched if they overlapped 60%, following Holle and Rein (2015). Cohen's kappa values of 0.61−0.80 were interpreted as substantial agreement, and values of 0.81−1.00 were interpreted as almost perfect agreement, following Landis and Koch (1977). In case of disagreement between the main coder and the reliability coder, coding from the main coder was used.

2.3.1. Questions

For the question data, video recordings were used from the free conversations, the discussions, and the holiday planning. We identified the question-response sequences using an existing coding manual (Stivers & Enfield, 2010) as the basis. Overall, we took a holistic approach to identifying questions, taking into account the utterance's phrasing (is it phrased like a question, e.g., word order, tags, WH-questions), intonation (does it sound like a question?), visual cues (does it look like a question?), context (does it make sense that this is a question in the context of the conversation?), pausing (does the speaker wait for a response to the question?), and addressee behavior (did the speaker respond as if it was a question?). When most of these factors pointed in the direction of a question, and as a whole the coder intuitively felt that an utterance was a question, then an utterance was coded as a question. We excluded questions that were not designed to obtain a response from the interlocutor, that is, self-directed questions (e.g., "What was it again?") and questions in reported speech (e.g., "I said: 'Why not?'") (Stivers & Enfield, 2010). While these utterances have the form of a question, in the conversations they did not function as questions in terms of aiming for a response from the interlocutor.

Responses to the questions were identified as anything that was verbalized in response to the question. Responses had to be contingent upon the question, for example, if after a question, the addressed speaker simply continued their previous talk, this was not coded as a response. Responses could be answers (e.g., "Yes," "No") or nonanswers (e.g., "I don't know," "What?"). Responses were annotated to include sounds such as "uh" but not nonverbal sounds, such as laughter, sighs, and lip smacks. The coding of utterances as questions and responses was reliable: based on 12% of the data, substantial agreement was reached between coders MtB and MG (questions: raw agreement 75%, modified Cohen's kappa 0.74; responses: raw agreement 73%, modified Cohen's kappa 0.73).

2.3.2. Nonquestion turns

Because nonquestion turns are very frequent in conversations, coding all nonquestion turns for all of our data (free conversations, discussions, holiday planning conversations) was not feasible. Therefore, we focused on the free conversations, as that was our most unconstrained conversation context. Within each free conversation, we coded nonquestion turns using a 5-min excerpt. To get these excerpts, we first randomly selected one of the minutes between the start of the conversation and 5 min before the end (e.g., if the conversation lasted 21 min, a random minute was selected between 0:00 and 16:00, such as 12:00). Second, to get a coherent excerpt, we determined the conversation topic at that point (e.g., at minute 12:00 they talked about a specific party), and searched for the start of that topic (e.g., 10:37). Finally, the excerpt was selected from the start of the topic until 5 min later (e.g., 10:37−15:37).

Within the 5-min excerpts, nonquestion turns were coded. Nonquestion turns were defined as stretches of speech by one speaker that were not a question or a response to a question (to avoid overlap with the question-response dataset), with the exception of self-directed questions and questions in reported speech (see Section 2.3.1). The nonquestion turns could, for example, include short statements (e.g., "I like your dress"), but also more elaborate stories. Backchannels (Yngve, 1970; e.g., "hm-hm," "yeah," "okay") were not coded as turns: they serve functions, such as displaying attention, comprehension, or stance, while the speaker passes up the opportunity to take a full turn (Bangerter & Clark, 2003; Edelsky, 1981; Tolins & Fox Tree, 2016). The start of a nonquestion turn was the start of the speech. The end of the nonquestion turn was when the speaker stopped talking because the other speaker uttered more than a backchannel (i.e., a full turn). Intraspeaker silences were coded as part of a turn (ten Bosch, Oostdijk, & Boves, 2005).

For each coded nonquestion turn, we also coded the next turn to be able to calculate turn transition gaps/overlaps for nonquestion turn sequences. A next turn was defined as the stretch of speech uttered by the other speaker immediately after the first speaker's turn. As above, backchannels were not included. Although questions were excluded from the category of nonquestion turns, they could sometimes be next turns after nonquestion turns (e.g., after a statement, the next turn was a question, and the turn transition gap/overlap between them was used for analysis). Collaborative completions (Lerner, 1991) were coded as next turns only if they were acknowledged by the first speaker (in line with Maynard [1997] who argues that minor additions are backchannels).

2.3.3. Gestures

We coded representational gestures, which are gestures that "depict semantic information by virtue of handshape, placement, or motion" (Alibali et al., 2001, pp. 173). Representational gestures included iconic gestures which depict concrete referents (e.g., actions or objects; McNeill, 1992), and metaphoric gestures which depict abstract concepts (McNeill, 1992). For an example of such a representational gesture, see how the participant on the left gesturally depicts the concept "crossed" in Fig. 1. It also included deictic gestures (McNeill, 1992), which are pointing gestures to concrete entities (e.g., the door of the room one is in) or which locate meaning abstractly in gesture space (e.g., while saying "Will we go to your house or to mine?" making two pointing gestures with the left and right hand referring to fictive locations in gesture space). The category of representational gestures excluded gestures that fulfill pragmatic (Kendon, 2004; including beats, McNeill, 1992) or interactive (Bavelas, Chovil, Coates, & Roe, 1995; Bavelas, Chovil, Lawrie, & Wade, 1992) functions. It also excluded emblems which have a conventional form and meaning that is culturally specified (McNeill, 1992).

For the question data, we looked through the videos in their entirety, and for each representational gesture, we determined whether it related in meaning to a nearby question. For example, someone said "And then I lost you, and then I arrived there, and then you arrived with Thijs and Niek, and Sam arrived with Koen, *how did you get [separated]?*" while producing a gesture involving the hands moving apart (stroke timing indicated by square brackets). The gesture was judged to mean "separated," so it was part of the question in italics, and not part of the speech preceding the question. This way, we did not decide on an a priori time window about how far away from questions we might still expect related representational gestures. We found 439 representational gestures that were judged to be part of questions. For the 5-min excerpts, we annotated all representational gestures that were part of nonquestion turns ($n = 359$).

The gesture coding was reliable. Coder MtB coded the gestures in the free conversations (questions and nonquestion turns), and got 82% agreement with independent coder MG based on 11% of the gestures (*n* = 71). Coder ME coded the gestures in the discussions and holiday planning conversations (questions). They achieved good reliability with coder MtB and then proceeded to code the discussions and holiday planning conversations. This procedure of reliability-before-coding has precedence in the gesture literature (e.g., Bavelas et al., 1992, 1995, 2014). Coder ME got 85% agreement with coder MtB using 21% of the gestures occurring during questions in the free conversations (*n* = 60). Cohen's kappa could not be calculated because there was only one gesture category.

2.3.4. *Gesture phases*

Next, the gestures were segmented into gesture phases using a frame-by-frame method in which first dynamic and static phases were segmented based on whether the image of the hands is blurry or clear (Seyfeddinipur, 2006). Then, the segmented phases were coded as one of the following: preparation, pre-stroke hold, stroke, post-stroke hold, or retraction (Kita, van Gijn, & van der Hulst, 1998). Crucial for the current analyses are the onset of the gesture as a whole, and the onset of the stroke phase, in which the meaning of the gesture is expressed. Overall, the first frame of a gesture was typically the first frame of the preparation, in which the hands moved away from their resting position and were blurry. The first frame of a stroke was the first frame in which the meaning of the gesture was expressed. If the gesture contained multiple, directly consecutive strokes, the onset of the first stroke was annotated for analysis. The last frame of a gesture was the first frame in which the hands were motionless in their rest position.

The gesture phase coding was reliable. For the question data from the free conversations, average disagreement between coder MS and independent coder MG on when gestures started was 37 ms (raw agreement 89%), based on 21% of the gestures (*n* = 60). For strokes, average disagreement was 55 ms (raw agreement 88%, modified Cohen's kappa 0.71). For the nonquestion turns, average disagreement between coder MtB and independent coder MG on when gestures started was 12 ms (raw agreement 99%), based on 21% of the gestures (*n* = 76). For strokes, average disagreement was 66 ms (raw agreement 96%, modified Cohen's kappa 0.90). Coder ME coded the gestures in the discussions and holiday planning (questions). Before starting this, they got average disagreement on gesture onset of 38 ms (raw agreement 90%) with coder MS based on 21% of the question gestures during free conversations (*n* = 60). For strokes, average disagreement was 55 ms (raw agreement 86%, modified Cohen's kappa 0.64).

