

Speaker gaze and disfluencies during storytelling toward a robot or a human: The role of partner type and temperamental shyness

Nils F. Tolksdorf

Institute for German as a
Foreign Language,
SMARTcognition Lab,
Heidelberg University

tolksdorf@idf.uni-heidelberg.de

Adriana Hanulíková

Institute for German as a
Foreign Language,
SMARTcognition Lab,
Heidelberg University

hanulikova@idf.uni-heidelberg.de

Abstract

Storytelling in face-to-face dialogue involves coordinating multiple modalities, including speech and gaze. When the addressee is a social robot rather than a human, narrators' speech and gaze behavior may differ. However, few studies have examined both modalities within the same narratives, and it remains unclear how different aspects of multimodal behavior vary depending on whether the interaction partner is a social robot or a human. In this study, we examined how adults narrated picture-based stories to a social robot and a human interlocutor. Two modalities of narrative production were considered: pausing, separated into filled and unfilled pauses, and gaze behavior during the narration, directed either toward the story pictures or toward the listener. Critically, we examined whether shyness, considered as a continuous dimension of individual variation, relates to pausing and gaze behavior. Results revealed that speakers produced fewer filled pauses when narrating to the robot than to the human, while unfilled pauses did not differ. Additionally, higher shyness was associated with longer, though not more frequent, filled pauses in human-directed narration, although shyness did not reliably predict the magnitude of the partner effect. We discuss how our findings may contribute to a more nuanced understanding of multimodal interactions with embodied agents.

1 Introduction

Human communicative interaction in face-to-face settings is characterized as a multimodal process (Levinson, 2016). When telling a story, for example, speakers produce speech that includes disfluencies, such as filled and unfilled pauses, while also coordinating their gaze between the story pictures and their interlocutor. Importantly, these modalities are adjusted to the addressee (Holler and Levinson, 2019; Hanulíková et al., 2026). Speakers have been shown to tailor their behavior to a model of their

partner, and communicative behaviors directed to artificial partners such as social robots appear to form a distinct register with features shared with computer- and child-directed speech (Fischer et al., 2011; Jacka et al., 2025; Hanulíková et al., 2026). The characteristics of speech and gaze that a robot elicits may therefore reveal how speakers perceive their partner, which can in turn inform the design of systems that interpret and produce human multimodal behavior more contingently (Hessels and Fang, 2026). Although research comparing interactions with robotic and human partners has grown (e.g., Demir-Lira et al., 2020; Vollmer et al., 2018), less is known about speakers' multimodal behavior in communicative tasks, such as narrative settings towards a social robot compared to human-human interaction, as well as the direction of potential adjustments a person might take when interacting with a robotic partner. On the one hand, a robot may be perceived as less socially evaluative than a human partner (Tolksdorf et al., 2026) and as a machine that requires precise, unambiguous input. In this case, speakers may hesitate less, and looking at the robot's face may carry less social cost. On the other hand, interpersonal interaction is a co-constructed achievement in which partners monitor their addressees and rely on their feedback (Clark and Krych, 2004). A robotic interaction partner that is less familiar and provides fewer listener cues could lead to more disfluent production (Tolksdorf et al., 2021a) and potentially decrease the reciprocity of partner-directed gaze. Crucially, temperamental traits can decisively influence how a person coordinates social interactions. Yet, the influence of such individual differences (e.g., shyness) on multimodal behavior remains understudied. As a prevalent phenomenon and continuous temperamental dimension involving heightened attention to the social evaluative aspects of interaction (Sosa-Hernandez et al., 2023; Viertel and Tolksdorf, 2024), shyness may relate to how

speakers regulate pausing and gaze directed at their partner. While previous research on the influence of shyness on multimodal behavior in human-robot interaction has primarily focused on children (e.g., [Tolksdorf et al., 2021b](#); [Tykhonenko et al., 2026](#)), little research has examined this topic in adults. The present pilot study addresses these gaps by comparing filled and unfilled pausing and gaze allocation within the same narrations, when interacting with either a social robot or a human partner, and by relating both modalities to individual levels of shyness. The aims of this work are as follows:

- To investigate whether filled and unfilled pauses, as well as gaze behavior directed toward story pictures and interaction partners, vary when a picture-based story is narrated to a robot versus a human interaction partner.
- To examine whether individual differences in shyness are associated with variation in hesitation and gaze behavior.

