Cognition 256 (2025) 106049

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Beat gestures and prosodic prominence interactively influence language comprehension

Ambra Ferrari [a,b,*,1], Peter Hagoort [a,b]

[a] Max Plank Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[b] Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, 6525 EN Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Beat gestures
Prosody
Language comprehension
Multimodal communication
Pragmatics
Metacognition

ABSTRACT

Face-to-face communication is not only about 'what' is said but also 'how' it is said, both in speech and bodily signals. Beat gestures are rhythmic hand movements that typically accompany prosodic prominence in conversation. Yet, it is still unclear how beat gestures influence language comprehension. On the one hand, beat gestures may share the same functional role of focus markers as prosodic prominence. Accordingly, they would drive attention towards the concurrent speech and highlight its content. On the other hand, beat gestures may trigger inferences of high speaker confidence, generate the expectation that the sentence content is correct and thereby elicit the commitment to the truth of the statement. This study directly disentangled the two hypotheses by evaluating additive and interactive effects of prosodic prominence and beat gestures on language comprehension. Participants watched videos of a speaker uttering sentences and judged whether each sentence was true or false. Sentences sometimes contained a world knowledge violation that may go unnoticed ('semantic illusion'). Combining beat gestures with prosodic prominence led to a higher degree of semantic illusion, making more world knowledge violations go unnoticed during language comprehension. These results challenge current theories proposing that beat gestures are visual focus markers. To the contrary, they suggest that beat gestures automatically trigger inferences of high speaker confidence and thereby elicit the commitment to the truth of the statement, in line with Grice's cooperative principle in conversation. More broadly, our findings also highlight the influence of metacognition on language comprehension in face-to-face communication.

1. Introduction

Natural face-to-face communication relies not only on speech but also on bodily signals such as manual gestures (Holler and Levinson, 2019; Kita and Emmorey, 2023; Özyürek, 2014; Vigliocco et al., 2014). Beat gestures are rhythmic hand movements that do not have a clear referential component (Dimitrova et al., 2016; McNeill, 1992). Compared to other types of co-speech gestures, beat gestures are by far the most frequent; yet, it is still an open question how beat gestures influence language comprehension (Dargue et al., 2019; Kita and Emmorey, 2023).

On the one hand, beat gestures may function as visual focus markers during sentence processing. Accordingly, the co-occurrence of a beat gesture with a target word embedded in discourse improves later word recall (Igualada et al., 2017; Llanes-Coromina et al., 2018). Furthermore, ERP evidence suggests that beat gestures, albeit not carrying any semantic information, may facilitate the semantic integration of words into sentence context by indexing word saliency (Biau and Soto-Faraco, 2013; Wang and Chu, 2013). Relative to no hand movement, beat gestures elicit a reduction of N400 amplitude (Wang and Chu, 2013), which reflects the difficulty of word integration into sentence context due to semantic surprisal (Wang et al., 2009, 2011). Similar reductions of N400 amplitude can be found when a word is focus-marked via linguistic devices such as question context (Wang et al., 2009), syntactic structure (Cowles et al., 2007) and prosodic prominence (Swerts et al., 2002). Together, it is then conceivable that beat gestures drive attention towards the concurrent speech and highlight its content. Thus, in line with the pragmatic synchrony rule (McNeill, 1992), speech prominence and beat gestures would have the same pragmatic function. However, at present no direct behavioural evidence corroborates this interpretation. A working hypothesis would be that, if a sentence violates world knowledge, beat gestures would highlight such anomaly and facilitate

* Corresponding author at: Neurobiology of Language Department, Max Plank Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands.
E-mail address: ambra.ferrari@mpi.nl (A. Ferrari).
[1] Present address: CIMeC, University of Trento, Corso Bettini 31, 38068 Rovereto, Italy.

https://doi.org/10.1016/j.cognition.2024.106049
Received 2 February 2024; Received in revised form 24 October 2024; Accepted 13 December 2024
Available online 24 December 2024

its recognition.

On the other hand, beat gestures may contribute to the pragmatic interpretation of discourse regarding the trustworthiness of the information being communicated. Specifically, beat gestures may express how confident the speaker is about what they are saying (Brennan and Williams, 1995). Accordingly, bodily signals such as gaze shifts, head nods and eyebrow movements influence how listeners grade the speaker's level of knowledge in question answering (Kuhlen et al., 2015; Roseano et al., 2016; Swerts and Krahmer, 2005). For example, speakers come across as more confident when keeping eye contact with the listener as opposed to when looking away (Swerts and Krahmer, 2005). Similarly, beat gestures may trigger inferences of high speaker confidence (Prieto et al., 2018). Consequently, they may drive the expectation that the sentence content is correct and thereby elicit the commitment to the truth of the statement. Then, an alternative working hypothesis would be that, if a sentence violates world knowledge, beat gestures would overshadow such anomaly and impair its recognition.

In the present study, we directly disentangled these two hypotheses by evaluating whether and how prosodic prominence, beat gestures and their combination influence language comprehension. Participants watched videos of a speaker producing sentences sometimes containing a slight violation of world knowledge, which may go unnoticed ('semantic illusion', Erickson and Mattson, 1981), and they judged whether each sentence was true or false. We independently manipulated the presence of prosodic prominence and beat gestures in correspondence to the word that dictated the presence of the semantic illusion. This allowed us to test for additive and interactive effects of prosodic prominence and beat gestures on language comprehension. In line with its role as focus marker, we expected prosodic prominence to highlight world knowledge violations and thereby decrease the semantic illusion, as previously shown (Kristensen et al., 2013; Wang et al., 2011). If beats shared this same functional role, they would additively decrease the semantic illusion. If they instead functioned as cues of high speaker confidence, they would counteract the prosodic effect and thereby increase the illusion. Importantly, we contrasted beat gestures with grooming gestures (e.g. scratching your own body) that were closely matched in terms of kinematics but were not perceived as meaningfully related to the speech they accompanied. If the communicative value of the hand gesture is critical, the effects found for beat gestures would not generalize to grooming gestures.