2.3.5. *Lexical affiliates*

For each gesture, its lexical affiliate was determined, which was defined as the word(s) deemed to correspond most closely to the gesture in meaning (Schegloff, 1984). We acknowledge that the notion of the lexical affiliate is a tricky one, since gestures depict meaning holistically, which often cannot be pinned down to the meaning of a single lexical item. Despite being established in the literature, the notion of lexical affiliate thus has to be seen as an attempt to approximate what the gestural depiction is about or relates to.

Our strategy for determining the lexical affiliate was to first see what information the gesture depicted, and to then choose the corresponding word(s). For example, if a gesture depicted an action, we chose the corresponding action verb. To illustrate, when the gesture imitated writing while the speaker said "Hij heeft die verhalen dan als antwoord op die vragen geschreven" (He has written those stories then in response to those questions), the word "geschreven" (written) was chosen as lexical affiliate. Often, gestures do not just depict either an action or an object due to their holistic nature. If a gesture could be related to multiple words in speech, coders selected the word the gesture was foregrounding with its depiction in the given conversational context. For example, suppose someone said "No, the ball was really big" with prosodic emphasis on the word "big." If while saying "big," they made a gesture with two hands as if holding a really big ball, the lexical affiliate could either be "ball" or "big." In this case, the size aspect was judged to be foregrounded, so coders would pick "big" as the lexical affiliate.

Following Bergmann et al. (2011), we excluded articles from lexical affiliate selection and we omitted prepositions as far as possible if no information was lost. Moreover, the lexical affiliate did not involve the amount of entities (Bergmann et al., 2011), unless number information was foregrounded in the gesture (e.g., two fingers to represent "two," or two hands next to each other in a cupped manner, palms pointing down, to represent "two cars"). 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, for example, the lexical affiliate would be "bridge" instead of "that bridge." Note that demonstratives were not fully excluded from lexical affiliate selection: if the demonstrative was the best lexical affiliate (e.g., pointing + "what is this?"), it was selected as the lexical affiliate. For exploratory analyses on the difference between such underspecified lexical affiliates (e.g., "thing," "that," "it," "here") and more specified lexical affiliates (e.g., "backpack") in terms of gesture-speech timing and conversational response times, see Tables S2−S5.

For 42 question gestures (9.57%), the lexical affiliate could not be determined, either because the information in the gesture was not present in speech due to the gesture's complementary nature (38, 8.66%) or because the gesture was too ambiguous (4 gestures, 0.91%). An example of such a gesture with no lexical affiliate was when a speaker said "But within one…?" and then made large a circle with both hands to metaphorically depict an overarching theme. This gesture is part of the question but adds information to speech. For 39 gestures in nonquestion turns (10.87%), the lexical affiliate could not be determined, because the information in the gesture was not present in speech (36 gestures, 10.03%) or the gesture was too ambiguous (3 gestures, 0.84%).

The lexical affiliate coding was reliable. Coder MtB coded the lexical affiliates in the free conversations, and got 95.4% agreement with independent coder MG for the questions (modified Cohen's kappa 0.95) based on 21% of the data. Moreover, they got 95.5% agreement for the nonquestion turns (modified Cohen's kappa 0.95), based on 21% of the data. Before starting to code the data from the discussions and holiday planning, coder ME coded the gestures in the discussions and holiday planning (questions) and got 77.5% agreement with coder MtB using 21% of question data from the free conversations (modified Cohen's kappa 0.77).

2.3.6. Response times

To investigate whether questions/nonquestion turns with gestures got faster responses than those without gestures, we needed the response times for analysis. For the analysis of response times to questions (*number of questions* = 5768), we excluded questions for which the response time could not be reliably calculated in terms of a gap between pragmatically contingent vocal contributions, such as questions that did not get a verbal response (e.g., nonverbal response like nodding, or rhetorical questions, $n = 947$). We also excluded cases of multiple questions in a row that got a single response ($n = 756$), and questions that got multiple responses ($n = 35$). For these, the linking between question and response—and thus the calculation of the response time—is not unequivocally possible. Finally, we also excluded cases where the speaker who asked the question continued talking after the question end ($n = 149$), since this may have impacted on the response time.

For the response times to nonquestion turns (*number of nonquestion turns* = 147), there were also cases for which the response time could not be reliably calculated and which were, therefore, excluded from the response time analysis. This was the case when there was no clear turn transition ($n = 13$; e.g., when both speakers started speaking simultaneously, or when there was a long pause in which the speakers appeared to search for a topic to talk about, i.e., lapses, Hoey, 2020). To prevent overlap between the question data and the data from nonquestion turns, we also excluded turns that ended in a question ($n = 5$). For example, one speaker said "Sometimes [I now already had to go up such a small bri]dge and then I was [really already] with all my power on my toes [uhhh] heavily pushing, let alone eh if there eh would have been snow, it was also, but in my opinion, I don't know if really more snow has fallen here?" and produced one gesture depicting the bridge and two depicting the action of pushing (stroke timing for each gesture indicated by square brackets). If we look at the response to this utterance ("I also don't know"), it is a response to the question ("I don't know if really more snow has fallen here?") and not to the speech before that which included the gestures. Including these cases in the nonquestion data would have created overlap between the question and nonquestion data. Also excluded were cases where the gesture occurred with speech in between a question and a response ($n = 3$). For example, one speaker asked "Is Liz then even back already?" and then started counting on her fingers "January, February, March," after which the other person responded with "No."

For the response times comparison, we also coded nonquestion turns without gestures and their next turns. For feasibility, we did not code all nonquestion turns without gestures in the excerpts. Instead, for each nonquestion turn sequence with gesture, we selected a nonquestion turn sequence without gesture that occurred right before (50%) or after the gesture turn sequence (50%), because conversation topic is known to affect response times (Strömbergsson, Hjalmarsson, Edlund, & House, 2013). In total, we used 126 nonquestion turns with gestures (and their next turns) and 126 nonquestion turns without gestures (and their next turns) for the analyses.

Next, for all questions/nonquestion turns and their response/next turns, the conversational response times were calculated. Using the python package pympi (version 1.70; Lubbers & Torreira, 2018), response times were calculated as the distance between the end of a question/nonquestion turn and the start of its response/next turn. Negative response times indicated overlaps, positive response times indicated gaps (S. G. Roberts et al., 2015).

2.3.7. Placeholders

It is possible that faster responses are the result of the use of turn-initial placeholders (S. G. Roberts et al., 2015), such as “uh,” which recipients can use to minimize the response time while taking extra time to plan the response. To check whether faster responses to questions with gestures could be the result of these placeholders, we recalculated response times with the turn-initial placeholders “uh,” “um,” “hm,” and “well” (Dutch: “nou”) not being counted as the first word of the response (S. G. Roberts et al., 2015; Strömbergsson et al., 2013). For the response “Um well I think so,” the response would start at “I.” We used both the original response times and these recalculated response times without placeholders for the analyses.

We coded placeholders for the question data and the nonquestion turn data from the free conversations. For each question with a gesture, we selected a question without gesture from the same speaker that occurred right before or after it, for comparison (as was done for the nonquestion turns, see Section 2.3.6. Response Times). For questions with ($n = 80$) and without gestures ($n = 79$), respectively, 5.0% and 15.2% of the responses started with a placeholder. For the nonquestion turns with ($n = 126$) and without gestures ($n = 126$), respectively, 1.6% and 0.8% of the next turns started with a placeholder.

2.4. Analysis

First, we asked whether gesture onsets and gesture strokes preceded 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 asked whether questions and nonquestion turns with gestures got faster responses, and whether this is especially so for gestures that preceded their lexical affiliate more (i.e., greater predictive potential).

We fitted linear-mixed effects models using the lme4 package (version 1.1.30; Bates, Mächler, Bolker, & Walker, 2015) in R (version 4.2.1; R Core Team, 2019). *p*-Values were obtained with the package lmerTest (version 3.1.3; Kuznetsova, Brockhoff, & Christensen, 2017). To analyze whether gesture-lexical affiliate asynchrony and stroke-lexical affiliate asynchrony related to response times (Section 3.4), we ran multiple tests on the same data and corrected for multiple comparisons using the False Discovery Rate, implemented in R with the function *p.adjust*. When gesture presence was a factor, no gesture was set as the reference level. The factor question/nonquestion turn was sum-to-zero contrast coded (0.5 question; −0.5 nonquestion turn). We used the maximal random effects structures (Barr, Levy, Scheepers, & Tily, 2013), which included random intercepts and slopes for fixed effects by participant nested within dyad. Conversation topic (free conversation, discussion, holiday planning) was also included as random factor. For the analyses of gesture-speech asynchrony, we used models with an intercept, question/nonquestion turn as factor, and the maximal random effects structure. Here a significant intercept means that the gesture-speech asynchrony significantly differs from zero, zero being that gestures and their lexical affiliate start at the same time. For the analysis of response times, the interaction between the factors gesture presence and question/nonquestion turn was added to the model.