2 Related Work

2.1 Speaking to a robot versus a human partner

A body of research has shown that speakers may adapt aspects of their speech to their addressees, depending on partner representations, communicative demands, and the interactional context ([Fischer, 2016](#); [Brennan and Hanna, 2009](#); [Ferreira, 2019](#)). However, although previous studies have documented linguistic and prosodic adaptations to different human addressees, comparatively little is known about differences in disfluency and gaze related to the partner during semi-naturalistic narrative production with embodied artificial partners, such as social robots. In particular, speech to artificial partners is reported to share features with other partner-adapted registers ([Fischer et al., 2011](#)). In addition, prior research suggests that backchannels and mutual monitoring in human-human interaction can support fluent production ([Clark and Krych, 2004](#)). Because robotic systems typically provide fewer such cues, narration to a robot may be comparatively more disfluent, consistent with earlier findings that children involved their caregiver more, and that the interactional flow was more often disrupted with a robot than with a human partner ([Serholt, 2018](#)). At the same time, speakers may treat a machine partner as requiring precise, unambiguous input ([Pang et al., 2025](#)), and

accordingly produce more controlled, less disfluent speech. Given that these findings may predict opposite directions, we leave the direction of partner effects open a priori.

2.2 Disfluencies as cognitive-communicative signals

Filled and unfilled pauses can serve multiple functions in interaction. [Clark and Fox Tree \(2002\)](#) argued that markers such as *uh* and *uhm* can act as conventionalized signals of an upcoming delay in turn-taking, and prior work converges on a dual role for these fillers in speech planning and listener-oriented signalling ([Corley and Stewart, 2008](#)). Unfilled pauses, by contrast, have been more closely linked to cognitive planning ([Zellner, 1994](#)). If speakers adapt their listener-directed behavior primarily to the type of interaction partner, then partner effects may be more likely for filled pauses than for unfilled ones. However, if reduced listener feedback leads to increased disruption in speech production, then both pause types may be affected. We therefore analyse filled and unfilled pauses separately and ask whether partner type and speaker characteristics act differently on each.

2.3 Gaze during storytelling

Speaker gaze in conversation has been described as serving regulatory, monitoring, and expressive functions ([Kendon, 1967](#)). In storytelling settings, narrators' gaze toward recipients has been associated with the elicitation of listener feedback ([Bavelas et al., 2002](#)), although more recent eye-tracking evidence suggests that this association is less uniform than initially proposed ([Zima, 2020](#)). Gaze toward interaction partners during hesitations can also invite collaboration, for instance when searching for a word ([Auer and Zima, 2026](#)), whereas gaze aversion has been observed during hesitation phases and before complex segments of talk ([Doherty-Sneddon and Phelps, 2005](#); [Auer and Zima, 2026](#)). In line with this notion, gaze directed at a partner primarily serves to elicit feedback. A robot offering fewer cues may attract fewer initiations of gaze from its partner. Additionally, looking at a human's face may carry more social-evaluative weight, whereas gazing at a robot's face may be less costly and thus occur more frequently. Importantly, most existing evidence stems from face-to-face conversations without task-related visual materials. In picture-elicited narration, speakers can be expected to direct much of their gaze at

the pictures, so that partner-directed gaze becomes a comparatively infrequent behavior whose distribution across different partner types is not well documented (Hessels and Fang, 2026).