2. Materials and methods

2.1. Participants

The minimally required sample size was N = 90, based on a-priori power analysis in G*Power (Faul et al., 2009) with power of 0.8 to detect a small effect size of Cohen's d = 0.3 at alpha = 0.05. We recruited 120 volunteers either via Academic Prolific (prolific.com) or the recruitment database of the Max Plank Institute for Psycholinguistics (Nijmegen, Netherlands). Following a-priori criteria (see section 'Exclusion criteria'), 110 participants were included in the analysis and results (38 males; mean age 26, range 18–45 years), considering the notion that online studies may produce noisier results than laboratory studies. All volunteers were native speakers of Dutch, with (corrected-to-)normal vision and hearing, no language-related disorders and no history of neurological or psychiatric conditions. They provided written informed consent and received financial reimbursement for their participation. The study followed institutional guidelines of the local ethics committee (CMO region Arnhem-Nijmegen, Netherlands).

2.2. Stimuli

We used videos of a speaker producing a sentence in Dutch. In each video, the actor performed either a beat gesture, a grooming gesture or no gesture. Crucially, grooming gestures closely matched the beat gestures' kinematic trajectory, distance, duration and thus speed. To increase ecological validity, we used different exemplars of beat and grooming gestures (Fig. 1A). In a factorial design (see section 'Experimental design'), we introduced prosodic prominence or a manual gesture in correspondence to a specific word within each sentence. For prosody, we manipulated acoustic features that are known to influence perceived emphasis in Dutch (de Pijper and Sanderman, 1994; Kristensen et al., 2013; Streefkerk et al., 1999). For videos with a manual gesture, the movement apex was synchronised with the onset of the corresponding word, in line with the typical temporal alignment between speech and co-speech gestures (Dimitrova et al., 2016; Leonard and Cummins, 2011). Each video started and finished with a silent frame of the speaker holding hands in a still position (Fig. 1B). See Supplementary section 1.1 for more information.

2.3. Experimental design

We independently manipulated the semantic congruence of spoken sentences (to create the semantic illusion), the presence of prosodic prominence and the presence of a manual gesture. Thus, the task conformed to a 2 (Congruence: yes, no) × 2 (Prosodic prominence: present, absent) × 3 (Gesture: beat, grooming or no gesture) repeated measures factorial design (Fig. 1C).

Within each sentence, we manipulated a specific word (defined as the critical word, CW) along the 3 independent factors: Congruence (congruent: C+; incongruent: C−), Prosodic prominence (present: P+; absent: P−) and Gesture (beat: B; grooming: G; no gesture: N). Hence, there were 12 versions of each experimental sentence (Table 1). The congruence of the CW was manipulated so it either fitted the sentence context (C+) or violated general world knowledge (C−), as confirmed through a validation procedure (see Supplementary section 1.3). Moreover, we added C− filler sentences to vary the position of prosodic prominence and gestures relative to CW onset, thereby avoiding its easy identification. These fillers contained an incongruent CW, but prosodic prominence and gestures co-occurred with one of the other words in the sentence. To obtain the same number of congruent and incongruent sentences in each list, we also added C+ fillers. Here, prosodic prominence and gestures coincided with a word selected from one random point along the sentence. We created two separate filler lists (C+ and C− respectively), which were different from the experimental sentences and were presented to all participants. We created twelve separate experimental sentence lists, such that each participant listened to only one version of the same experimental sentence (Table 1). In a Latin square design, each list ([15 experimental sentences/version × 12 versions] + [15 fillers/version × 8 versions (only sentences with gestures)] = 300 sentences) was presented to the same number of participants (12 lists × 10 subjects/list).

2.4. Experimental procedure

Participants watched videos of a speaker producing a sentence (Fig. 1D) and reported as accurately and fast as possible by button press whether the sentence was true or false in a yes/no forced choice task. Trials were presented in pseudorandom order without consecutive repetitions of the same condition.

To assure attention to the screen and to the sentence content throughout the experiment, participants performed a semantic matching task (Dimitrova et al., 2016). On catch trials (10% of all trials counterbalanced across conditions), a single-word question (e.g. *'Bible?'*) was visually displayed in the centre of the screen after the video's conclusion. Participants reported by button press whether the word was semantically related to the preceding sentence in a yes/no forced choice task. Questions never referred to the CWs (thus, the violations) and required an equal number of yes/no responses. Participants responded with high accuracy (93% ± 1%, across-participants' mean ± SEM), confirming their vigilance during the experiment.