When maximal models led to convergence issues, we first increased the number of iterations, then used the estimates from the nonconverged fit as the new starting values, and finally compared the estimates from different optimizers using allFit(). If all optimizers converged to highly similar values, then we considered the convergence warnings to be false positives, following the lme4 package documentation (RDocumentation, 2021). For all models reported below, this last step showed highly similar values, and convergence warnings were treated as false positives. Therefore, all final models contained the maximal random effects structure. For the final models, visual inspection was used to check for violations of homoscedasticity or residual normality and linearity.

To see which data points were used for which analysis, see Table 1. All data and analysis scripts are available on OSF (https://osf.io/f9qm6/).

3. Results

3.1. Did gesture onsets precede lexical affiliate onsets?

The overwhelming majority of gestures started before their lexical affiliate (94%), and on average gestures started 644 ms before ($SD = 648$ ms). This pattern was similar across questions (Fig. 2A) and nonquestion turns (Fig. 2B), with gestures starting on average 685 ms before their lexical affiliates during questions ($SD = 653$ ms) and 593 ms before during nonquestion turns ($SD = 639$ ms). Within the questions, the timing was similar across conversation topics, with gestures starting on average 724 ms before in the free conversations ($SD = 730$ ms), 594 ms before in the discussions ($SD = 540$ ms), and 706 ms before in the holiday planning conversations ($SD = 627$ ms). The linear mixed effects model revealed that overall, gesture onsets significantly preceded lexical affiliate onsets ($\beta = -642.28$, $SE = 27.06$, $t = -23.74$, $p < .001$). Whether the gesture occurred during a question or nonquestion turn did not impact this timing ($\beta = -90.07$, $SE = 59.68$, $t = -1.51$, $p = .14$).

3.2. Did stroke onsets precede lexical affiliate onsets?

It could be that the gesture onsets preceded lexical affiliate onset, simply because the gesture preparation preceded the lexical affiliate, while the gesture stroke occurred during the lexical affiliate. To test whether strokes also preceded their lexical affiliates, we ran a similar analysis for the gesture stroke timing. The majority of strokes (60%) started before their lexical affiliate, and on average strokes started 193 ms before ($SD = 616$ ms). This pattern was similar across questions (Fig. 2C) and nonquestion turns (Fig. 2D), with strokes starting on average 201 ms before their lexical affiliates during questions ($SD = 620$ ms) and 184 ms before during nonquestion turns ($SD = 612$ ms). Within the questions, the timing was

Table 1
Overview of datasets and number of data points used for each analysis

	Numbers	*Used for*
Question data		
Total set of questions	1. 5768 questions 2. containing 439 gestures 3. of which 397 had a lexical affiliate	1. - 2. - 3. Did gesture onsets precede lexical affiliate onsets? (3.1) + Did stroke onsets precede lexical affiliate onsets? (3.2)
Questions after exclusions[a]	3881 questions, of which 205 have at least one gesture (with or without lexical affiliate)	Did questions and nonquestion turns with gestures get faster responses? (3.3)
Placeholder coding	80 questions with gesture, 79 questions without (from the free conversations)	Did questions and nonquestion turns with gestures get faster responses? (3.3.1)
Questions after exclusions[a] with gestures with lexical affiliates	191 questions with at least one gesture with a lexical affiliate	Did questions with more preceding gestures get faster responses? (3.4)
Nonquestion turn data		
Total set of turns	1. 147 turns 2. containing 359 gestures 3. of which 320 had a lexical affiliate	1. - 2. - 3. Did gesture onsets precede lexical affiliate onsets? (3.1) + Did stroke onsets precede lexical affiliate onsets? (3.2)
Turns after exclusions[a]	252 nonquestion turns, of which 126 contain at least one gesture (with or without lexical affiliate)	Did questions and nonquestion turns with gestures get faster responses? (3.3)
Placeholder coding	252 nonquestion turns, of which 126 contain at least one gesture (with or without lexical affiliate)	Did questions and nonquestion turns with gestures get faster responses? (3.3.1)

Note. The hyphen (-) indicates that the corresponding data selection described in the column Numbers was not used for any of the reported analyses.
[a]For details on the exclusions, see Section 2.3.6. Response times.

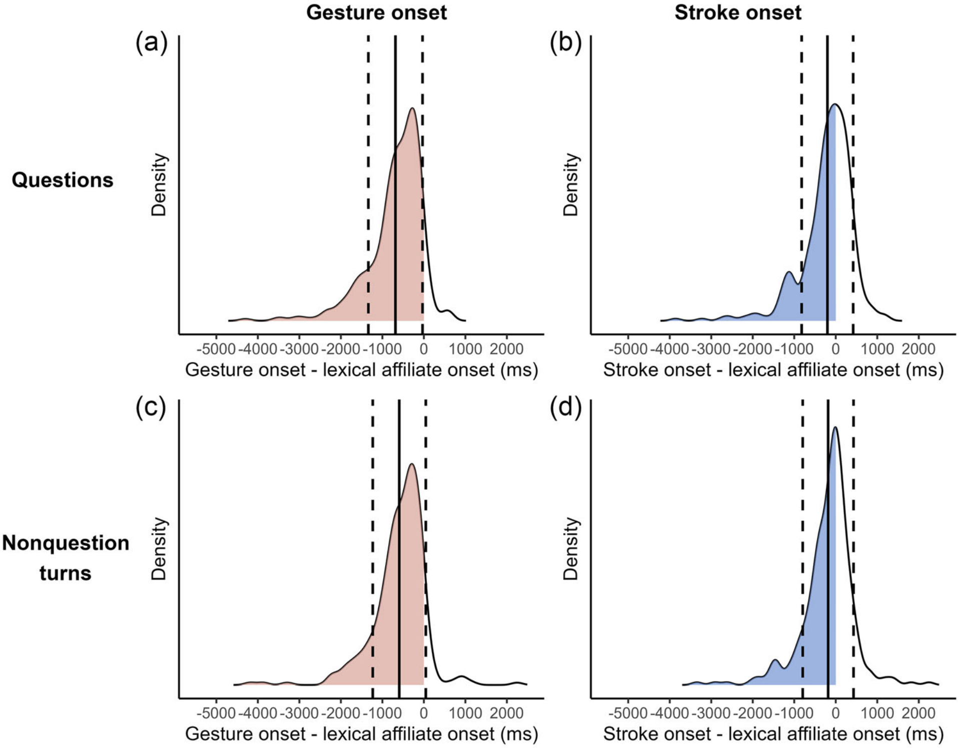

Fig. 2. Gesture onsets and gesture strokes tended to precede the onset of their lexical affiliate for gestures in questions as well as nonquestion turns. The figures show density plots of gesture-speech temporal asynchrony for gesture onset (A/B) and stroke onset (C/D), for gestures in questions (A/C) and nonquestion turns (B/D). The solid lines indicate the mean asynchronies, the dashed lines indicate a distance of one standard deviation from the mean. The colored sections indicate the gestures/strokes that started before lexical affiliate onset.

similar across the conversation topics, with strokes starting on average 232 ms before in the free conversations ($SD = 685$ ms), 149 ms before in the discussions ($SD = 534$ ms), and 202 ms before in the holiday planning conversations ($SD = 598$ ms). The linear mixed effects model revealed that stroke onset significantly preceded lexical affiliate onset ($\beta = -200.52$, $SE = 27.73$, $t = -7.23$, $p < .001$). Whether the gesture occurred during a question or nonquestion turn did not impact this timing ($\beta = -10.16$, $SE = 63.61$, $t = -0.16$, $p = .87$). Thus, not only gesture onsets as a whole, but also gesture strokes typically started before their corresponding information in speech (Fig. 3).

3.3. Did questions and nonquestion turns with gestures get faster responses?

To investigate whether questions and nonquestion turns with gestures got faster responses, we analyzed response times to questions and nonquestion turns with and without representational gestures (for descriptives, see Table 2). Questions with gestures got around

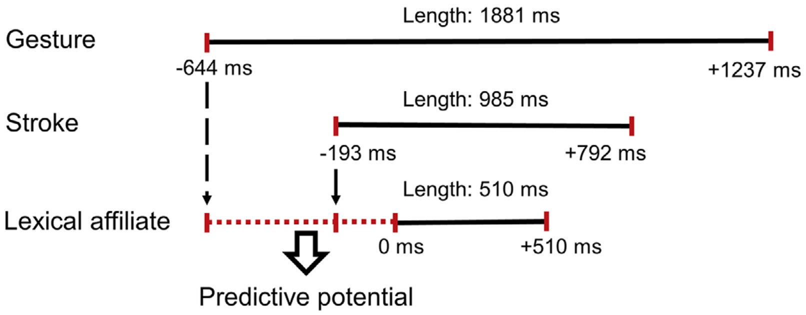

Fig. 3. Mean timing relations between representational gestures, their strokes, and their lexical affiliates. Values were taken from all gestures together: from questions and nonquestion turns. In the time window indicated by the red dotted line, gestures have predictive potential and could possibly be used by recipients to predict the lexical affiliate.