2.4 Shyness and socially evaluative influences on production

Temperamental shyness is characterized by high sensitivity to novel situational factors, including social aspects such as the concern about being evaluated negatively (Viertel and Tolkstdorf, 2024). In this respect, social robots have been discussed as potential partners that can lessen the evaluative load within interaction (Peter and van Straten, 2024). The available evidence is, however, nuanced. In a dyadic peer paradigm investigating temporal coordination during interaction, conversational response times were related to children’s observed social engagement, but not their self-reported shyness (Wilson et al., 2024). In adults with social anxiety disorder, a clinical construct distinct from temperamental shyness, the proportion of hesitation disfluencies was similar to that of controls, though silent pausing was more prevalent (Silber-Varod et al., 2016). Evidence on the influences of shyness on multimodal behavior in human-robot interaction largely stems from children and modalities such as proxemics, expressiveness, and response latencies (e.g., Tolkstdorf et al., 2021b; Tykhonenko et al., 2026). Taken together, these findings can motivate the expectation that shyness, if it relates to production, may do so selectively — for pausing, possibly more in unfilled than in filled pauses. For gaze, partner-directed gaze is one of the more socially exposed behaviors in a narration task and may therefore be a plausible source of variation related to shyness. Because directly comparable evidence is limited, we examine this question without making a strong directional prediction. Since the social-evaluative demands of a robot and a human partner differ, shyness-related patterns may be partner-dependent.

2.5 The present study

This work is part of a broader project examining how speech production (e.g., lexical retrieval) in interaction with a robot compares with human–human interaction and depending on language background. Language background is a focus of the broader project, and is addressed here only exploratorily because of its potential relevance to hesitation behavior (Section 4.6). In the present

study, adults narrated picture-based stories to a social robot and to a human partner. We focus on examining whether the characteristics of the interaction partner (human versus robot) are associated with differences in hesitation behavior and gaze behavior during narration when overt listener feedback is minimized. Additionally, we investigate whether temperamental shyness is associated with disfluencies and partner-directed gaze and whether such associations appear selective for the type of social partner (robot versus human).

3 Method

3.1 Participants

24 adults (19 female, 5 male; mean age 26.98 years, $SD = 6.2$) participated in the pilot study. 13 were first-language (L1) speakers of German and 11 were speakers with German as a second language (L2). Due to technical errors with the robot, two participants completed only the human-directed condition and are excluded from within-participant partner contrasts.

3.2 Partner conditions

In the robot condition, participants narrated to the Navel social robot (Toussaint et al., 2023). The robot’s verbal behavior was generated via its built-in speech-to-text API and LLM integration, prompt-engineered to follow a standardized narrative-elicitation protocol (Gagarina et al., 2019): it invited the participant to tell the story, did not interrupt during the narration, and thanked the participant at the end. Both conditions were designed to be as comparable as possible, and during the narration, the robot remained in a stationary listening posture with its gaze directed ahead (see Figure 1A). The robot did not provide contingent verbal or nonverbal backchannel signals. In the human condition, participants narrated to a human experimenter whose listening behavior remained consistent as well. The experimenter did not initiate or return eye contact or produce backchannel signals and looked at the story pictures during the narration. Thus, both partners provided minimal, non-contingent listening behavior. This pattern was a deliberate design decision. Since the robot could not produce contingent listener responses, an unconstrained human listener might introduce differences in feedback, which would make it difficult to distinguish the effects of partner type from those of feedback. Therefore, holding listener responsive-

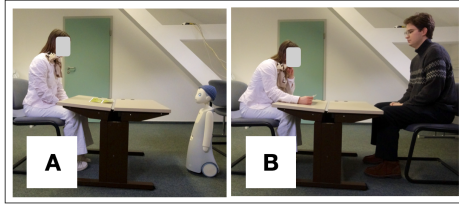


Figure 1: Experimental setup for robot- or human-directed storytelling interactions. (A) robot condition, (B) human condition.

ness constant at a minimal level may help to discern effects attributable to the partner itself rather than to the feedback it provides. However, this approach comes at the cost of naturalness. A listener who withholds backchannels and eye contact diverges from typical conversational storytelling, where listeners respond audibly and visibly (Bavelas et al., 2002; Zima, 2020).