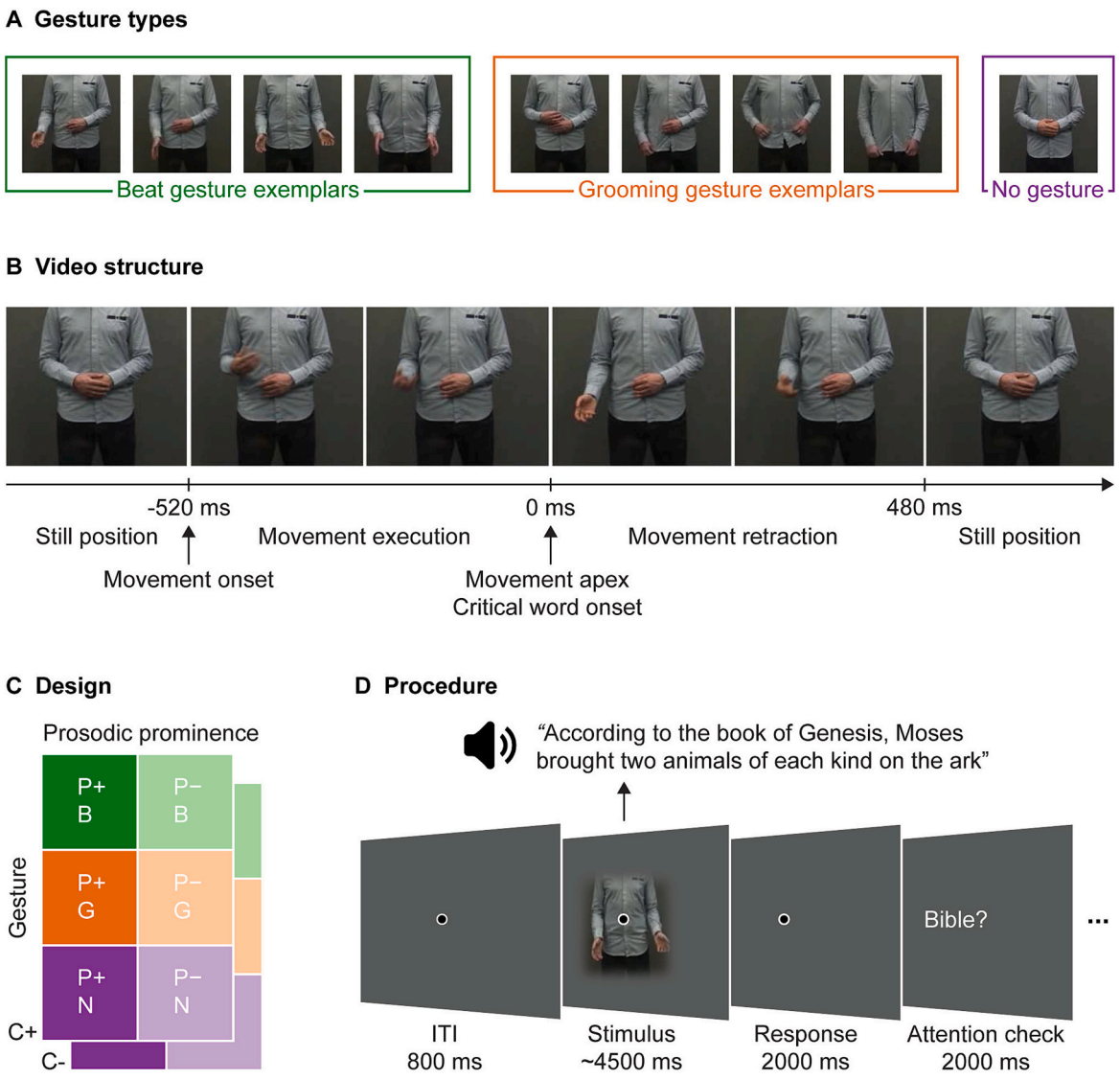

Fig. 1. Stimuli, experimental design and procedure. A) Participants watched videos of a speaker producing a sentence. The speaker performed either a beat gesture, a grooming gesture (i.e. adjusting their clothing) or no gesture, thus keeping in still position. To increase ecological validity, we employed four exemplars of beat and grooming gestures respectively (exemplars were randomly paired to sentences assuring a uniform distribution within conditions and participant). The still frames show the apex of each movement. B) Each video started and finished with a silence gap of 500 ms with the speaker in still position. For videos without a hand movement, the actor stayed in this position through the sentence. For videos containing a movement (beat or grooming), the apex was synchronised with the onset of the corresponding word (i.e. critical word, CW; 0 ms). Accordingly, all movements started 520 ms before the CW and finished 480 ms afterwards. C) We manipulated the CW along 3 independent factors: Congruence (congruent: C+; incongruent: C−), Prosodic prominence (present: P+; absent: P−) and Gesture (beat: B; grooming: G; no gesture: N). The congruence of the CW was manipulated so it either fitted the sentence context (C+) or violated general world knowledge (C−). D) Participants reported whether the sentence uttered by the speaker was true or false in a yes/no forced choice task (response window: 2000 ms; ITI: 800 ms). Sentences were presented in Dutch to native speakers (here is the English translation). To assure attention to the audiovisual stimuli, we introduced an oddball semantic matching task: on 10% of all trials, participants reported whether a word presented on screen was semantically related to the sentence they had just heard (response window: 2000 ms; ITI: 800 ms).

Participants completed 5 blocks, each lasting ~8 minutes ([3 experimental sentences/version × 12 versions] + [3 fillers/version × 8 versions]). Before the experiment, they underwent a practice session with an independent set of 8 sentences to familiarize themselves with the experimental procedure.

2.5. Experimental setup

The experiment was conducted online via Gorilla Experiment Builder (Anwyl-Irvine et al., 2020; gorilla.sc). To optimize the audio quality and minimize distractions, we screened participants for headphone use via a validated online procedure including dichotic pitch perception and binaural beat perception (Milne et al., 2021). Participants provided responses with their right hand using specific keyboard keys (counterbalanced across participants). Specific a-priori exclusion criteria were employed to control for proper task execution (see section 'Exclusion criteria').

2.6. Statistical analysis

All analyses were limited to trials without missed responses (i.e. no answer within 2000 ms), premature responses (i.e. RTs < 100 ms) or response outliers (i.e. $|\,RT\,| > 3$ SD above each participant's mean across conditions). Only few trials were discarded (7.0% ± 0.8%, across participants' mean ± SEM).

We considered three complementary dependent variables. First, we

Table 1
Exemplification of the 12 versions of one sentence. Critical words (CWs) are in bold; CWs with prosodic prominence are in CAPITALS; for CWs accompanied by a manual gesture, the apex of the movement was synchronised with CW onset (see Fig. 1A-B for gesture types and video structure). The C+ and C− sentences were identical except for the CWs. For an optimal semantic illusion, we constructed sentences such that the congruent and incongruent CWs were semantically related (*'Noah'* and *'Moses'* in the example). Sentences were presented in Dutch to native speakers; here is the English translation (see Supplementary section 3.2 for the complete sentence list).

Example sentence by condition
Congruent, Prosodic prominence, Beat
In the book of Genesis, **NOAH** brought two animals of each kind on the ark.
Congruent, Prosodic prominence, Grooming
In the book of Genesis, **NOAH** brought two animals of each kind on the ark.
Congruent, Prosodic prominence, No gesture
In the book of Genesis, **NOAH** brought two animals of each kind on the ark.
Congruent, No prosodic prominence, Beat
In the book of Genesis, **Noah** brought two animals of each kind on the ark.
Congruent, No prosodic prominence, Grooming
In the book of Genesis, **Noah** brought two animals of each kind on the ark.
Congruent, No prosodic prominence, No gesture
In the book of Genesis, **Noah** brought two animals of each kind on the ark.
Incongruent, Prosodic prominence, Beat
In the book of Genesis, **MOSES** brought two animals of each kind on the ark.
Incongruent, Prosodic prominence, Grooming
In the book of Genesis, **MOSES** brought two animals of each kind on the ark.
Incongruent, Prosodic prominence, No gesture
In the book of Genesis, **MOSES** brought two animals of each kind on the ark.
Incongruent, No prosodic prominence, Beat
In the book of Genesis, **Moses** brought two animals of each kind on the ark.
Incongruent, No prosodic prominence, Grooming
In the book of Genesis, **Moses** brought two animals of each kind on the ark.
Incongruent, No prosodic prominence, No gesture
In the book of Genesis, **Moses** brought two animals of each kind on the ark.