Table 2
Means, ranges, and standard deviations (SD) for response times (RT) to questions and nonquestion turns: overall, with representational gestures and without representational gestures

	Overall mean RT	With representational gestures	Without representational gestures
Questions	362 ms (*SD* = 650 ms, range = [−6191 ms, 4348 ms], *n* = 3881)	114 ms (*SD* = 727 ms, range = [−2979 ms, 2658 ms], *n* = 205)	376 ms (*SD* = 642 ms, range = [−6191 ms, 4348 ms], *n* = 3676)
Nonquestion turns	249 ms (*SD* = 757 ms, range = [−2240, 3385 ms])	231 ms (*SD* = 810 ms, range = [−1865, 3385 ms])	267 ms (*SD* = 702 ms, range = [−2240, 3076 ms])

260 ms faster responses than questions without gestures (Fig. 4). However, nonquestion turns with and without gestures got similarly fast responses (Fig. 5). The linear mixed effects model revealed that questions with gestures got significantly faster responses, but nonquestion turns did not ($\beta = -246.83$, $SE = 106.08$, $t = -2.33$, $p = .02$). For exploratory analyses of response times to questions with and without gestures split by conversation topic, see Table S1.

3.3.1. Placeholders

To check whether faster responses to questions with gestures could be the result of recipients quickly responding with turn-initial placeholders (“uh,” “um,” “hm,” “well”), for the subset of free conversations, we recalculated response times without including any turn-initial placeholders (see method). In this recalculated data, the average response time to questions was 226 ms (*SD* = 720 ms, range = [−2979 ms, 2577 ms]). The average response time to questions with representational hand gestures was 43 ms (*SD* = 732 ms, range = [−2979 ms, 2577 ms]). The average response time to questions without representational hand gestures was 409 ms (*SD* = 661 ms, range = [−1827 ms, 2462 ms]). The model revealed that questions with gestures get significantly faster responses, where questions with gestures get responses that are around 361 ms faster ($\beta = -361.29$, $SE = 100.20$, $t = -3.61$, $p < .001$).

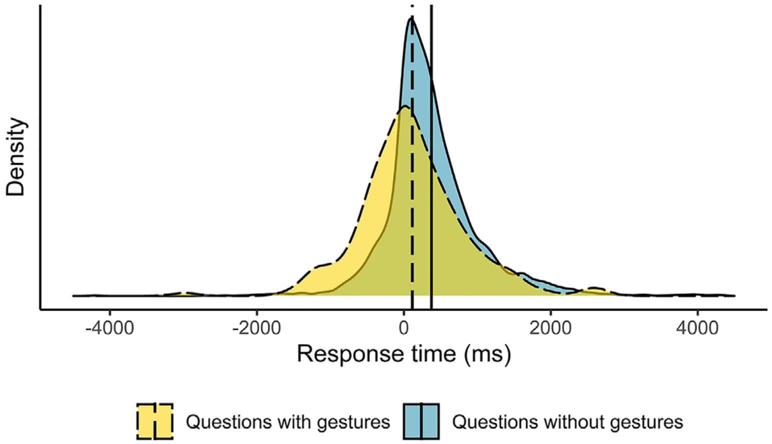

Fig. 4. Questions with gestures got faster responses than questions without gestures. The plot shows the distribution of the response times for questions with (dashed line) and without gestures (solid line), in milliseconds. Negative values indicate overlap between question and response, positive values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

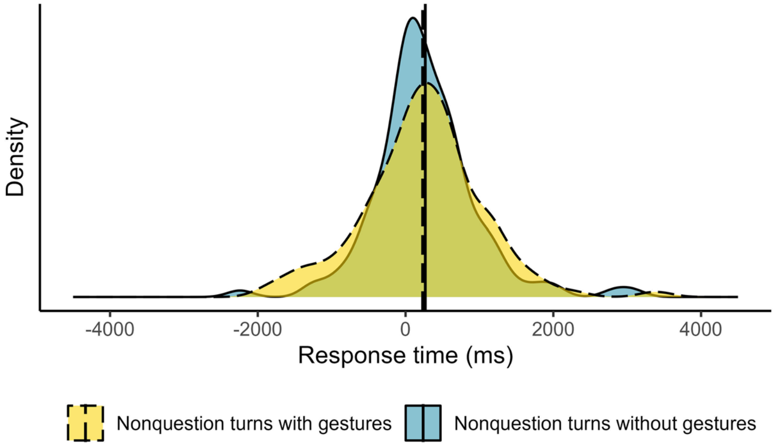

Fig. 5. Distribution of the response times for nonquestion turns with (dashed line) and without gestures (solid line), in milliseconds. Negative values indicate overlap between turn and next turn, positive values indicate gaps. The vertical lines represent the mean response time to nonquestion turns with (dashed line) and without gestures (solid line).

Thus, questions with gestures did not simply get faster responses because recipient quickly respond with “uh,” “um,” or “well.”

For the nonquestion turns, only 3 out of 252 next turns started with a placeholder, and, therefore, the response times were barely affected by the exclusion of placeholders. Therefore, we deemed it not necessary to repeat the analysis without including turn-initial placeholders.

3.3.2. Nonquestion turns: Follow-up analysis

One possibility for why questions with gestures got faster responses, but nonquestion turns with gestures did not, could be that the nonquestion turns were often much longer than the questions. Therefore, gestures at the start of a turn may be relatively further away from the turn transition. Indeed, gestures occurring during questions on average started 2353 ms before question offset (*SD* = 1497 ms), while gestures occuring during nonquestion turns on average started 13,215 ms before turn offset (*SD* = 13,052 ms). To investigate whether nonquestion turns with gestures closer to turn offset did get faster responses than nonquestion turns without gestures, we ran the same analysis on a subset of nonquestion turns for which the gestures were similarly far from the turn offset as the gestures in questions were from question offset.

For each gesture, we calculated the distance in milliseconds between the gesture onset and the turn offset. Next, we selected gestures/nonquestion turns such that this distance distribution had a similar mean and standard deviation compared to the gestures during questions (*M* = 2353 ms before question offset, *SD* = 1497 ms). For each gesture, we calculated whether adding it to the subset would bring the mean and standard deviation of the distance closer to the mean and standard deviation of the distance for gestures during questions. If adding a gesture decreased the difference in means + the difference in SDs, it was added. To achieve the subset with as many gestures as possible, we started with the gestures closest to turn offset.

For the subset of nonquestion turns, the gesture started on average 2302 ms before turn offset (*SD* = 1076 ms), which is much more comparable to the questions. This subset contained 58 nonquestion turns with gesture and 58 nonquestion turns without gesture for comparison (out of 126 nonquestion turns with gesture and 126 nonquestion turns without gesture). Nonquestion turns with gesture could contain multiple gestures, but they had at least one gesture that fell within this gesture onset—question offset range.

In this subset, the average response time to nonquestion turns was 207 ms (*SD* = 801 ms, range = [−2240, 2824 ms]). The average response time to nonquestion turns with representational hand gestures was 154 ms (*SD* = 831 ms, range = [−1758, 1847 ms]). The average response time to nonquestion turns without representational hand gestures was 260 ms (*SD* = 774 ms, range = [−2240, 2824 ms]). There seems to be a small beneficial effect for nonquestion turns with gestures. However, a model with response time as dependent variable and gesture presence as fixed effect revealed no significant difference in response time between nonquestion turns with and without representational gestures in this subset ($\beta = -107.59$, $SE = 138.00$, $t = -0.78$, $p = .44$). Thus, even when the gestures were comparably close to the turn end, nonquestion turns with gestures did not get faster responses. Therefore, we did not perform the follow-up analysis of whether gesture-speech asynchrony relates to response times to nonquestion turns.

3.4. Did questions with more preceding gestures get faster responses?

One reason why questions with gestures might get faster responses is because recipients might use the information in the gesture to predict the lexical affiliate. If this is the case, then the response time to questions with gestures might be related to the degree to which the gestures precede (and thus possibly predict) their lexical affiliate.