3.3 Procedure and materials

Participants completed a single laboratory session. Narratives were elicited with the Multilingual Assessment Instrument for Narratives (MAIN; Gagarina et al., 2019), using parallel six-picture stories. Following the MAIN protocol, participants selected one of three envelopes, a procedure intended to create the impression that the partner did not already know the chosen story, and then narrated the story while viewing the picture sequence, which remained visible throughout. Narrations were video-recorded, and the duration of each narration was measured to calculate rates per minute. The order and assignment of stories were balanced across social partners, as was the order in which participants told stories to the robot or the human.

3.4 Disfluency and gaze coding

Disfluencies were annotated as filled pauses (e.g., *uh*, *uhm*) and unfilled (silent) pauses within the narration. Each pause was annotated with its onset and offset. From this annotation, we derived two measures per narration. First, as frequency measures, we computed filled and unfilled pauses per minute by dividing each participant’s pause counts by the corresponding narration duration. Second, we calculated the mean duration of pauses for each participant and partner condition. Gaze was annotated manually, frame by frame, from the video recordings in ELAN (Wittenburg et al., 2006). Gaze targets are coded as (a) the picture sequence and (b) the partner’s face region (the robot’s face re-

gion in the robot condition). From this annotation we derive, per narration, the number of partner-directed gaze events per minute. We report event rates because they indicate how often speakers look at each target, which most directly corresponds to the gaze-initiation behavior that motivated our research questions. In the present picture-elicited narration, speakers direct most of their gaze toward the stimulus material. Consequently, partner-directed looks occur as brief departures from the pictures. Therefore, their frequency provides insight into how often the speaker turns toward the partner, whereas duration-based measures are dominated by the picture-directed baseline. Additionally, rates rather than raw counts were used because narration durations varied across participants. All coding was performed by a single trained coder.

3.5 Assessment of shyness

Temperamental shyness was assessed using the “Shyness and Sociability Scales for Adults” (Asendorpf, 1997; Asendorpf and Wilpers, 1998). This standardized self-report questionnaire was administered at the end of the session and provided a continuous score that was used for analyses.

4 Preliminary Results

4.1 Analytic approach

Pause rates (filled and unfilled pauses per minute) and gaze-event rates per minute were analyzed using nonparametric, rank-based ANOVA-type statistics (ATS; Brunner et al., 2018; Noguchi et al., 2012), with partner (robot, human) as a within-participant factor. For each comparison, we report the ANOVA-type statistic (ATS), its associated F-approximation (for which the denominator degrees of freedom are set to infinity), and the relative treatment effects (RTEs) for each condition. An *RTE* above .50 indicates that observations from that condition tend to take higher ranks in the pooled sample. The same rank-based approach was applied to participants’ mean filled- and unfilled-pause durations in each condition. For the analysis of pause durations, duration values are defined only for participants who produced at least one pause of the relevant type in a given condition; participants without such pauses were excluded from the corresponding duration analysis (18 pairs for filled and 14 pairs for unfilled pauses). These analyses therefore concern mean pause duration among participants who produced at least one pause of the relevant type in

both conditions. Additionally, two further participants could not be included in the gaze coding, so gaze analyses are based on 20 complete pairs.

Associations with shyness were examined in two ways. First, we report Spearman rank correlations within each partner condition, describing how shyness relates to behavior within a partner condition. Additionally, to directly examine whether the magnitude of the partner difference varied with shyness, we calculated the within-participant difference between the human and robot conditions for each measure and correlated these difference scores with shyness. A significant association would indicate that the magnitude of the difference between partner conditions varied systematically with individual differences in shyness. Given the small sample size and the distributional characteristics of the behavioural measures, we used a rank-based analysis of within-participant difference scores as a parsimonious approach to examining whether the magnitude of partner differences varied with shyness. All shyness analyses are exploratory, and we did not apply correction for multiple testing, so individual p -values are not interpreted as confirmatory evidence. To examine whether partner effects differed by language background, we additionally included language background as a between-participant factor in the rank-based ATS model, with partner as a within-participant factor for filled and unfilled pauses.

