Article:

Holler, Judith, Kendrick, Kobin H. orcid.org/0000-0002-6656-1439 and Levinson, Stephen C. (2017) Processing language in face-to-face conversation: Questions with gestures get faster responses. Psychonomic Bulletin and Review. pp. 1-9. ISSN 1069-9384

Psychon Bull Rev
DOI 10.3758/s13423-017-1363-z

BRIEF REPORT

Processing language in face-to-face conversation: Questions with gestures get faster responses

Judith Holler[1,2] · Kobin H. Kendrick[1,3] · Stephen C. Levinson[1,4]

Abstract The home of human language use is face-to-face interaction, a context in which communicative exchanges are characterised not only by bodily signals accompanying what is being said but also by a pattern of alternating turns at talk. This transition between turns is astonishingly fast—typically a mere 200-ms elapse between a current and a next speaker's contribution—meaning that comprehending, producing, and coordinating conversational contributions in time is a significant challenge. This begs the question of whether the additional information carried by bodily signals facilitates or hinders language processing in this time-pressured environment. We present analyses of multimodal conversations revealing that bodily signals appear to profoundly influence language processing in interaction: Questions accompanied by gestures lead to shorter turn transition times—that is, to faster responses—than questions without gestures, and responses come earlier when gestures end before compared to after the question turn has ended. These findings hold even after taking into account prosodic patterns and other visual signals, such as gaze. The empirical findings presented here provide a first glimpse of the role of the body in the psycholinguistic processes underpinning human communication.

✉ Judith Holler
Judith.holler@mpi.nl

[1] Language & Cognition Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[2] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, The Netherlands

[3] Department of Language and Linguistic Science, University of York, York, UK

[4] Centre for Language Studies, Radboud University Nijmegen, Nijmegen, The Netherlands

Keywords Multimodal communication · Gesture · Language processing · Turn-taking

Introduction

The human language faculty sets us apart from other species. Its cognitive workings and social uses have intrigued scholars at least since the ancient Greeks. And yet, in many respects, we are still only beginning to discover how this unrivalled cognitive machinery functions and allows us to communicate with others. Conversation is one of the most fundamental human activities, yet the cognitive processes that underpin it are surprisingly poorly understood due to a long-standing focus on the processing of utterances in isolation.

Language processing and turn-taking in conversation

A striking feature of conversation is its temporal structure. The surface pattern is one of alternating bursts of vocalisation, with gaps between them averaging around just 200 ms (Stivers et al., 2009). This poses a psycholinguistic puzzle (Levinson, 2016): Judging by language production experiments, it takes a minimum of 600 ms to produce a simple one-word utterance (Indefrey & Levelt, 2004), implying that the turn-taking system shaping human conversation must be built on a considerable amount of parallel predictive processing. While listening to an ongoing turn, next speakers must predict its unfolding content and its end point to be able to begin planning their response early and to launch it on time. Experimental evidence suggests that processing spoken turns in conversation indeed rests quite considerably on prediction (e.g., Magyari & de Ruiter, 2012). In addition, quantitative analyses of conversational corpora have opened up a new domain of empirical enquiry, overcoming some of the well-

Published online: 08 September 2017

known limitations of traditional psycholinguistic paradigms by investigating human communication *in interactive situ*. Levinson and Torreira (2015) and Stivers et al. (2009) have provided quantitative confirmation of the turn-timing principle of minimal gaps and overlaps first observed by conversation analysts (Sacks, Schegloff, & Jefferson, 1974). Further, the fast timing of turns in conversation appears to be influenced not only by prediction but also by the ease with which turn content is cognitively processed (Roberts, Torreira, & Levinson, 2015) and the availability of turn-final 'go signals' (Levinson & Torreira, 2015).

Language and the body

The primary site of human language use is face-to-face conversation, suggesting that human language should be conceptualized as a fundamentally multimodal phenomenon (e.g., Bavelas & Chovil, 2000; Clark, 1996; Goldin-Meadow, 2003; Kendon, 2004, 2014; Levinson & Holler, 2014; Mondada, 2016; McNeill, 1992). Bodily signals, in particular, manual gestures, add a significant amount of meaning to what is being said. In certain contexts, the manual modality carries about 50% to 70% of the information constituting the overall message a speaker is encoding (Gerwing & Allison, 2009; Holler & Beattie, 2003; Holler & Wilkin, 2009). This information is taken up by recipients (e.g., Holler, Shovelton, & Beattie, 2009; S. D. Kelly, Barr, Church, & Lynch, 1999) and readily integrated with the information from the spoken channel (e.g., Kelly, Healey, Özyürek, & Holler, 2015; Kelly, Kravitz, & Hopkins, 2004; Willems, Özyürek, & Hagoort, 2007). Importantly, receiving gestural information in addition to speech appears to facilitate information processing in experimental settings as evidenced by faster reaction times to speech-plus-gesture compared to speech-only stimuli (Holle, Gunter, Rüschemeyer, Hennenlotter, & Iacoboni, 2008; Kelly, Özyürek, & Maris, 2010, Note 2; Nagels, Kircher, Steines, & Straube, 2015; Wu & Coulson, 2015).