Out of the 205 questions that were produced with a gesture, 191 were produced with a gesture that had a lexical affiliate. Thus, for these questions, we had both a response and a gesture-speech asynchrony value. When a question was produced with multiple gestures (this was the case for 44 out of the 191 questions), we related the gesture-speech asynchrony value of the first gesture that had a lexical affiliate to the response time, because the recipient had most time to process the first gesture, and thus we reasoned that the first gesture might be most likely to relate to response times. We ran two linear mixed effects models with response time as dependent variable and gesture-speech asynchrony as fixed effect. In the first model, asynchrony was calculated as stroke onset—lexical affiliate onset. Stroke-lexical affiliate asynchrony did not relate to response times ($\beta = 0.15$, $SE = 0.13$, $t = 1.21$, $p = .48$).

Because it is unknown whether only gesture strokes might be used for prediction, or whether gesture preparations already contain relevant semantic information, in the second model, we calculated asynchrony as gesture onset (as a whole)—lexical affiliate onset. Gesture-lexical affiliate asynchrony also did not relate to response times ($\beta = 0.11$, $SE = 0.13$, $t = 0.82$, $p = .48$).

4. Discussion

Face-to-face conversation involves rapid turn-taking. Theories propose that to respond so quickly, listeners predict aspects of the unfolding turn, such that response planning can start as soon as possible (e.g., Levinson & Torreira, 2015). During face-to-face conversation, interlocutors also communicate with co-speech hand gestures, which carry a significant amount of semantic information. This makes them a strong contender for playing a role in facilitating prediction and fast responding (Holler & Levinson, 2019). In this study, representational gestures and their most meaningful parts (strokes) typically started before the corresponding semantic information in speech, thereby providing them with predictive potential. Moreover, gestures were associated with fast responding during question−response interactions. However, we found no evidence for the idea that how much a gesture preceded its lexical affiliate (i.e., its predictive potential) related to how fast responses were given.

4.1. *Gestures have predictive potential*

Representational gestures as a whole as well as their most meaningful parts (stroke phases) started before the most closely corresponding information in speech. The semantic information these gestures shared with speech was thus already visible in the gesture before it was vocalized. This finding replicates earlier work indicating that gestures typically precede their lexical affiliates (Bergmann et al., 2011; Bernardis & Gentilucci, 2006; Church et al., 2014; de Kok et al., 2016; Donnellan et al., 2022; Ferré, 2010; Graziano et al., 2020; Kendon, 1980; Levelt et al., 1985; Morrel-Samuels & Krauss, 1992; Schegloff, 1984; Urbanik & Svennevig, 2021). Importantly, we show for the first time that this pattern also holds in a corpus of natural, face-to-face conversation in which a large number of participants talked in a range of communicative situations (corroborating earlier observations using a smaller number of speakers [Ferré, 2010; Urbanik & Svennevig, 2021] or in which participants talked about specific objects [Donnellan et al., 2022]). Our corpus captured a wide range of gesturally depicted meanings and communicative intentions. Future analyses may shed light on how exactly specific communicative intentions may be associated with different or similar patterns of gesture-speech timing.

Moreover, previous results are mixed regarding the extent to which strokes precede their lexical affiliates during natural, face-to-face conversations. Our data contribute further data showing that strokes typically preceded speech, here in Dutch conversations. Our study is an important advance to the literature, given that most language use occurs in such face-to-face conversations, and that it is the context in which language evolved and is acquired (Holler & Levinson, 2019). Although there was variation in the gesture-speech asynchrony, overall the pattern that gesture onsets—and to a lesser extent, strokes—preceded their lexical affiliate was very stable across conversation topics (free conversations, discussions, holiday planning conversations). Moreover, we showed for the first time that this pattern is also stable across turn types (questions, nonquestion turns). Generalizability of gesture-speech timing data across people, communicative situations, and languages is very important if we want to draw conclusions about whether prediction is a possible mechanism underlying multimodal language processing.

Complementing the previous literature, our data show that it is possible that listeners use the gestural information to predict the lexical affiliate, demonstrating this effect here for casual, unprompted conversation. This possibility would be in line with the idea that multimodality facilitates language processing during face-to-face conversation (Drijvers & Holler, 2023). Multimodal language processing involves the challenge of having to integrate a plethora of information coming from different sources, displaced in time and space. Rather than slowing down language processing, the specific temporal organization of the different signals may be at the very heart of this facilitation.

4.2. Questions with gestures got faster responses

To test whether these gestures that preceded their lexical affiliates facilitate fast responding, we looked at response times during natural conversation. We provided additional evidence for the finding that questions with hand gestures got faster responses than questions without hand gestures (originally found in English three-party conversations; Holler et al., 2018). The current results extend this finding to dyadic conversations, and to Dutch, and show that the pattern also holds when considering only representational gestures that depict or refer to semantic information. Moreover, we show for the first time that responses are not simply faster because recipients use turn-initial placeholders (e.g., “uh”) to minimize the response time while taking extra time to plan the response.

A crucial question is whether the effect of gestures on response times is causal. No causal claims can be made on the basis of corpus data, and it is thus unclear whether the gestures really facilitate fast responding, or whether there might be an underlying confounding variable explaining the pattern. For example, questions with representational gestures may also contain more other visual signals, such as interactive gestures or facial signals, or they may contain more concrete words. To test for a causal effect of gestures on response speed, experimental studies could manipulate the presence and timing of gestures while keeping the speech of the question exactly the same across conditions. If the effect is indeed causal, future studies could investigate the mechanism(s) underlying this multimodal facilitation. It could, for example, be that gestures draw attention to what is being said, that there is additional semantic information in the gestures which facilitates message processing, or that—as a result of their timing—gestures facilitate prediction of upcoming semantic information (Holler et al., 2018).

4.3. Nonquestion turns with gestures did not get faster responses

Interestingly, we found no evidence that nonquestion turns with gestures got faster responses. A follow-up analysis showed that this was the case even when gestures appeared as close to turn end as they did in questions. However, the sample size for this follow-up analysis was small, and further research using larger samples should shed more light on how a gesture's timing within a turn impacts how much it speeds up responses.

If nonquestion turns with gestures truly do not get faster responses, this could perhaps be because compared to questions, there is less of a pressure to produce a next turn quickly, or even to produce one at all. Questions, on the other hand, come with the normative expectation that they are responded to, and the extent of the lag with which such responses occur carries pragmatic meaning (such as longer delays being associated with dispreferred responses, [Kendrick & Torreira, 2015; F. Roberts et al., 2006; Schegloff, 2007]). In line with these different normative expectations, in our data responses to nonquestion turns started less often with a turn-initial placeholder (e.g., "uh") than responses to questions. Perhaps due to this different pressure, gestures play a smaller role in reducing the lags between nonquestion turns and turns that follow them. Another possibility is that when participants responded quickly to long turns, they did so by only producing short backchannels (e.g., "yeah"). As we only coded full turns, such backchannels were excluded from our analysis. More fine-grained analyses of the turn structure, such as coding these backchannels or looking at turn-constructional units (Sacks et al., 1974) rather than full turns, could provide more insight into gesture's role during fast responding to nonquestion turns (Holler & Kendrick, 2015; Kendrick et al., 2023). Regardless of the specific reason, nonquestion turns with gestures did not get faster responses in our dataset, these results show a more refined picture of when gestures result in faster responses and when they do not. Our results highlight the need for more (qualitative) research into the circumstances under which gestures result in faster responses, focusing on aspects of the gestures as well as the interactions.

4.4. Predictive potential did not relate to response times

We hypothesized that when a gesture precedes its lexical affiliate more, the recipient may have more opportunity to use gestural information to predict the lexical affiliate and may thus respond faster. However, in our dataset, we found no evidence for this novel hypothesis. This could be because during conversation, many factors influence response times, including the sequential organization of turns, speech rate, syntactic complexity (S. G. Roberts et al., 2015), word order (Pekarek Doehler, 2021), and question format (Holler et al., 2018). Together, these and other factors may have masked a relation between gesture-speech asynchrony and response times in these corpus data. Therefore, controlled experiments are necessary to test whether gesture-speech asynchrony by itself can influence the time course of responding. Such experiments could also investigate when listeners start planning their response, instead of only measuring when responses are articulated, as was done in this corpus study. One possibility is that gestures with more predictive potential do in fact allow for earlier planning, but that the articulation of the response is held until the perception of turn-final cues (e.g., upcoming syntactic closure or gesture retraction) are detected (Barthel, Meyer, & Levinson, 2017; Holler et al., 2018; Levinson & Torreira, 2015). Experimental studies of the cognitive processes underpinning conversational turn-taking, and response planning in particular, are needed to shed more light on this issue.