4.2 Descriptive overview

Table 1: Descriptive overview per partner condition (medians with interquartile ranges). Medians are reported because the analyses are rank-based and the distributions are skewed.

	Human	Robot
Narration duration (min)	1.07 [0.92–1.57]	1.12 [0.85–1.37]
<i>Pauses per minute</i>		
Filled	8.26 [4.72–13.55]	3.15 [1.23–10.13]
Unfilled	1.97 [0.69–3.44]	1.74 [0.87–2.81]
<i>Mean pause duration (s)</i>		
Filled	0.79 [0.60–0.96]	0.70 [0.59–0.85]
Unfilled	1.13 [0.91–1.55]	1.05 [0.72–1.84]
<i>Gaze events per minute</i>		
Partner-directed	1.04 [0.29–2.18]	3.02 [0.29–6.38]
Picture-directed	1.13 [0.85–2.75]	3.44 [0.95–6.71]

4.3 Partner effects on pausing

Figure 2 shows pause rates in both conditions. Filled-pause rates were lower in robot-directed than in human-directed narration, and the ATS indi-

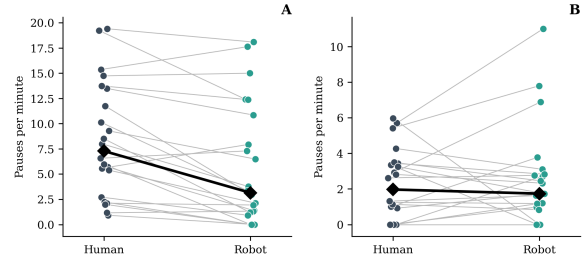


Figure 2: Filled (A) and unfilled (B) pause rates per participant and partner condition. Grey lines connect each participant’s two narrations; black line marks the condition medians.

cated a difference between conditions, $F(1, \infty) = 18.88$, $p < .001$, with relative treatment effects of $RTE = .58$ (human) and $.42$ (robot). Unfilled-pause rates did not differ between conditions, $F(1, \infty) = 0.04$, $p = .848$, $RTE = .51$ (human) and $.49$ (robot). Together, in this sample the partner effect appeared to be confined to filled pauses while the data provide no comparable indication for unfilled pauses. Pause durations showed no partner differences, for filled pauses, $F(1, \infty) = 0.93$, $p = .335$, $RTE = .53$ (human) vs. $.47$ (robot), or for unfilled pauses, $F(1, \infty) = 1.39$, $p = .239$, $RTE = .48$ (human) vs. $.52$ (robot).

4.4 Partner effects on gaze

To investigate partner-directed gaze-event rates per participant and condition, we ran another ATS that did not indicate a significant difference, $F(1, \infty) = 2.46$, $p = .117$, $RTE = .45$ (human) and $.55$ (robot), although median rates were descriptively higher toward the robot ($Mdn = 3.02$ vs. 1.04 events per minute). Picture-directed gaze events were more frequent in robot-directed than in human-directed narration; the ATS indicated a difference between conditions, $F(1, \infty) = 3.94$, $p = .047$, $RTE = .44$ (human) and $.56$ (robot). Descriptively, gaze-event rates were higher in the robot condition for both targets. However, only the picture-directed comparison yielded a nominally significant difference.

4.5 Role of shyness

Table 2 reports Spearman correlations between shyness and the disfluency and gaze measures within each partner condition, together with the correlation between shyness and the within-participant difference between conditions.

Within the human condition, exploratory analyses suggested a possible association between higher

Table 2: Spearman correlations between shyness and each measure, per partner condition and for the within-participant difference (Δ = Human – Robot). p -values are uncorrected for multiple testing and are reported as exploratory.