Language processing situated in multimodal, face-to-face conversation

These diverse effects of gesture make it plausible that gestures also play a pivotal role in facilitating the remarkably fast system of turn-taking in conversation. Despite face-to-face conversation being the prime site for language use, there have been relatively few investigations in this domain. Duncan (1972) suggested a turn-end-signalling-model, in which the presence of bodily cues influences the likelihood of another speaker taking the turn. Further, conversation analysts have shown how gestures may give interlocutors cues to the prolongation, upcoming ending or beginning of a speaker's turn (Mondada, 2006; Schegloff, 1984; Streeck & Hartge, 1992). If so, we might expect gestures to affect the timing of turns in conversation, but to date this remains a gap in our understanding of human communicative behaviour.

Further, if gestures indeed influence the cognitive processing underlying turn-taking, then we should be able to show, first, that quite a substantial number of turns have a gestural component. Second, this gestural presence should have a direct effect on turn timing. As shown above, the gap between turns is shorter when turn content is easier to process. If the information provided by gesture indeed facilitates the processing of communicative messages, as suggested by experiments (Holle et al., 2008; Kelly et al., 2010; Nagels et al., 2015; Wu & Coulson, 2015), then next turns should follow faster, at least when they respond to preceding turns. Third, the timing of a specific aspect of gesture may also be of relevance, namely the temporal relationship between the termination of gesture and the termination of the spoken turn. In line with Duncan (1972), Levinson and Torreira (2015) argued that gestures may contribute to signalling turn ends, acting as an additional 'go signal'. Thus, if the gestural movement terminates prior to the spoken turn they accompany, this may affect the prediction of upcoming turn completion, resulting in faster turn transitions.

The present study

Here, we report quantitative analyses based on a rich corpus of multimodal conversation data (Holler & Kendrick, 2015), allowing us to move beyond traditional psycholinguistic methods based on studying comprehension and production independently and typically out of conversational and multimodal context. The study focuses on a type of communicative action that pervades conversation across languages (Stivers et al., 2009): question–response sequences. A question makes a response mandatory, making it an ideal prerequisite for comparing the speed with which next turns are issued. Further, the type of action a turn accomplishes has been shown to matter for the timing of turns (Roberts et al., 2015)—thus, mixing questions with other types of actions may blur the picture. As a first enquiry into whether bodily movements influence linguistic responses in conversation, the present analysis focuses on questions accompanied by two kinds of gestures that frequently accompany spontaneous speech: head and hand gestures. The results from this study inform models of human language use and the cognitive processes that underpin it.

Method

Corpus and participants

The present analyses are based on the Eye-Tracking in Multimodal Interaction Corpus (EMIC; Holler & Kendrick, 2015; see Fig. 1), which consists of 10 groups of acquainted native-English speakers engaged in dyadic and triadic casual conversations. For the analyses reported here, the conversations of seven triads were analysed. Of these, two were all male, two were all female, and four were mixed male–female groupings (age range: 19–68 years, $M = 30$).

Apparatus

In addition to eye-tracking glasses (SMI) filming the participants' behaviour from a first-person perspective, the interactions were filmed with three high-definition video cameras. Participants also wore a head-mounted directional microphone providing precise recordings of their vocal behaviour (for further details on laboratory setup and equipment, see Holler & Kendrick, 2015).