4.5. Gesture and prediction

Although our analysis of predictive potential and response times did not show evidence for prediction, our gesture-speech timing data show that prediction is certainly a possible mechanism underlying multimodal language processing during face-to-face conversation and thus provide a crucial empirical foundation for controlled experiments looking into this mechanism. Indeed, some experimental studies already hint at the possibility of gestures possibly being used to predict upcoming speech input. In one study, participants shadowed speech from face-to-face conversations in an audiovisual context, an audiovisual context without visible speech, and an audio-only context (Drijvers & Holler, 2023). Participants would sometimes already utter words before they were heard, that is, they were predicting these words. Such prediction happened to a similar extent in all conditions, but participants predicted words earliest in the audiovisual context, later in the condition without visible speech, and latest in the audio-only context. These results suggest that bodily signals beyond visible speech play a role in facilitating linguistic prediction. However, participants perceived all visible signals speakers produced; the specific role of representational hand gestures was not tested in this study.

Another piece of evidence in line with the idea that gestures might be used for prediction of upcoming semantic information comes from Zhang, Frassinelli, Tuomainen, Skipper, and Vigliocco (2021). In their experiment, participants watched videos in which an actress uttered two-sentence passages, while their electroencephalography was recorded. When words were accompanied by representational gestures, the N400 event-related potential was less negative than when the same words were not accompanied by gesture. This effect on the N400 was especially strong for less predictable words. Because the N400 is sensitive to predictability (Dambacher, Kliegl, Hofmann, & Jacobs, 2006; Frank, Otten, Galli, & Vigliocco, 2015; Wlotko & Federmeier, 2013), these findings were taken to show that gestures affect the predictability of words. However, to investigate whether listeners use gestures for prediction, it is crucial that the gestural information precedes the corresponding information in speech (Fritz, Kita, Littlemore, & Krott, 2021), as is done in studies investigating whether lip movements facilitate language processing and prediction (van Wassenhove, Grant, & Poeppel, 2005; Venezia, Thurman, Matchin, George, & Hickok, 2016). One interesting study did present preceding gestures to listeners, and showed a decrease in auditory cortex activity for processing lexical affiliates when they were preceded by an iconic gesture compared to no gesture (Skipper, 2014). Still, from these results, one cannot distinguish prediction from integration (Mantegna et al., 2019; Nieuwland et al., 2020). One possibility is to measure prediction after the information that could be used for prediction is perceived, but before the target word is heard (Rommers, Dickson, Norton, Wlotko, & Federmeier, 2017; Terporten, Schoffelen, Dai, Hagoort, & Kösem, 2019; Wang, Hagoort, & Jensen, 2017). That way, it is possible to identify whether listeners indeed use the preceding gestural information to predict the to be conveyed lexical affiliate. Such paradigms would provide a more direct and more controlled test of prediction than looking at response times in conversation, as was done in the current study.

4.6. Limitations

The present study is limited in that it only focused on representational hand gestures, while during face-to-face conversation, many other communicative bodily signals are used (Holler & Levinson, 2019) which may also play a role in prediction and fast responding. These bodily signals include but are not limited to eye gaze, facial signals (Nota, Trujillo, & Holler, 2021), head gestures (Holler et al., 2018), and torso movements (Trujillo & Holler, 2021). Whether these signals, or combinations of signals, can be or are used to predict aspects of upcoming speech (including the social action, or turn end) is an open question. Moreover, while we only focused on the verbal aspect of responses, responses are also multimodal in nature. Future studies could, therefore, take into account the visual aspects of responses (e.g., nods) as well.

Moreover, the current study cannot speak to the question of why gestures are typically produced earlier than their lexical affiliates. Perhaps access to the gesture's motor representation is faster than lexical retrieval (Morrel-Samuels & Krauss, 1992), or gestures are planned before linguistic messages are formulated (de Ruiter, 2000). It could also be that speech and gesture are planned at the same time, but that speech production of the lexical affiliate is delayed for grammatical reasons (McNeill, 1985). Focusing on the listener, perhaps this gesture-speech timing is ideal for the listener's gesture-speech integration (Habets, Kita, Shao, Özyurek, & Hagoort, 2010; Obermeier & Gunter, 2015) or possibly speakers make themselves predictable for the recipient (Lelonkiewicz & Gambi, 2020; Urbanik & Svennevig, 2021; Vesper, van der Wel, Knoblich, & Sebanz, 2011). Further research is necessary to distinguish between these explanations.

Furthermore, the current study only tested dyads of acquainted participants, who had shared knowledge coming into the conversations. How their level of connection impacted the turn-taking and the role of gestures is an interesting question for future research, given that there is some evidence that friendship and connection may impact turn-taking timing (Coates, 1994; Templeton, Chang, Reynolds, Cone LeBeaumont, & Wheatley, 2022, 2023). Moreover, how the presence of shared knowledge may affect these processes could be studied by manipulating whether the knowledge participants talk about is shared or not. One interesting possibility is that if speakers produce gestures early to make themselves predictable for the recipient, they might do this even more for information that is not shared, to facilitate fast comprehension.

Another limitation of the present work is that it only focused on a single language, namely Dutch. However, gesture-speech timing could differ across languages: for example, a previous study on Chinese conversation found gestures to be less preceding of their corresponding speech (Chui, 2005) than what we found here. To identify possible cross-linguistic differences and ensure they are not the result of differences in coding, future studies investigating multiple languages with an identical approach are essential.

Finally, an interesting question concerns how the complementarity/redundancy of gesture to speech may impact gesture-speech timing, prediction, and fast responses. Fully complementary gestures depict meaning not present in speech (also called “obligatory gestures” [De Ruiter, Bangerter, & Dings, 2012] or “mixed syntax gestures” [Slama-Cazacu, 1976]), but gestures can also have underspecified lexical affiliates, such as “this” or “thing.” Exploratory analyses of our data suggest that when the lexical affiliate is underspecified, stroke onset may be closer to lexical affiliate onset (Tables S2 and S3). This way, the listener may have less time for prediction, reducing their chances of incorrectly predicting a specified lexical affiliate. Further research is necessary to replicate this pattern in larger datasets and test for any differences in a quantitative manner. Moreover, future studies could investigate how gesture complementarity relates to response times (for exploratory analyses on our data, see Tables S4 and S5).

4.7. Conclusion

To conclude, the early timing of naturally produced gestures during face-to-face conversation allows for the gestural information to be predictive of the upcoming lexical affiliate—in the majority of cases, the temporal organization of multimodal utterances equips representational hand gestures with “predictive potential.” This provides a crucial empirical basis for studies investigating whether, how, and under which circumstances listeners may use these gestures to make linguistic predictions. Moreover, we provided additional evidence for the finding that gestures are associated with fast responding during question−response interactions. These results are in line with the idea that multimodality may facilitate rapid turn-taking in face-to-face conversation, despite the challenge of having to integrate a plethora of information coming from different sources, displaced in time and space. The specific temporal organization of the different signals may be at the very heart of this facilitation.

Conflict of interest statement

The authors declare no conflict of interest.

References

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, *44*(2), 169–188. https://doi.org/10.1006/jmla.2000.2752

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science*, *27*(2), 195–225. https://doi.org/10.1207/s15516709cog2702_3

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barthel, M., & Levinson, S. C. (2020). Next speakers plan word forms in overlap with the incoming turn: Evidence from gaze-contingent switch task performance. *Language, Cognition and Neuroscience*, *35*(9), 1183–1202. https://doi.org/10.1080/23273798.2020.1716030

Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final “go-signals”. *Frontiers in Psychology*, *8*, 393. https://doi.org/10.3389/fpsyg.2017.00393

Barthel, M., Sauppe, S., Levinson, S. C., & Meyer, A. S. (2016). The timing of utterance planning in task-oriented dialogue: Evidence from a novel list-completion paradigm. *Frontiers in Psychology*, *7*, 1858. https://doi.org/10.3389/fpsyg.2016.01858

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, E., D’Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., Tzeng, A., & Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, *10*(2), 344–380. https://doi.org/10.3758/BF03196494

Bavelas, J. B. (2022). *Face-to-face dialogue: Theory, research, and applications*. Oxford University Press.