	Human		Robot		Δ	
	r_s	p	r_s	p	r_s	p
<i>Pauses per minute</i>						
Filled	+.08	.696	-.10	.670	+.26	.235
Unfilled	+.18	.399	+.11	.626	-.25	.269
<i>Mean pause duration (s)</i>						
Filled	+.41	.046	+.13	.597	+.25	.318
Unfilled	-.27	.285	-.14	.602	+.11	.702
<i>Gaze events per minute</i>						
Partner-directed	+.03	.914	-.30	.203	+.50	.026

shyness and longer mean filled-pause durations, $r_s = +.41$, $p = .046$, whereas the corresponding robot-condition coefficient was weaker, $r_s = +.13$, $p = .597$. As illustrated in Figure 3, the difference-score test, however, did not indicate that shyness predicts the magnitude of the partner effect on filled-pause duration, $r_s = +.25$, $p = .318$. The within-condition pattern therefore describes an association in human-directed narration. Because the difference-score test was not significant, the present data provide no evidence that shyness relates to pausing differently depending on partner type.

For partner-directed gaze, the difference score was nominally associated with shyness, $r_s = +.50$, $p = .026$. Because the difference score was defined as human – robot, and partner-directed gaze was descriptively more frequent toward the robot, this positive association indicates that the relative increase in partner-directed gaze in the robot condition was smaller among participants with higher shyness scores. However, a comparable coefficient was obtained for picture-directed gaze ($r_s = +.48$, $p = .032$), which is consistent with a general difference in gaze shifting rather than a partner-specific effect. We report it as an exploratory observation for future testing.

These associations should further be considered in light of the linguistic heterogeneity of the sample. Shyness scores were comparable between L1 and L2 speakers ($M = 14.69$, $SD = 4.01$ vs. $M = 12.18$, $SD = 3.46$; $p = .110$), and as a relatively stable temperamental characteristic that can manifest across linguistic contexts (Viertel and Tolkendorf, 2024), shyness may not be expected to relate to pausing in a manner specific to L1 or L2 speech. Given the small sample, language back-

ground nevertheless remains a potential source of heterogeneity that should be examined more systematically in future work.

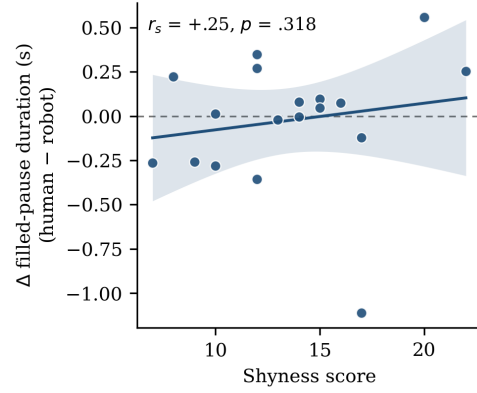


Figure 3: Association between shyness and the within-participant difference in mean filled-pause duration between conditions (human – robot). Positive values indicate longer filled pauses toward the human partner. The dashed line marks no partner difference; the shaded band is the 95% confidence interval of the linear fit.

4.6 Language background

For filled-pause rates, the ATS indicated a nominal L2 main effect, $F(1, 14.51) = 4.79$, $p = .046$, with L2 speakers at relatively higher filled-pause ranks overall, and an L2 $\times$ Partner interaction, $F(1, \infty) = 5.44$, $p = .020$: the reduction of filled pauses toward the robot was descriptively concentrated among L1 speakers (RTE human vs. robot: .50 vs. .29; median within-participant difference 3.27 pauses per minute), whereas the difference was smaller among L2 speakers ($RTE = .68$ vs. .61; median difference 1.37). For unfilled-pause rates, no effect involving language background approached significance. Interestingly, the reduction in filled pauses when speaking to the robot was evident in both language groups, and was descriptively more pronounced among L1 speakers. Exploratory analyses suggested that language background moderated the magnitude of partner differences in filled-pause rates, with larger human-robot differences observed among L1 speakers than L2 speakers. However, because subgroup sizes were small and no within-group tests were conducted, this pattern should be interpreted cautiously. Therefore, the main partner effect should not be assumed to be uniform across language backgrounds.