Procedure

Data collection

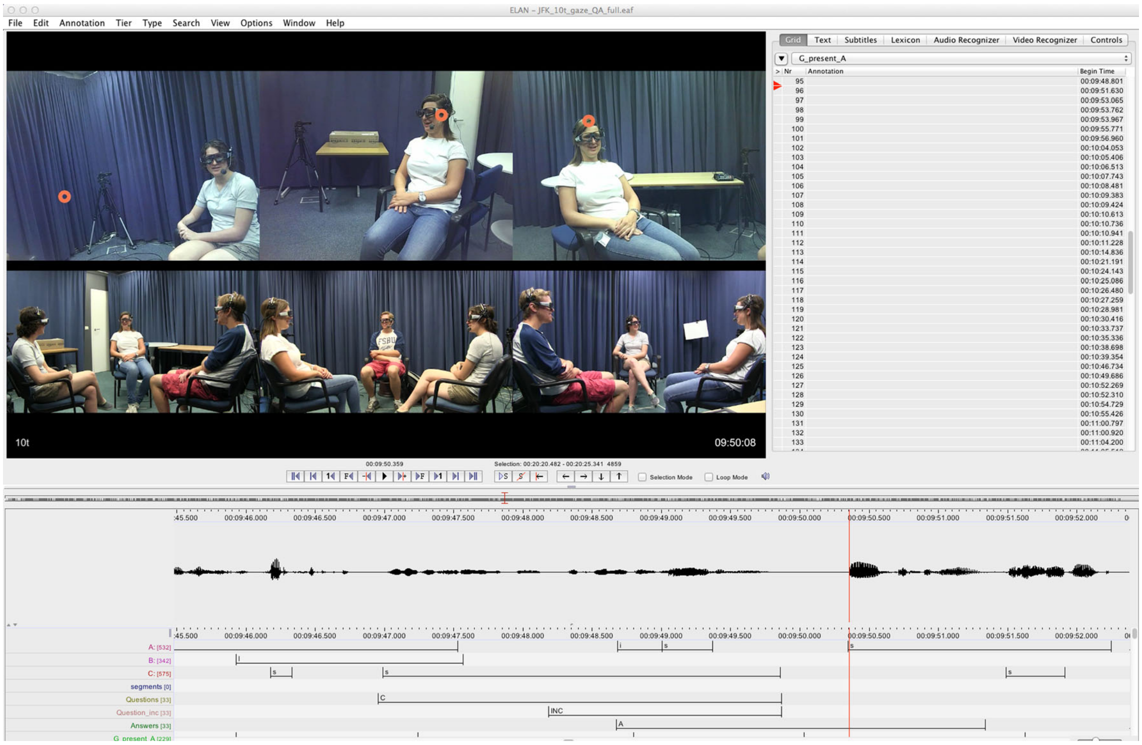
After the equipment was set up for recording, experimenters left the recording suite for 20 min. During this time participants conversed freely. The study was approved by the Social Sciences Faculty Ethics Committee, Radboud University Nijmegen.

Analytic focus

The present analysis focussed on question–response sequences, due to their prevalence in conversation, and because questions clearly make a specific next social action relevant (i.e., a response, typically an answer). This allows us to measure participants' speed of responding in comparable sequential environments, reducing the noise that may be induced by inclusion of a wide variety of different communicative actions. In this study only triadic conversations were used, as the presence of more participants opens up extra coordination problems—who is being addressed, and who will respond.

Question–response sequences Two hundred and eighty-one question–response sequences were identified (see Holler & Kendrick, 2015). The acoustic signal constituting the questions and responses was measured in Praat (Version 5.3.82; Boersma & Weenink, 2014).

Gestures For each question, the occurrence of gestures was annotated (using the software ELAN 4.61; Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006). Gestures were defined as communicative movements of head or hand (and torso) that speakers produced as part of conveying the question (or the response); this included (a) iconic and metaphoric gestures (McNeill, 1992); (b) deictic gestures



G_present_A

00:09:50.359

Selection: 00:20:20.482 - 00:20:25.341 4859

Nr	Annotation	Begin Time
95		00:09:48.801
96		00:09:51.630
97		00:09:53.065
98		00:09:53.762
99		00:09:53.967
100		00:09:55.771
101		00:09:56.960
102		00:10:04.053
103		00:10:05.406
104		00:10:06.513
105		00:10:07.743
106		00:10:08.481
107		00:10:09.383
108		00:10:09.424
109		00:10:10.613
110		00:10:10.736
111		00:10:10.941
112		00:10:11.228
113		00:10:14.836
114		00:10:21.191
115		00:10:24.143
116		00:10:25.086
117		00:10:26.480
118		00:10:27.259
119		00:10:28.981
120		00:10:30.416
121		00:10:33.737
122		00:10:35.336
123		00:10:38.698
124		00:10:39.354
125		00:10:46.734
126		00:10:49.686
127		00:10:52.269
128		00:10:52.310
129		00:10:54.729
130		00:10:55.426
131		00:11:00.797
132		00:11:00.920
133		00:11:04.200

Fig. 1 Still image showing an extract from the multisource synchronised audio-video recordings (and annotation software ELAN) serving as the basis for the present analyses.

(McNeill, 1992), and (c) pragmatic/interactive gestures (Bavelas, Chovil, Lawrie, & Wade, 1992; Bavelas, Chovil, Coates, & Roe, 1995; Kendon, 2004). An independent coder, blind to hypotheses, identified all gestures meeting the above criteria. Reliability was established for 22.8% of the question–response sequences (n = 64). This yielded a reliability of 76.7% for gesture identification indicating a high degree of agreement.

Measuring the timing of verbal and visual behaviours

Three timing measurements were established (using ELAN). The first focused on gestures only. Often, gestures consist of three phases: preparation, stroke, and retraction (Kita, van Gijn, & van der Hulst, 1998). The focus here was on the onset of the gesture retraction, which was determined through frame-by-frame inspection of the movement based on an established method (Seyfeddinipur, 2006). Identification of the retraction onset for head gestures is not as objectively possible, presumably due to the capital articulator being considerably more constrained than the hands in the size, range, complexity, and velocity of its movements. For this reason, we restricted our retraction annotations to manual gestures (however, our third timing measure described below—used for testing the effect of gesture on response speed—is based on all manual and all head gestures).