evaluated the probability of correct responses across semantically congruent (C+) and incongruent (C−) sentences, i.e. 'yes' responses to C+ sentences and 'no' responses C− sentences. Critically, this measure is agnostic about the nature of participants' errors. Hence, two subsequent analyses evaluated participants' responses for C+ and C− sentences respectively. We anticipated participants to be highly accurate for C+ sentences, which were preselected to be unproblematic and compatible with general world knowledge (see Supplementary section 1.3). Therefore, we expected errors to occur primarily for C− sentences and evaluated the probability of semantic illusion responses, i.e. 'yes' responses to C− sentences. This measure specifically targeted our phenomenon of interest: if beat gestures functioned as visual focus markers, they would decrease the semantic illusion; if they instead functioned as cues of high speaker confidence, they would increase the illusion. For completeness, we additionally analysed response times for correct trials (see Supplementary section 1.5).

Responses to individual trials were entered into Bayesian linear mixed models testing for main effects and interactions in our 2 (Prosodic prominence: present, P+; absent, P−) × 3 (Gesture: beat, B; grooming, G; no gesture, N) factorial design. Using reduced-rank coding, Prosodic prominence was represented by the contrast [present (P+) > absent (P−)]; Gesture was represented by the contrasts [no gesture (N) > beat (B)] and [grooming (G) > beat (B)]; the Prosodic prominence × Gesture interaction was represented by the contrasts [present (P+) > absent (P−)] > [no gesture (N) > beat (B)] and [present (P+) > absent (P−)] > [grooming (G) > beat (B)]. For any significant interactions, we ran post-hoc comparisons via follow-up models. All experimental factors of interest included the full random effects structure (both intercepts and slopes) for subjects as well as for items (i.e. sentences). Thus, we were able to generalize our conclusions across the populations from which our participants and our sentences were drawn. Models were constructed using the Python package Bambi v0.13.0 (Capretto et al., 2022; see Supplementary section 1.4 for details on model specification, estimation and evaluation). Results are primarily summarized using 94% Highest Density Intervals (HDI).

2.7. Exclusion criteria

We excluded participants from all analyses if they did not show the general ability to discriminate between semantically congruent (C+) and incongruent (C−) sentences, which we anticipated based on the materials preselection (see Supplementary section 1.3). For each participant, 'yes' responses to C+ sentences were labelled as hits, while 'yes' responses to C− sentences were labelled as false alarms. It follows that:

$$d' = Z(P_H) - Z(P_{FA})$$

with P_H = proportion of hits, P_{FA} = proportion of false alarms. A $d' \leq 0$ would indicate that a participant could not successfully discriminate between C+ and C− sentences. Based on this criterion, 10 participants were excluded from the analyses.

3. Results

3.1. Accuracy

We first evaluated the probability of correct responses across both congruent (C+) and incongruent (C−) sentences, i.e. 'yes' responses to C+ sentences and 'no' responses C− sentences. Accuracy was significantly higher when prosodic prominence was present versus absent; however, adding beat gestures to prominence decreased accuracy and effectively counteracted the prosodic effect (Fig. 2A). This translated into a significant Prosodic prominence × Gesture interaction (Fig. 2B): the posterior estimates of both interaction contrasts (contrast 1: [P+ > P−] > [N > B]; contrast 2: [P+ > P−] > [G > B]) were significantly above zero (contrast 1: mean = 0.18, 94% HDI = [0.00, 0.37]; contrast 2: mean = 0.14, 94% HDI = [0.00, 0.32]). Importantly, the Bayesian analysis allowed us to generalize our conclusions across both the population from which our participants were drawn and the population from which our sentences were drawn, with contrasts 1 and 2 respectively 96.70% and 93.21% likely to be above zero (Fig. S3A; Table 2). Follow-up models confirmed that accuracy was significantly higher when prosodic prominence was accompanied by a still position (P+ [N > B]: mean = 0.21, 94% HDI = [0.05, 0.38]) or, crucially, when it was accompanied by a grooming gesture (P+ [G > B]: mean = 0.16, 94% HDI = [0.03, 0.30]) rather than a beat gesture. Instead, accuracy did not differ between grooming and no gesture (P+ [N > G]: mean = 0.05, 94% HDI = [−0.09, 0.19]). Accordingly, the respective posterior probability distributions were above zero with 99.38% (P+ [N > B]), 98.98% (P+ [G > B]) and 74.50% (P+ [N > G]) probability (Fig. S3B; Table 2).

Critically, accuracy is agnostic about the nature of participants' errors. As anticipated, participants' responses were ~ 90% correct for C+ sentences across all conditions, without any significant effects of prosodic prominence or manual gesture (Fig. S5; Table S1). Hence, we expected the interactive effect of prominence and beat gestures to originate from the semantic illusion.

3.2. Semantic illusion

In line with accuracy, we found an interactive effect of prosodic prominence and beat gestures on the probability of semantic illusion responses, i.e. 'yes' responses to C− sentences. Illusion responses significantly decreased when prosodic prominence was present versus absent; however, adding beat gestures to prominence increased the illusion and effectively counteracted the prosodic effect (Fig. 3A). This translated into a significant Prosodic prominence × Gesture interaction (Fig. 3B): the posterior estimates of both interaction contrasts (contrast 1: [P+ > P−] > [N > B]; contrast 2: [P+ > P−] > [G > B]) were significantly below zero (contrast 1: mean = −0.33, 94% HDI = [−0.59, −0.08]; contrast 2: mean = −0.20, 94% HDI = [−0.44, 0.00]). The