Bavelas, J. B., Chovil, N., Coates, L., & Roe, L. (1995). Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, *21*(4), 394–405. https://doi.org/10.1177/0146167295214010

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes*, *15*(4), 469–489. https://doi.org/10.1080/01638539209544823

Bavelas, J. B., Gerwing, J., & Healing, S. (2014). Effect of dialogue on demonstrations: Direct quotations, facial portrayals, hand gestures, and figurative references. *Discourse Processes*, *51*(8), 619–655. https://doi.org/10.1080/0163853X.2014.883730

Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *Proceedings of the 2nd Gesture and Speech in Interaction Conference*. https://pub.uni-bielefeld.de/record/2392953

Bernardis, P., & Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia*, *44*(2), 178–190. https://doi.org/10.1016/j.neuropsychologia.2005.05.007

Bernardis, P., Salillas, E., & Caramelli, N. (2008). Behavioural and neurophysiological evidence of semantic interaction between iconic gestures and words. *Cognitive Neuropsychology*, *25*(7/8), 1114–1128. https://doi.org/10.1080/02643290801921707

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia*, *109*, 295–310. https://doi.org/10.1016/j.neuropsychologia.2017.12.028

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, *5*(1), 12881. https://doi.org/10.1038/srep12881

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, *37*(6), 871–887. https://doi.org/10.1016/j.pragma.2004.10.016

Church, R. B., Kelly, S., & Holcombe, D. (2014). Temporal synchrony between speech, action and gesture during language production. *Language, Cognition and Neuroscience*, *29*(3), 345–354. https://doi.org/10.1080/01690965.2013.857783

Coates, J. (1994). No gaps, lots of overlap: Turn-taking patterns in the talk of women friends. In D. Graddol, J. Maybin, & B. Stierer (Eds.), *Researching language and literacy in social context: A reader* (pp. 177–192). Multilingual Matters Ltd.

Corps, R. E., Crossley, A., Gambi, C., & Pickering, M. J. (2018). Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition*, *175*, 77–95. https://doi.org/10.1016/j.cognition.2018.01.015

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103. https://doi.org/10.1016/j.brainres.2006.02.010

de Kok, I., Hough, J., Schlangen, D., & Kopp, S. (2016). Deictic gestures in coaching interactions. In *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human–Machine Interaction* (pp. 10–14). https://doi.org/10.1145/3011263.3011267

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284–311). Cambridge University Press. https://doi.org/10.1017/CBO9780511620850.018

De Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, *4*(2), 232–248. https://doi.org/10.1111/j.1756-8765.2012.01183.x

de Ruiter, J. P., Mitterer, Holger., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*(3), 515–535. https://doi.org/10.1353/lan.2006.0130

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Donnellan, E., Özder, L. E., Man, H., Grzyb, B., Gu, Y., & Vigliocco, G. (2022). Timing relationships between representational gestures and speech: A corpus based investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*. https://escholarship.org/uc/item/7w349725

Drijvers, L., & Holler, J. (2023). The multimodal facilitation effect in human communication. *Psychonomic Bulletin & Review*, *30*, 792–801. https://doi.org/10.3758/s13423-022-02178-x

Drijvers, L., Özyürek, A., & Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping*, *39*(5), 2075–2087. https://doi.org/10.1002/hbm.23987

Drijvers, L., Vaitonytė, J., & Özyürek, A. (2019). Degree of language experience modulates visual attention to visible speech and iconic gestures during clear and degraded speech comprehension. *Cognitive Science*, *43*(10), e12789. https://doi.org/10.1111/cogs.12789

Edelsky, C. (1981). Who's got the floor? *Language in Society*, *10*(3), 383–421.

Enfield, N. J. (2009). *The anatomy of meaning: Speech, gesture, and composite utterances*. Cambridge University Press.

Ferré, G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous French. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (pp. 86–91).

Ferré, G. (2019). Gesture/speech alignment in weather reports. In *Proceedings of the 6th Gesture and Speech in Interaction Conference*. https://doi.org/10.17619/UNIPB/1-805

Feyereisen, P. (1997). The competition between gesture and speech production in dual-task paradigms. *Journal of Memory and Language*, *36*(1), 13–33. https://doi.org/10.1006/jmla.1995.2458

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Fritz, I., Kita, S., Littlemore, J., & Krott, A. (2021). Multimodal language processing: How preceding discourse constrains gesture interpretation and affects gesture integration when gestures do not synchronise with semantic affiliates. *Journal of Memory and Language*, *117*, 104191. https://doi.org/10.1016/j.jml.2020.104191

Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology*, *6*, 751. https://doi.org/10.3389/fpsyg.2015.00751

Gisladottir, R. S., Chwilla, D. J., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PLoS ONE*, *10*(3), e0120068. https://doi.org/10.1371/journal.pone.0120068

Graziano, M., Nicoladis, E., & Marentette, P. (2020). How referential gestures align with speech: Evidence from monolingual and bilingual speakers. *Language Learning*, *70*(1), 266–304. https://doi.org/10.1111/lang.12376

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*(4), 274–279. https://doi.org/10.1111/1467-9280.00255

Habets, B., Kita, S., Shao, Z., Özyurek, A., & Hagoort, P. (2010). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*(8), 1845–1854. https://doi.org/10.1162/jocn.2010.21462

He, Y., Gebhardt, H., Steines, M., Sammer, G., Kircher, T., Nagels, A., & Straube, B. (2015). The EEG and fMRI signatures of neural integration: An investigation of meaningful gestures and corresponding speech. *Neuropsychologia*, *72*, 27–42. https://doi.org/10.1016/j.neuropsychologia.2015.04.018

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

Hoey, E. M. (2020). *When conversation lapses: The public accountability of silent copresence*. Oxford University Press.

Holle, H., Gunter, T. C., Rüschemeyer, S.-A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *Neuroimage*, *39*(4), 2010–2024. https://doi.org/10.1016/j.neuroimage.2007.10.055

Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, *47*(3), 837–847. https://doi.org/10.3758/s13428-014-0506-7

Holler, J., Bavelas, J., Woods, J., Geiger, M., & Simons, L. (2022). Given-new effects on the duration of gestures and of words in face-to-face dialogue. *Discourse Processes*, *59*(8), 619–645. https://doi.org/10.1080/0163853X.2022.2107859

Holler, J., & Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica*, *146*(1), 81–116. https://doi.org/10.1515/semi.2003.083

Holler, J., & Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction: Optimizing recipiency. *Frontiers in Psychology*, *6*, 98. https://doi.org/10.3389/fpsyg.2015.00098

Holler, J., Kendrick, K. H., & Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic Bulletin and Review*, *25*(5), 1900–1908. https://doi.org/10.3758/s13423-017-1363-z

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*, 255. https://doi.org/10.3389/fpsyg.2011.00255

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1–2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267. https://doi.org/10.1177/0956797609357327

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–227). De Gruyter.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press. https://doi.org/10.1017/CBO9780511807572

Kendrick, K. H., Holler, J., & Levinson, S. C. (2023). Turn-taking in human face-to-face interaction is multimodal: Gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1875), 20210473. https://doi.org/10.1098/rstb.2021.0473

Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes*, *52*(4), 255–289. https://doi.org/10.1080/0163853X.2014.955997

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and sign language in human–computer interaction* (Vol. 1371, pp. 23–35). Springer. https://doi.org/10.1007/BFb0052986

Krason, A., Fenton, R., Varley, R., & Vigliocco, G. (2021). The role of iconic gestures and mouth movements in face-to-face communication. *Psychonomic Bulletin & Review*, *29*, 600–612. https://doi.org/10.3758/s13423-021-02009-5

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lelonkiewicz, J. R., & Gambi, C. (2020). Making oneself predictable in linguistic interactions. *Acta Psychologica*, *209*, 103125. https://doi.org/10.1016/j.actpsy.2020.103125

Lerner, G. H. (1991). On the syntax of sentences-in-progress*. *Language in Society*, *20*(3), 441–458. https://doi.org/10.1017/S0047404500016572

Levelt, W. J. M., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language*, *24*(2), 133–164. https://doi.org/10.1016/0749-596X(85)90021-X

Levinson, S. C. (2016). Turn-taking in human communication – Origins and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*, 731. https://doi.org/10.3389/fpsyg.2015.00731

Liu, J., & Kavakli, M. (2011). Temporal relation between speech and co-verbal iconic gestures in multimodal interface design. In *Proceedings of the 2nd Gesture and Speech in Interaction Conference*.