The second timing measure concerned the relation of gesture retraction and end of the verbal component of the question. We determined whether the gesture retraction began prior to or following the termination of the spoken question component (manual gestures only, and excluding those manual gestures without retractions due to another gesture following).

The third timing measure concerned the gap (in milliseconds) between the end of the spoken question and the onset of the spoken response. Positive values indicate silence between question and response, and negative values indicate overlap.

Measuring the associations between questions, gestures, prosody, and gaze

To be able to draw conclusions about the unique contribution of gestures to the timing of responses to questions we also coded the questions in our data for a range of prosodic variables as well as for questioners' gaze direction. With regard to prosodic patterns, questions were coded for F0 (Hz) (fundamental frequency), minimum pitch (Hz) (i.e., the lowest pitch level of a question), maximum pitch (Hz) (i.e. the pitch peak of a question), average intensity (dB) (i.e. amplitude), and maximum intensity (dB) (i.e. the amplitude peak of a question). Because previous research has suggested a link between gesture apex and pitch peak of the verbal utterance a gesture accompanies (e.g., Leonard & Cummins, 2011), and because gestural retractions tend to directly follow the point of maximum excursion (i.e., the apex) of a gesture, we were also interested in the relationship between gesture retraction onsets and pitch peaks. For this purpose, we also measured the time stamps of gesture retraction onset and pitch peaks. Finally, we also coded the questioner's gaze direction during the question, since previous research has argued for a role of gaze direction in foreshadowing upcoming turn completion (Duncan, 1972; Kendon, 1967). Gaze direction was categorized into 'always on responder', 'averts and returns to responder', 'averts but does not return to responder', or 'unclear' (for those cases where the eye-tracking data was not reliable [e.g., gaze fixations were missing for a number of frames during the question, or the calibration was off] or where the two independent coders did not agree). Note that the latter category included almost half of our data points (n = 144), meaning we should remain cautious regarding the interpretation of the gaze data in our sample.

Statistical analyses

We fitted linear mixed effects models to our data using lme4 (Version 1.1-12) package (Bates, Maechler, Bolker, & Walker, 2015) in R (Version 3.2.3; R Core Team, 2015). The main analyses are based on models with gap duration as the dependent variable; presence of gesture, or presence of gesture retraction, as a fixed effect; and questioner, respondent, and conversation as random intercepts. The analyses that check for the influence of prosodic patterns include F0, minimum pitch, maximum pitch, average intensity, or maximum intensity as additional fixed effects. The analyses that check for the influence of gaze include gaze direction (four levels, 'gaze always on responder' = reference level) as an additional fixed effect to the fixed effect of gesture. All model comparisons were made using the 'anova' function.

Table 1 Proportion of questions and responses accompanied by manual and/or head gesture(s) (left-hand column), one or more manual gestures (centre column), and one or more head gestures (right-hand column)

	Gesture (one or multiple)	Manual gesture (one or multiple)	Head gesture (one or multiple)
Questions	61% (N = 171)	30% (N = 83)	46% (N = 130)
Responses	67% (N = 189)	16% (N = 45)	60% (N = 168)

Note that the middle and right-hand columns are not additive due to utterances being accompanied by both hand and head gestures in several cases

Table 2 Use of gesture types by speakers of questions and responses

	Representational	Deictic	Pragmatic	Total
Questions	40	20	206	266
Responses	20	12	219	251

Results

Frequency of gestures with question–response sequences

Out of the 281 questions, the majority were accompanied by gesture (>60%), and responses to these questions were accompanied by gesture only marginally more frequently (see Table 1). A gestural contribution to questions and responses in more than 60% of cases is rather substantial for spontaneous conversation. Most gestures speakers produced (namely, 82%, *n* = 425) appeared to fulfil pragmatic functions (see Table 2). Of these, 16% (*n* = 136) were headshakes and nods when responses were given, and 33% (*n* = 68) when questions were being asked (interestingly, this latter group of headshakes and nods fulfil functions other than responding yes or no).

The effect of question-associated gestures on the timing of responses

One question we set out to test was whether responses to questions accompanied by gesture are faster than responses to questions without gesture. Our data suggest that this is indeed the case. Whereas gaps following questions without gestures (*N* = 110) were most frequently on the order of 200 ms those following questions with a gestural component (*N* = 171) were considerably shorter, most frequently around 0 ms (see Fig. 2). This relationship between gesture and gap duration was statistically significant (β = −299.82, *SE* = 67.21, *t* = −4.46, *p* < .0001).