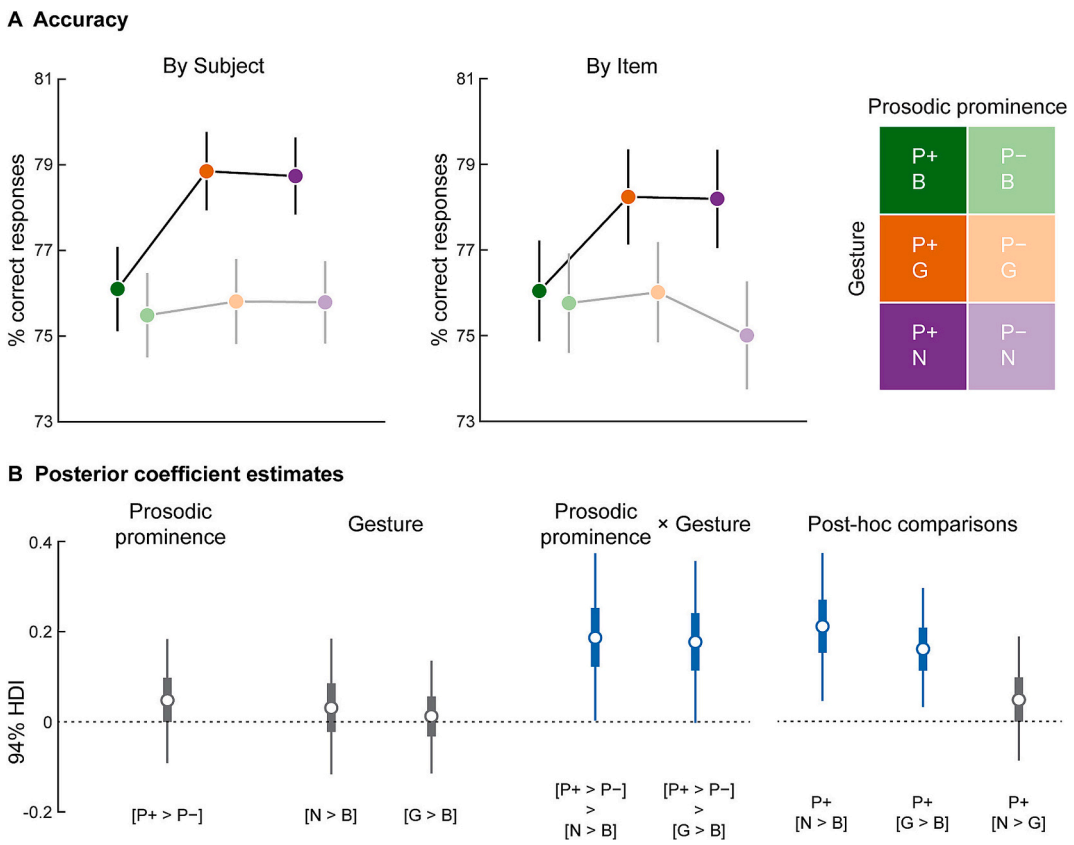

Fig. 2. Accuracy results. A) Across participants' (left) and across items' (right) mean correct responses ± SEM shown as a function of Prosodic prominence (present: P+; absent: P−) and Gesture (beat: B; grooming: G; no gesture: N). B) Posterior coefficient estimates of fixed effects of the model based on the GLM equation: $\text{logit}(C) \sim 1 + \text{Prosody*Gesture} + (1 + \text{Prosody*Gesture} \mid \text{Subject}) + (1 + \text{Prosody*Gesture} \mid \text{Item})$, where C = probability of correct response across C+ and C− sentences. White dots represent the estimated posterior means, thick lines indicate the central quartiles, thin lines represent the 94% Highest Density Intervals (HDI). Estimates indicate significant results when they do not overlap with zero (dashed horizontal line) as highlighted in blue. In essence, when prosodic prominence was present (versus absent) accuracy was significantly higher in the presence of a still position or, crucially, a grooming gesture rather than a beat gesture.

Table 2
Results of Bayesian linear mixed models for accuracy. Posterior coefficient estimates are summarized in terms of mean, standard deviation (SD), lower (3%) and upper (97%) bounds of 94% Highest Density Intervals (HDI). Estimates were obtained using Markov Chain Monte Carlo simulations with four chains, and chain convergence was quantified through the rank-normalized $\hat{r}$ (Vehtari et al., 2021): upon convergence, the between-chain and within-chain variances are identical ($\hat{r}$=1). Models' fit was assessed by the coefficient of determination R^2 for Bayesian linear regression models (Gelman et al., 2019). Labels represent experimental factors Prosodic prominence (present: P+; absent: P−) and Gesture (beat: B; grooming: G; no gesture: N).

Effect	Mean	SD	94% HDI	$\hat{r}$	R^2
Main model					0.36 ± 0.01
Intercept	1.30	0.08	(1.15, 1.45)	1.00	
[P+ > P−]	0.05	0.07	(−0.09, 0.18)	1.00	
[N > B]	0.03	0.08	(−0.12, 0.18)	1.00	
[G > B]	0.01	0.07	(−0.1, 0.13)	1.00	
[P+ > P−] > [N > B]	0.18	0.10	(0.00, 0.37)	1.00	
[P+ > P−] > [G > B]	0.14	0.09	(0.00, 0.32)	1.00	
Follow-up model 1					0.36 ± 0.01
Intercept	1.32	0.08	(1.17, 1.48)	1.00	
P+: [N > B]	0.21	0.09	(0.05, 0.38)	1.00	
Follow-up model 2					0.36 ± 0.01
Intercept	1.33	0.09	(1.17, 1.49)	1.00	
P+: [G > B]	0.16	0.07	(0.03, 0.30)	1.00	
Follow-up model 3					0.36 ± 0.01
Intercept	1.45	0.08	(1.30, 1.62)	1.00	
P+: [N > G]	0.05	0.07	(−0.09, 0.19)	1.00	

Bayesian analysis allowed us to generalize our conclusions across both the population from which our participants were drawn and the population from which our sentences were drawn, with contrasts 1 and 2 respectively 99.37% and 94.73% likely to be below zero (Fig. S4A; Table 3). Follow-up models confirmed that illusion responses significantly decreased when prosodic prominence was accompanied by a still position (P+ [N > B]: mean = −0.25, 94% HDI = [−0.46, −0.06]) or, crucially, when it was accompanied by a grooming gesture (P+ [G > B]: mean = −0.21, 94% HDI = [−0.39, −0.03]) rather than a beat gesture. Instead, illusion responses did not differ between grooming and no gesture (P+ [N > G]: mean = −0.03, 94% HDI = [−0.21, 0.16]). Accordingly, the respective posterior probability distributions were above zero with 99.21% (P+ [N > B]), 98.77% (P+ [G > B]) and 60.90% (P+ [N > G]) probability (Fig. S4B; Table 3).