Lubbers, M., & Torreira, F. (2018). *Pympi-ling: A Python module for processing ELANs EAF and Praats TextGrid annotation files* (1.69) [Computer software]. Retrieved from https://pypi.python.org/pypi/pympi-ling

Magyari, L., & de Ruiter, J. P. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, *3*, 376. https://doi.org/10.3389/fpsyg.2012.00376

Magyari, L., De Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in Psychology*, *8*, 211. https://doi.org/10.3389/fpsyg.2017.00211

Mantegna, F., Hintz, F., Ostarek, M., Alday, P. M., & Huettig, F. (2019). Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. *Neuropsychologia*, *134*, 107199. https://doi.org/10.1016/j.neuropsychologia.2019.107199

Maynard, S. K. (1997). Analyzing interactional management in native/non-native English conversation: A case of listener response. *International Review of Applied Linguistics in Language Teaching*, *35*(1), 37–75.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, *92*(3), 350–371. https://doi.org/10.1037/0033-295X.92.3.350

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

Meyer, A. S., Alday, P. M., Decuyper, C., & Knudsen, B. (2018). Working together: Contributions of corpus analyses and experimental psycholinguistics to understanding conversation. *Frontiers in Psychology*, *9*, 525. https://doi.org/10.3389/fpsyg.2018.00525

Miller, J. (1986). Timecourse of coactivation in bimodal divided attention. *Perception & Psychophysics*, *40*(5), 331–343. https://doi.org/10.3758/BF03203025

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex*, *14*(4), 452–465. https://doi.org/10.1093/cercor/bhh007

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*(3), 615–622. https://doi.org/10.1037/0278-7393.18.3.615

Nagels, A., Kircher, T., Steines, M., & Straube, B. (2015). Feeling addressed! The role of body orientation and co-speech gesture in social communication. *Human Brain Mapping*, *36*(5), 1925–1936. https://doi.org/10.1002/hbm.22746

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Husband, E. M., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., & Von Grebmer Zu Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., & Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, *11*(8), 1017. https://doi.org/10.3390/brainsci11081017

Obermeier, C., & Gunter, T. C. (2015). Multisensory integration: The case of a time window of gesture–speech integration. *Journal of Cognitive Neuroscience*, *27*(2), 292–307. https://doi.org/10.1162/jocn_a_00688

Obermeier, C., Holle, H., & Gunter, T. C. (2011). What iconic gesture fragments reveal about gesture–speech integration: When synchrony is lost, memory can help. *Journal of Cognitive Neuroscience*, *23*(7), 1648–1663. https://doi.org/10.1162/jocn.2010.21498

Pekarek Doehler, S. (2021). Word order affects response latency: Action projection and the timing of responses to question-word questions. *Discourse Processes*, *58*(4), 328–352. https://doi.org/10.1080/0163853X.2020.1824443

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software]. Retrieved from https://www.R-project.org/

RDocumentation. (2021). *Convergence: Assessing convergence for fitted models*. Retrieved from https://www.rdocumentation.org/packages/lme4/versions/1.1-27.1/topics/convergence

Roberts, F., Francis, A. L., & Morgan, M. (2006). The interaction of inter-turn silence with prosodic cues in listener perceptions of “trouble” in conversation. *Speech Communication*, *48*(9), 1079–1093. https://doi.org/10.1016/j.specom.2006.02.001

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology*, *6*, 509. https://doi.org/10.3389/fpsyg.2015.00509

Romei, V., Murray, M. M., Merabet, L. B., & Thut, G. (2007). Occipital transcranial magnetic stimulation has opposing effects on visual and auditory stimulus detection: Implications for multisensory interactions. *Journal of Neuroscience*, *27*(43), 11465–11472. https://doi.org/10.1523/JNEUROSCI.2827-07.2007

Rommers, J., Dickson, D. S., Norton, J. J. S., Wlotko, E. W., & Federmeier, K. D. (2017). Alpha and theta band dynamics related to sentential constraint and word expectancy. *Language, Cognition and Neuroscience*, *32*(5), 576–589. https://doi.org/10.1080/23273798.2016.1183799

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735. https://doi.org/10.2307/412243

Schegloff, E. A. (1984). On some gestures’ relation to talk. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 266–296). Cambridge University Press.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge University Press.

Senkowski, D., Molholm, S., Gomez-Ramirez, M., & Foxe, J. J. (2006). Oscillatory beta activity predicts response speed during a multisensory audiovisual reaction time task: A high-density electrical mapping study. *Cerebral Cortex*, *16*(11), 1556–1565. https://doi.org/10.1093/cercor/bhj091

Seyfeddinipur, M. (2006). *Disfluency: Interrupting speech and gesture* [Ph.D. dissertation, Radboud University.]. Retrieved from https://pure.mpg.de/rest/items/item_59337/component/file_2291735/content

Sjerps, M. J., Decuyper, C., & Meyer, A. S. (2020). Initiation of utterance planning in response to pre-recorded and “live” utterances. *Quarterly Journal of Experimental Psychology*, *73*(3), 357–374. https://doi.org/10.1177/1747021819881265

Sjerps, M. J., & Meyer, A. S. (2015). Variation in dual-task performance reveals late initiation of speech planning in turn-taking. *Cognition*, *136*, 304–324. https://doi.org/10.1016/j.cognition.2014.10.008

Skipper, J. I. (2014). Echoes of the spoken past: How auditory cortex hears context during speech perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130297. https://doi.org/10.1098/rstb.2013.0297

Slama-Cazacu, T. (1976). Nonverbal components in message sequence: “Mixed syntax.” In W. C. McCormack & S. A. Wurm (Eds.), *Language and man: Anthropological issues* (pp. 217–227). Mouton & Co.

Stivers, T., & Enfield, N. J. (2010). A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, *42*(10), 2620–2626. https://doi.org/10.1016/j.pragma.2010.04.002

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., Peter de Ruiter, J., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Strömbergsson, S., Hjalmarsson, A., Edlund, J., & House, D. (2013). Timing responses to questions in dialogue. In *Proceedings of INTERSPEECH* (pp. 2584–2588).

Suied, C., & Viaud-Delmon, I. (2009). Auditory-visual object recognition time suggests specific processing for animal sounds. *PLoS ONE*, *4*(4), e5256. https://doi.org/10.1371/journal.pone.0005256

Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., & Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, *119*(4), e2116915119. https://doi.org/10.1073/pnas.2116915119

Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., & Wheatley, T. (2023). Long gaps between turns are awkward for strangers but not for friends. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1875), 20210471. https://doi.org/10.1098/rstb.2021.0471

ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, *47*(1–2), 80–86. https://doi.org/10.1016/j.specom.2005.05.009

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. PsyArXiv. https://doi.org/10.31234/osf.io/b5zq7

Terporten, R., Schoffelen, J.-M., Dai, B., Hagoort, P., & Kösem, A. (2019). The relation between alpha/beta oscillations and the encoding of sentence induced contextual information. *Scientific Reports*, *9*(1), 20255. https://doi.org/10.1038/s41598-019-56600-x

Tolins, J., & Fox Tree, J. E. (2016). Overhearers use addressee backchannels in dialog comprehension. *Cognitive Science*, *40*(6), 1412–1434. https://doi.org/10.1111/cogs.12278

Torreira, F., Bögels, S., & Levinson, S. C. (2015). Breathing for answering: The time course of response planning in conversation. *Frontiers in Psychology*, *6*, 284. https://doi.org/10.3389/fpsyg.2015.00284

Trujillo, J. P., & Holler, J. (2021). The kinematics of social action: Visual signals provide cues for what interlocutors do in conversation. *Brain Sciences*, *11*(8), Article 8. https://doi.org/10.3390/brainsci11080996

Urbanik, P., & Svennevig, J. (2021). Action-depicting gestures and morphosyntax: The function of gesture-speech alignment in the conversational turn. *Frontiers in Psychology*, *12*, 689292. https://doi.org/10.3389/fpsyg.2021.689292

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Venezia, J. H., Thurman, S. M., Matchin, W., George, S. E., & Hickok, G. (2016). Timing in audiovisual speech perception: A mini review and new psychophysical data. *Attention, Perception, & Psychophysics*, *78*(2), 583–601. https://doi.org/10.3758/s13414-015-1026-y

Vesper, C., van der Wel, R. P. R. D., Knoblich, G., & Sebanz, N. (2011). Making oneself predictable: Reduced temporal variability facilitates joint action coordination. *Experimental Brain Research*, *211*(3–4), 517–530. https://doi.org/10.1007/s00221-011-2706-z

Wang, L., Hagoort, P., & Jensen, O. (2017). Language prediction is reflected by coupling between frontal gamma and posterior alpha oscillations. *Journal of Cognitive Neuroscience*, *30*(3), 432–447. https://doi.org/10.1162/jocn_a_01190

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (pp. 1556–1559).

Wlotko, E. W., & Federmeier, K. D. (2013). Two sides of meaning: The scalp-recorded N400 reflects distinct contributions from the cerebral hemispheres. *Frontiers in Psychology*, *4*, 181. https://doi.org/10.3389/fpsyg.2013.00181

Wu, Y. C., & Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, *42*(6), 654–667. https://doi.org/10.1111/j.1469-8986.2005.00356.x

Yngve, V. H. (1970). On getting a word in edgewise. *Papers from the Sixth Regional Meeting* (pp. 567–578). Chicago Linguistic Society.

Zellers, M., Gorisch, J., House, D., & Peters, B. (2019). Timing properties of hand gestures and their lexical counterparts at turn transition places. In *Proceedings from FONETIK*. https://doi.org/10.5281/zenodo.3246021

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*(1955), 20210500. https://doi.org/10.1098/rspb.2021.0500

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information