We then tested whether this pattern held, based on several subsets of the data, to gain insight into the generalizability of the effect. First of all, the pattern held when the model was run on just hand gestures (β = −304.81, *SE* = 73.61, *t* = −4.14, *p* < .0001) and on just head gestures (β = −195.70, *SE* = 68.64, *t* = −2.85, *p* = .005).

Secondly, we generated a subset consisting of polar (= yes/no) questions only (i.e., polar questions [*n* = 228] as opposed to wh-questions [*n* = 48]; *n* = 5 classified as ‘other’). Responses to these two types of questions tend to differ in complexity, which can influence response times. For polar questions, those with gestures were still responded to faster than questions without gesture (β = −325.06, *SE* = 75.05, *t* = −4.33, *p* < .0001). The effect was not significant for the set of wh-questions, but bear in mind that these statistics were based on a small dataset.

Finally, we split our questions into two sets, one with increments (Schegloff, 2016)—that is, linguistic add-ons to syntactically and prosodically complete units (e.g., “How are you finding it [by the way]” [= increment]; *n* = 146)—and one without. In the case of increments nonfinal possible completion points may provide cues that elicit early responses, and this was indeed the case in our data (gaps for questions with increments: mode = −350 ms, median = −164 ms, mean = −154 ms; gaps for questions without increments: mode = 75 ms, median = 123 ms, mean = 163 ms). Because these responses occur fast in relation to the actual end of the question, we were interested whether the gesture effect would hold on the subset of questions without increments. It did indeed (β = −213.57, *SE* = 77.82, *t* = −2.75, *p* = .007).

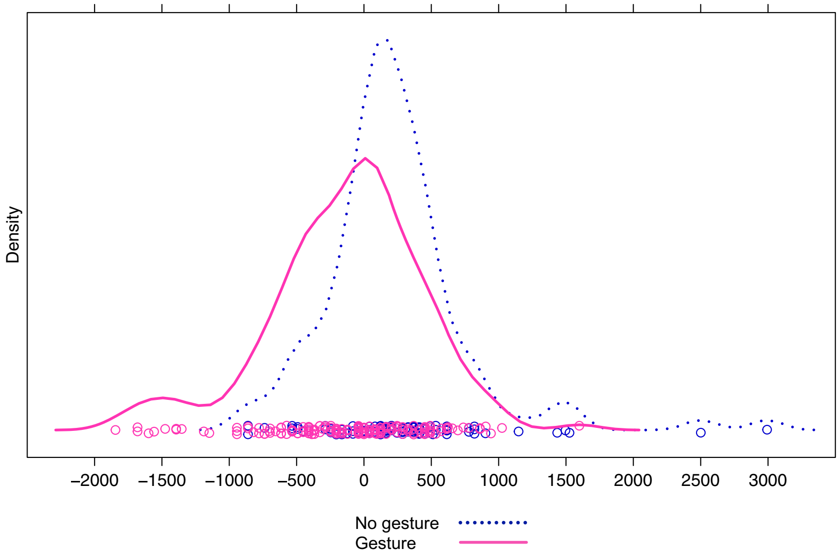


Fig. 2 Distribution of the duration of interturn gaps for questions with (*pink*) and without gestures (*dotted blue*), in milliseconds. Negative numbers indicate overlap; positive numbers indicate a gap between turns

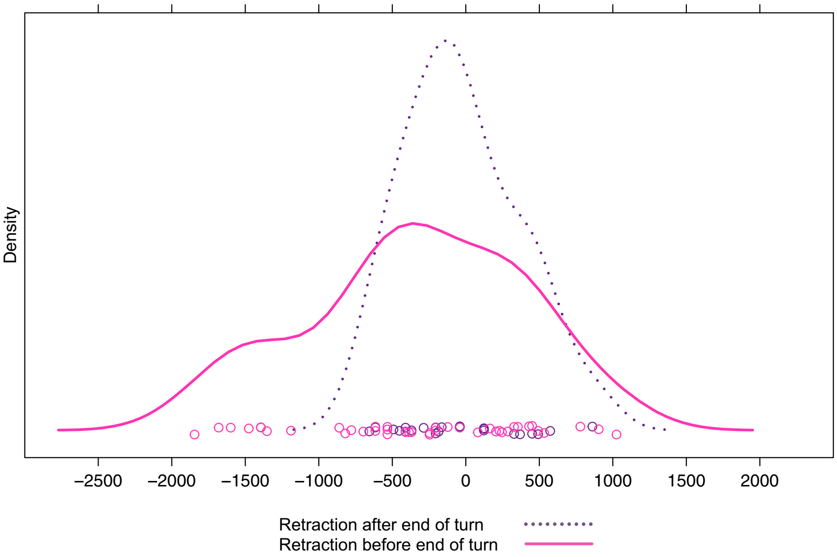


Fig. 3 Distribution of the duration of interturn gaps for questions with gestures retracting prior to the end of the spoken element of the question (*pink*) and with gestures retracting after the completion of the spoken element of the question (*dotted purple*), in milliseconds. Negative numbers indicate overlap; positive numbers indicate a gap between turns