In sum, prosodic prominence increased participants' comprehension accuracy (i.e. it decreased the semantic illusion). In line with previous research (Kristensen et al., 2013; Wang et al., 2011), this result suggests that prosodic prominence drives attention towards the speech content. Crucially, beat gestures counteracted the prosodic effect: combining beat gestures with prosodic prominence led to a higher degree of semantic illusion, making more world knowledge violations go unnoticed during language comprehension. Such effect did not generalize to grooming gestures, which were closely matched to beat gestures in terms of kinematics but were not perceived as meaningfully related to the concurrent speech. Notably, the present results generalized across all gesture exemplars used in the present study (see Supplementary section 1.7).

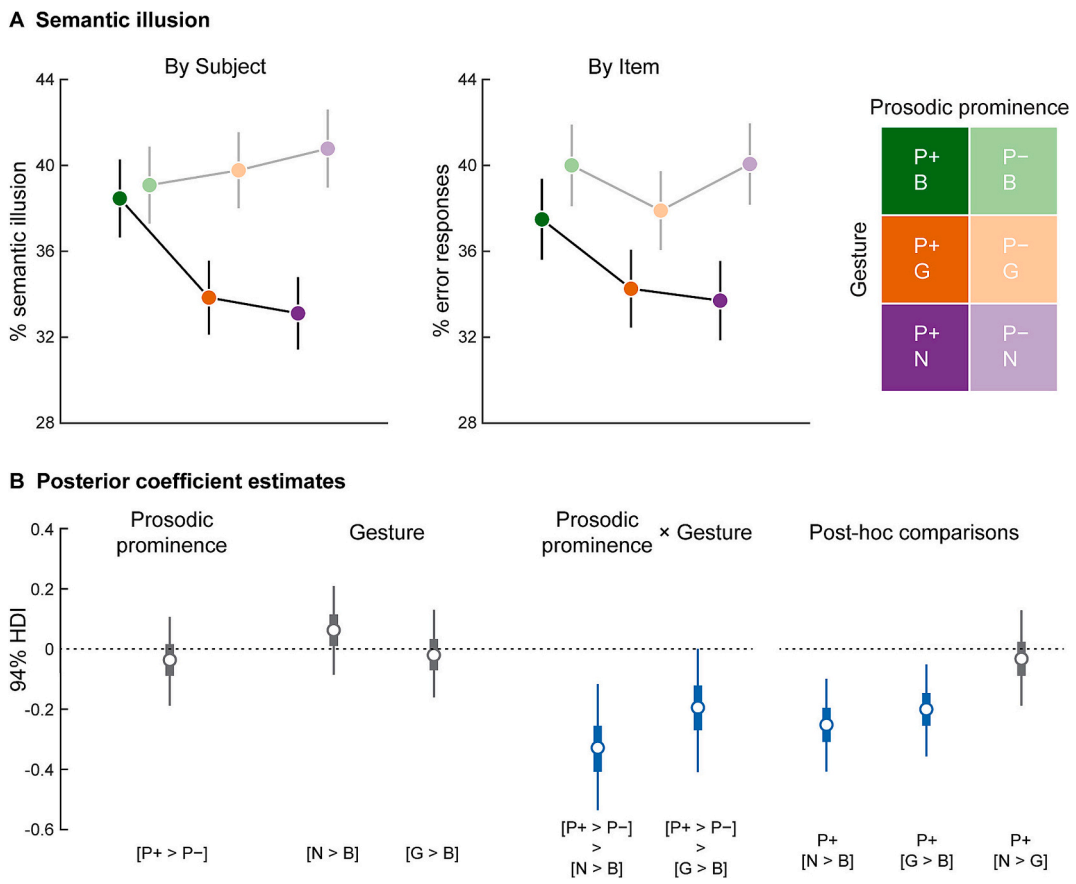

Fig. 3. Semantic illusion results. A) Across participants' (left) and across items' (right) mean semantic illusion responses ± SEM shown as a function of Prosodic prominence (present: P+; absent: P−) and Gesture (beat: B; grooming: G; no gesture: N). B) Posterior coefficient estimates of fixed effects of the model based on the GLM equation: $\text{logit}(S) \sim 1 +$ Prosody*Gesture $+ (1 +$ Prosody*Gesture $|$ Subject$) + (1 +$ Prosody*Gesture $|$ Item$)$, where S = probability of semantic illusion response in C− sentences. White dots represent the estimated posterior means, thick lines indicate the central quartiles, thin lines represent the 94% Highest Density Intervals (HDI). Estimates indicate significant results when they do not overlap with zero (dashed horizontal line) as highlighted in blue. In essence, when prosodic prominence was present (versus absent) the semantic illusion decreased in the presence of a still position or, crucially, a grooming gesture rather than a beat gesture.

Table 3
Results of Bayesian linear mixed models for semantic illusion. Posterior coefficient estimates are summarized in terms of mean, standard deviation (SD), lower (3%) and upper (97%) bounds of 94% Highest Density Intervals (HDI). Estimates were obtained using Markov Chain Monte Carlo simulations with four chains, and chain convergence was quantified through the rank-normalized $\hat{r}$ (Vehtari et al., 2021): upon convergence, the between-chain and within-chain variances are identical ($\hat{r}$=1). Models' fit was assessed by the coefficient of determination R^2 for Bayesian linear regression models (Gelman et al., 2019). Labels represent experimental factors Prosodic prominence (present: P+; absent: P−) and Gesture (beat: B; grooming: G; no gesture: N).