In addition, we tested whether the gesture effect might in fact be due to other confounding factors; questions accompanied by gestures may be associated with specific prosodic patterns that are different to those for questions not accompanied by gestures. Our analyses show that neither fundamental frequency (F0) (β = −0.20, *SE* = 0.88, *t* = −0.23, *p* > .05), nor minimum pitch (β = 0.69, *SE* = 0.91, *t* = 0.76, *p* > .05), nor maximum pitch (β = −0.49, *SE* = 0.28, *t* = −1.78, *p* > .05) were significant predictors of response speed when they were added as fixed effects to the fixed effect of gesture; neither was average amplitude (β = 0.26, *SE* = 0.54, *t* = 0.48, *p* > .05). However, maximum amplitude did emerge as a significant predictor when added to the fixed effect of gesture (β = −20.17, *SE* = 5.87, *t* = −3.44, *p* < .0001), as well as when used as the only predictor (β = −24.29, *SE* = 5.96, *t* = −4.08, *p* < .0001). A model comparison showed that a model including both gesture and maximum amplitude provided a better fit for our data than a model including only the presence of gesture, $\chi^2(1) = 11.56$, *p* < .001, or only maximum amplitude, $\chi^2(1) = 14.76$, *p* < .0001 (when comparing the two models with just gesture or just maximum amplitude, the models were deemed to be equally good, but the gesture model had a lower Akaike information criterion [AIC] and Bayesian information criterion [BIC]). In the combined model, both the presence of gesture and maximum amplitude remained significant predictors, thus suggesting these variables lead to additive but independent effects and that gestures do make a unique contribution to gap duration.

Furthermore, we tested for the possible confounding influence of the questioner's gaze behaviour during the questions. We added gaze direction as a predictor to the predictor of gesture which resulted in a model where gesture remained a significant predictor (β = −302.82, *SE* = 66.09, *t* = −4.52, *p* < .0001). The effect of gaze was not significant for any of the levels of the gaze variable except for 'gaze never returns to responder' (β = −267.36, *SE* = 134.68, *t* = −1.99, *p* < .05). Fitting a model that includes gesture and gaze compared to one including only gesture did not result in a better model, $\chi^2(3) = 6.93$, *p* = .074. On the whole, there was not a large amount of variation in gaze direction—speakers were constantly looking at the responder in the majority (58%) of cases, and they did so equally often for questions with gesture (42 questions) and without (38 questions).

The effect of gesture retractions on the timing of responses

Of those questions that had a gestural component (*n* = 170), 70 questions (41%) were accompanied by a manual gesture with a retraction phase (see Method), and for 43 of these (61%), the retraction of the gesture began prior to the end of the question turn. Our analysis focused on whether gestures whose retractions begin prior to the end of the verbal utterance may function as early turn completion signals, thus affecting the timing of responses. This was indeed the case (β = −327.52, *SE* = 141.70, *t* = −2.31, *p* = .024; see Fig. 3). This pattern held when considering only questions without increments, which by themselves may provide early cues of possible termination (β = −414.2, *SE* = 196.96, *t* = −2.11, *p* = .0493; but note that our sample is reduced to just 21 data points for this latter analysis because not all manual gestures have retractions).

Also note that the effect reported above (responses being faster for questions with than without gesture) holds even when we remove all those gestures with a retraction. This means that the overall effect of responses being faster for questions with gestures is independent from the retraction effect.

Because it is known that gesture apex tends to be associated with pitch peaks, and because the gesture apex tends to directly precede the onset of gesture retraction (at least in the absence of gesture holds), we were intrigued to see whether we find similar associations in our data for the kind of gestures in our sample. Indeed, we did: The temporal association between pitch peak and retraction onset for our manual gestures was highly significant, $r = .99, p < .0001$. While this means that the presence of a pitch peak and a co-occurring gesture retraction onset may lead to a combined perceptual effect, the effect of gesture retraction on gap duration we reported above cannot be explained by the presence of the pitch peak instead, since questions without gestures also have pitch peaks, of course. See Table 3 for an overview of values of central tendency for the various datasets used in the main analyses reported in this section.

Discussion

In conversation, time is of the essence. Across languages, interactants take turns at speaking, precisely timed with gaps between turns in the ballpark of milliseconds. While it has been acknowledged that natural human language usage is multimodal in nature, the study here makes plausible that gestures are interdigitated with speech in language processing in interaction: First, the data show that most question turns in conversation have visual communicative components. This clear prevalence of bodily signals in the direct environment of speaking turns is the prerequisite for postulating a significant role of gesture in face-to-face language processing. Second, we showed that question turns are responded to faster when the question has a gestural component than when it does not. Gestures sped up turn transitions by around 200 ms if we consider modal response times, a significant difference in the context of conversational turn-taking, where response latencies are themselves typically just 200 ms long (Stivers et al., 2009). We have ruled out the potentially confounding influence of some obvious variables (question type, linguistic turn structure, prosodic patterns, gaze direction), none of which lead to significant effects overall, with the exception of maximum amplitude. This means that questions with gestures were often uttered with higher amplitude peaks than questions without gestures—presumably, this is due to intercostal muscles being moved when making gestures, thus affecting a greater expulsion of air. While this is an interesting finding in itself, it does not mitigate the role of gesture we have postulated here—the effect of gesture being present is independent of the effect of amplitude peak. However, it helps us to further refine the picture since the best model fit is one that includes both gesture presence and amplitude peak as predictors, suggesting that the effect of these two variables on gap duration may be strongest when they occur together. Furthermore, while gaze direction did not have a significant effect overall, it is worth mentioning that the effect approached significance, and that one of the individual levels of this variable had a significant effect—interestingly, this related to cases where questioners did not return their gaze to the responder after averting it, which were often questions in which more than one person was addressed, suggesting that

Table 3 Values of central tendency (in ms) for turn transitions measured in the present study

Data set		Mode[a]	Median	Mean
All questions	with gesture	200	164	229
	without gesture	0	−41	−130
Questions	with hand gesture	−500	−205	−237
	without hand gesture	50	123	115
Questions	with head gesture	−50	−41	−131
	without head gesture	100	123	132
Polar questions	with gesture	200	−123	−194
	without gesture	300	124	203
No-increment questions	with gesture	0	41	65
	without gesture	200	164	293
Questions	with early gesture retraction	−370	−369	−336
	with late gesture retraction	−80	−41	−45

Note. Negative numbers indicate overlap, positive values a gap, and a zero value a gapless transition

[a] All modes are based on Gaussian kernel density estimates.

competition for the floor may have elicited early responses in those cases. The previously postulated role (Duncan, 1972; Kendon, 1967) of the questioner's gaze being on the responder or gaze returning to the responder towards the end of the turn, however, did not seem to play a significant role in our data. Thus, while gesture makes a unique contribution to the effect of early responses in our data even when taking gaze direction into account, further studies into the role of gaze during turn-taking and its interplay with gesture seem warranted, especially in the context of dyadic (rather than triadic) interactions (i.e., the context in which a potential function of gaze for turn-taking was first observed).

Of course, corpus data always carry the possibility that uncontrolled factors may play a role, too. However, the fact that our finding is in line with experimental evidence that gestures facilitate language processing (Holle et al., 2008; Kelly et al., 2010; Nagels et al., 2015; Wu & Coulson, 2015) where such factors are strictly controlled for mitigates this issue. Our results suggest that such findings generalise to multimodal language use in interactive situ. Third, considering those question-accompanying gestures that terminated with a retraction phase (a subset, but still accounting for a quarter of all data points), next speakers responded significantly faster when the retraction began before compared to after turn end. We also found that retraction onset strongly correlates with the pitch peak of a question, which further adds to the literature on the link between prosody and gesture (e.g., Leonard & Cummins, 2011), but it does not change our interpretation of gesture causing the early retraction effect (since questions without gestures have pitch peaks, too, of course).

As to the exact mechanisms underlying these findings, we must remain speculative at this stage. The effect of responses being faster for questions with gestures than for those without (irrespective of whether or when the gestures retracted) may be explained in terms of a gesture-induced processing advantage—shorter response times in conversation tend to be taken as evidence for reduced comprehension processing in the face of production (Bögels, Magyari, & Levinson, 2015; Roberts et al., 2015). According to such an interpretation, the additional semantic and/or pragmatic information conveyed by hand and head gestures facilitates message processing (next to many other functions co-speech gestures can fulfil). It is also possible that gestures, which often precede corresponding information in speech (e.g., Bergmann, Aksu, & Kopp, 2011; Pine, Lufkin, Kirk, & Messer, 2007; Schegloff, 1984), facilitate the prediction of upcoming information, thus facilitating language processing. Of course, a third possibility is that bodily movements just draw more attention to what is being said, thus enhancing processing of the linguistic message itself, without any visual–verbal integration taking place. As to the second effect we found, addresses may perceive the onset of a visual retraction as an early cue of upcoming turn completion, thus acting as a 'go signal'. Experimental research will allow us to tease apart these different potential mechanisms.