Effect	Mean	SD	94% HDI	$\hat{r}$	R^2
Main model					0.39 ± 0.01
Intercept	−0.59	0.11	(−0.81, −0.38)	1.00	
[P+ > P−]	−0.05	0.09	(−0.23, 0.12)	1.00	
[N > B]	0.07	0.09	(−0.10, 0.26)	1.00	
[G > B]	−0.02	0.09	(−0.19, 0.14)	1.00	
[P+ > P−] > [N > B]	−0.33	0.13	(−0.59, −0.08)	1.00	
[P+ > P−] > [G > B]	−0.20	0.12	(−0.44, 0.00)	1.00	
Follow-up model 1					0.39 ± 0.01
Intercept	−0.64	0.12	(−0.88, −0.43)	1.00	
P+: [N > B]	−0.25	0.11	(−0.46, −0.06)	1.00	
Follow-up model 2					0.38 ± 0.01
Intercept	−0.62	0.11	(−0.84, −0.41)	1.00	
P+: [G > B]	−0.21	0.09	(−0.39, −0.03)	1.00	
Follow-up model 3					0.38 ± 0.01
Intercept	−0.83	0.11	(−1.04, −0.61)	1.00	
P+: [N > G]	−0.03	0.10	(−0.21, 0.16)	1.00	

4. Discussion

Beat gestures are the most frequent and yet least investigated type of co-speech manual gestures. Consequently, their influence on language comprehension is at the centre of debate. On the one hand, beat gestures may share the same functional role of focus markers as prosodic prominence, following the pragmatic synchrony rule (McNeill, 1992). Accordingly, they would drive attention towards the concurrent speech and highlight its content, whether true or false. On the other hand, beat gestures may trigger inferences of high speaker confidence, possibly driving the expectation that the sentence content is correct. The present study directly disentangled these two hypotheses by evaluating additive and interactive effects of prosodic prominence and beat gestures on the semantic illusion (i.e. judging sentences with world knowledge violations as correct).

As previously shown (Kristensen et al., 2013; Wang et al., 2011), we found that prosodic prominence decreased the semantic illusion, namely it highlighted world knowledge violations. Thus, we corroborate the conclusion that, by controlling perceived emphasis (de Pijper and Sanderman, 1994; Streefkerk et al., 1999), prosodic prominence marks focus during sentence processing: under a "good-enough processing" framework (Ferreira et al., 2002), it orients our limited attentional resources to focused words in the sentence (Büring, 2007). Crucially, beat gestures counteracted the prosodic effect: when prominence was present, beat gestures led to a higher degree of semantic illusion. Hence, once prominence highlights the importance of a word, it is not necessarily processed in terms of accurate meaning. Instead, adding beat

gestures appears to drive the expectation that the word content is correct. The present results challenge current theories proposing that beat gestures necessarily share the same functional role of focus markers as prosodic prominence (McNeill, 1992). Instead, our findings indicate that beat gestures, combined with prominence, contribute to the pragmatic interpretation of discourse (Prieto et al., 2018). Specifically, they appear to steer the listener towards a judgement of trustworthiness of the information being communicated. According to Grice's cooperative principle in conversation (Grice, 1975), this may ultimately elicit the listener's commitment to the truth of the statement during language comprehension. While we propose that this effect is likely driven by inferences of high speaker confidence (Prieto et al., 2018; see also Supplementary section 1.6), it is important to note that beat gestures may not trigger such inferences directly but could do so through mediators such as perceived emotional security or engagement. Crucially, under either account, our results have clear theoretical implications: beat gestures can not only mark focus but also interact with prosodic prominence to function additionally as markers of speaker epistemic stance. Following recent literature (Lopez-Ozieblo, 2020; Shattuck-Hufnagel and Prieto, 2019), we propose that beat gestures can be multifunctional: they contextually interact with speech to co-determine the final perceived meaning.

In this study, we aligned all gestures to the word onset to maintain consistency across all conditions (Prosodic prominence: present/absent; Gesture: beat/grooming) and thereby reduce potential spurious differences (e.g. attentional differences) across conditions (Dimitrova et al., 2016). This was crucial because one of our hypotheses explored the attentional account of beat gestures as focus markers. However, this alignment choice may not fully reflect natural communication. In natural speech, the timing of gestures and speech varies significantly (Loehr, 2012; Rohrer et al., 2023), and beat gestures often align more closely with stressed syllables within a word (Leonard and Cummins, 2011; Wagner and Watson, 2010). Considering ecological validity, future research should examine whether different gesture-speech alignments affect communication.

Importantly, the pragmatic effect of beat gestures did not extend to grooming gestures, which were closely matched in terms of kinematics but were not perceived as meaningfully connected to the concurrent speech. As such, grooming gestures controlled for the presence of biological motion information, therefore ruling out the possibility that beat gestures simply distracted the listener because a salient visual event appeared in the conversational scene. This alternative interpretation is also incompatible with the specific type of error found for beat gestures: participants did not exhibit a decline in accuracy for all sentences, which may point to a general distraction effect; instead, they were systematically more likely to misjudge sentences carrying violations as being correct. Combined, these results strongly suggest that the communicative value of the hand movements is crucial in eliciting the listener's commitment to the truth of the statement.

More broadly, the present results provide original evidence on how metacognition influences language comprehension in face-to-face communication. Although confidence can be expressed at the lexical level (e.g. using epistemic adverbs such as "surely" or "perhaps"), previous research shows that bodily signals primarily disambiguate the speakers' awareness of their own possession of knowledge (Kuhlen et al., 2015; Roseano et al., 2016; Swerts and Krahmer, 2005). Beyond this, our findings suggest that listeners rapidly integrate bodily signals with speech to interpret utterances in light of speaker metacognitive knowledge and communicative goals. As such, we unveil a direct and compelling impact of expressed confidence on language comprehension: bodily communication has the power to shape our perception of truth.

Funding

This work was supported by the Max Plank Institute for Psycholinguistics.

CRediT authorship contribution statement

Ambra Ferrari: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Peter Hagoort:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization, Validation.

Declaration of competing interest

The authors declare no competing interests.

Data availability

All materials, data and code that are relevant to replicate the current findings are available on the Radboud Data Repository: https://doi.org/10.34973/6pyy-3x03.