The findings have a number of theoretical implications. Traditional language processing models tend to focus on verbal language alone, and on isolated utterances. Language produced in face-to-face conversation, the primordial site of language use, poses significant processing challenges that necessitate an interactional perspective. Here, we have shown that the multimodal environment of verbal language may crucially influence its processing in an interactive context. Our findings fit with gesture comprehension models (Kelly et al., 2010), at least if we assume that integration processes explain the gesture-induced facilitation observed here. Turn-taking models with a clear focus on the verbal modality (Sacks et al., 1974) should take note of the role gesture appears to play in turn coordination. The turn-taking model proposed by Duncan (1972) *is* multimodal in nature, but it reduces the role of gestural signals to a matter of presence or absence, whereas the current findings suggest that their temporal relationship with speech may play a crucial role. Future research may explore the role of gesture for turn-taking, and especially their temporal relation with speech, beyond question turns, with the aim to more fully refine existing turn-taking models.

Acknowledgements Open access funding provided by Max Planck Society. We would like to thank the Max Planck Society and the European Research Council for their financial support (INTERACT Grant #269484, awarded to S.C. Levinson), Nick Wood and Ludy Cilissen for their help with video processing, and Linda Drijvers and Marloes van der Groot for their assistance in annotating the corpus.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4.* Retrieved from the Institute for Statistics and Mathematics of WU website: http://CRAN.R-project.org/package=lme4

Bavelas, J. B., & Chovil, N. (2000). Visible acts of meaning: An integrated message model of language use in face-to-face dialogue. *Journal of Language and Social Psychology, 19,* 163–194.

Bavelas, J. B., Chovil, N., Coates, L., & Roe, L. (1995). Gestures specialized for dialogue. *Personality and Social Psychology Bulletin, 21,* 394–405.

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes, 15,* 469–489

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

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer (Version 5.3.82) [Computer program]. Retrieved from http://www.praat.org/

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports, 5,* 12881.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*, 283–292.

Gerwing, J., & Allison, M. (2009). The relationship between verbal and gestural contributions in conversation: A comparison of three methods. *Gesture*, *9*, 312–336.

Goldin-Meadow, S. (2003). *Hearing gesture*. Cambridge: Harvard 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*, 2010–2024.

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,* 81–116.

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

Holler, J., Shovelton, H., & Beattie, G. (2009). Do iconic gestures really contribute to the semantic information communicated in face-to-face interaction? *Journal of Nonverbal Behavior, 33,* 73–88.

Holler, J., & Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences the representation of semantic information in speech and gesture in a narrative task. *Language and Cognitive Processes, 24,* 267–289.

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92,* 101–144.

Kelly, S. D., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture and action during language comprehension. *Psychonomic Bulletin & Review, 22,* 517–523.

Kelly, S. D., Barr, D., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language, 40,* 577–592.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89,* 253–260.

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,* 260–267.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*, 22–63.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kendon, A. (2014). Semiotic diversity in utterance production and the concept of 'language'. *Philosophical Transactions of the Royal Society B*, *369*(1651)20130293. doi:10.1098/rstb.2013.0293

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Gesture and sign language in human-computer interaction. *Lecture Notes in Computer Science*, *1371*, 23–35.

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language & Cognitive Processes*, *26*, 1457–1471.

Levinson, S. C. (2016). Turn-taking in human communication, origins, and implications for language processing. *Trends in Cognitive Sciences, 20,* 6–14.

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B, 369 (1651)* 20130302*,* doi:10.1098/rstb.2013030

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology, 6,* 731.

Magyari, L., & de Ruiter, J. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology, 3, 376.*

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

Mondada, L. (2006). Participants' online analysis and multimodal practices: Projecting the end of the turn and the closing of the sequence. *Discourse Studies*, *8*, 117–129.

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics*, *20*, 336–366.

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,* 1925–1936.

Pine, K. J., Lufkin, N., Kirk, E., & Messer, D. (2007). A microgenetic analysis of the relationship between speech and gesture in children: Evidence for semantic and temporal asynchrony. *Language and Cognitive Processes*, *22*, 234–246.

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organisation on the timing of turn taking: A corpus study. *Frontiers in Psychology, 6,* 509.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50,* 696–735.

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

Schegloff, E. A. (2016). Increments. In J. D. Robinson (Ed.), *Accountability in social interaction* (pp. 239–263). Oxford: Oxford University Press.

Seyfeddinipur, M. (2006). *Disfluency: Interrupting speech and gesture* (Unpublished doctoral dissertation), Radboud University Nijmegen, Nijmegen.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 10587–10592.

Streeck, J., & Hartge, U. (1992). Previews: Gestures at the transition place. In P. Auer & P. A. di Luzio (Eds.), *The contextualization of language* (pp. 135–157). Amsterdam: Benjamins.

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex, 17,* 2322–2333.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006),* 1556–1559. http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:60436

Wu, Y. C., & Coulson, S. (2015). Iconic gestures facilitate discourse comprehension in individuals with superior immediate memory for body configurations. *Psychological Science*, *26,* 1717–1727.