Acknowledgements

The authors thank the Neurobiology of Language Department at the Max Plank Institute for Psycholinguistics and Donders Institute for Brain, Cognition and Behaviour for helpful discussions.

Appendix A. Supplementary materials and data

Supplementary materials and data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.106049.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language, 124*(2), 143–152. https://doi.org/10.1016/j.bandl.2012.10.008

Brennan, S. E., & Williams, M. (1995). The Feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language, 34*(3), 383–398. https://doi.org/10.1006/jmla.1995.1017

Büring, D. (2007). Semantics, intonation and information structure focus – background preliminaries on focus realization. In *The Oxford Handbook of Linguistic Interfaces*. Oxford University Press.

Capretto, T., Piho, C., Kumar, R., Westfall, J., Yarkoni, T., & Martin, O. A. (2022). Bambi: A simple interface for fitting bayesian linear models in Python. *Journal of Statistical Software, 103*(15). https://doi.org/10.18637/jss.v103.i15

Cowles, H. W., Kluender, R., Kutas, M., & Polinsky, M. (2007). Violations of information structure: An electrophysiological study of answers to wh-questions. *Brain and Language, 102*(3), 228–242. https://doi.org/10.1016/j.bandl.2007.04.004

Dargue, N., Sweller, N., & Jones, M. P. (2019). When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin, 145*(8), 765–784. https://doi.org/10.1037/bul0000202

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that Word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience, 28*(9), 1255–1269. https://doi.org/10.1162/jocn_a_00963

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior, 20*(5), 540–551. https://doi.org/10.1016/S0022-5371(81)90165-1

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11–15. https://doi.org/10.1111/1467-8721.00158

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *American Statistician, 73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Grice, H. P. (1975). Logic and Conversation. In *Speech Acts* (pp. 41–58). Brill. https://doi.org/10.1163/9789004368811_003.

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

Igualada, A., Esteve-Gibert, N., & Prieto, P. (2017). Beat gestures improve word recall in 3- to 5-year-old children. *Journal of Experimental Child Psychology, 156*, 99–112. https://doi.org/10.1016/j.jecp.2016.11.017

Kita, S., & Emmorey, K. (2023). Gesture links language and cognition for spoken and signed languages. *Nature Reviews Psychology, 2*(7), 407–420. https://doi.org/10.1038/s44159-023-00186-9

Kristensen, L. B., Wang, L., Petersson, K. M., & Hagoort, P. (2013). The interface between language and attention: Prosodic focus marking recruits a general attention network in spoken language comprehension. *Cerebral Cortex, 23*(8), 1836–1848. https://doi.org/10.1093/cercor/bhs164

Kuhlen, A. K., Bogler, C., Swerts, M., & Haynes, J.-D. (2015). Neural coding of assessing another person's knowledge based on nonverbal cues. *Social Cognitive and Affective Neuroscience, 10*(5), 729–734. https://doi.org/10.1093/scan/nsu111

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language & Cognitive Processes, 26*(10), 1457–1471. https://doi.org/10.1080/01690965.2010.500218

Llanes-Coromina, J., Vilà-Giménez, I., Kushch, O., Borràs-Comes, J., & Prieto, P. (2018). Beat gestures help preschoolers recall and comprehend discourse information. *Journal of Experimental Child Psychology, 172*, 168–188. https://doi.org/10.1016/j.jecp.2018.02.004

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology, 3*(1), 71–89. https://doi.org/10.1515/lp-2012-0006

Lopez-Ozieblo, R. (2020). Proposing a revised functional classification of pragmatic gestures. *Lingua, 247*, Article 102870. https://doi.org/10.1016/j.lingua.2020.102870

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

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods, 53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369*(1651). https://doi.org/10.1098/rstb.2013.0296

de Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *The Journal of the Acoustical Society of America, 96*(4), 2037–2047. https://doi.org/10.1121/1.410145

Prieto, P., Cravotta, A., Kushch, O., Rohrer, P. L., & Vilà-Giménez, I. (2018). Deconstructing beat gestures: A labelling proposal. In *Proceedings of the International Conference on Speech Prosody, 2018-June(June)* (pp. 201–205). https://doi.org/10.21437/SpeechProsody.2018-41

Rohrer, L. P., Delais-Roussarie, E., & Prieto, P. (2023). Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in English academic discourses. *Lingua, 293*, Article 103583. https://doi.org/10.1016/j.lingua.2023.103583

Roseano, P., González, M., Borràs-Comes, J., & Prieto, P. (2016). Communicating epistemic stance: How speech and gesture patterns reflect epistemicity and evidentiality. *Discourse Processes, 53*(3), 135–174. https://doi.org/10.1080/0163853X.2014.969137

Shattuck-Hufnagel, S., & Prieto, P. (2019). Dimensionalizing co-speech gestures. *Proceedings of the International Congress of Phonetic Sciences, 5*, 1490–1494.

Streefkerk, B. M., Pols, L. C. W., & ten Bosch, L. F. M. (1999). Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's. *EUROSPEECH, 99*, 551–554.

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language, 53*(1), 81–94. https://doi.org/10.1016/j.jml.2005.02.003

Swerts, M., Krahmer, E., & Avesani, C. (2002). Prosodic marking of information status in Dutch and Italian: A comparative analysis. *Journal of Phonetics, 30*(4), 629–654. https://doi.org/10.1006/jpho.2002.0178

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with Discussion). *Bayesian Analysis, 16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369*(1651), Article 20130292. https://doi.org/10.1098/rstb.2013.0292

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language & Cognitive Processes, 25*(7–9), 905–945. https://doi.org/10.1080/01690961003589492

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. *Neuropsychologia, 51*(13), 2847–2855. https://doi.org/10.1016/j.neuropsychologia.2013.09.027

Wang, L., Hagoort, P., & Yang, Y. (2009). Semantic illusion depends on information structure: ERP evidence. *Brain Research, 1282*, 50–56. https://doi.org/10.1016/j.brainres.2009.05.069

Wang, L., Bastiaansen, M., Yang, Y., & Hagoort, P. (2011). The influence of information structure on the depth of semantic processing: How focus and pitch accent determine the size of the N400 effect. *Neuropsychologia, 49*(5), 813–820. https://doi.org/10.1016/j.neuropsychologia.2010.12.035