5 Discussion and Limitations

In this paper, we describe a pilot study that investigated the effects of different social partners (robot vs. human) and levels of shyness on multimodal narrative production with respect to disfluencies and gaze behavior. We found that robot-directed narrations included fewer filled pauses than human-directed narrations, whereas unfilled-pause rates, mean pause durations, and partner-directed gaze showed no significant differences. In the present interactional setting, both partners acted as minimally responsive listeners. Neither produced backchannels nor initiated or returned eye contact. Because overt listener feedback was minimised in both conditions, differences in overt contingent feedback are unlikely to be the primary explanation. However, the present design does not isolate partner expectations from other differences associated with interacting with a human versus a robot. For pausing, the observed pattern appears not to support the expectation that reduced listener support increases disfluency. Instead, fewer filled pauses toward the robot may converge with findings of lower disfluency rates in human–computer interaction and with evidence that speakers adapt to perceived artificial-partner capabilities (Branigan et al., 2011). Analysis of participants’ gazes toward the robot or human revealed no differences, although partner-directed gazes were descriptively more frequent toward the robot. Because neither interaction partner returned the participant’s gaze, mutual gaze was not established following partner-directed glances in either condition. We explain these findings with the possibility that glances toward the robot may partly reflect a visual inspection of a novel embodied partner rather than communicative signaling (Admoni and Scassellati, 2017; Hessels and Fang, 2026). This explanation is perhaps corroborated by our observations that gaze event rates also increased for the story pictures in the robot condition; thus, the pattern may reflect more frequent gaze shifting overall. However, the current event-based gaze measure cannot distinguish between more looking and more switching. A duration-based measure could differentiate between the two: the proportion of narration time spent on each target indicates the overall allocation of visual attention, while the mean time spent looking at each target indicates how long attention is sustained once directed there. A high event rate combined with short dwell times would indicate frequent redirection of gaze. In

contrast, a high event rate with comparable dwell times would indicate more looking overall. We regard this as an informative next step for the gaze data. Moreover, higher shyness was associated with longer, not more frequent, filled pauses in human-directed narration. However, a direct test of whether shyness predicts the magnitude of the partner effect did not reach significance, so the present data do not establish that this association is specific to the human partner. The observed pattern raises the possibility that variation related to shyness may be more apparent in the temporal characteristics of pausing than in the pause frequency itself. Whether this variation appears to be partner-dependent, potentially because participants experience the partners differently in social or evaluative demands (Tolksdorf et al., 2026), remains an open question for future research.

While our study has taken initial steps to explore the effects of different social partners and shyness on multimodal behavior, several limitations need to be recognized. The sample is small and was gender-imbalanced. In addition, the human partner was a male experimenter, thus evaluative and gaze dynamics may vary with the gender composition of a dyad. A further limitation is that all disfluency and gaze annotation was performed by a single trained coder, so no inter-rater reliability estimate is available. Double-coding a subset of the recordings and reporting agreement is necessary to establish the reliability of these measures, and we regard this as a requirement for the larger confirmatory study. Another limitation concerns the extent to which the human condition can generalize. Because the human listener was limited to minimal, non-contingent behavior, our results characterize how speakers may adjust their behavior based on the type of partner when listener responsiveness is held constant. These results should not be interpreted as a full representation of natural human-to-human storytelling, in which the listener’s responses are available and have been argued to encourage fluency (Clark and Krych, 2004). Although in the present design overt listener feedback was minimised to increase comparability across conditions, the absence of feedback may not have been perceived equivalently, and whether the observed partner effect would change with a responsive human listener cannot be determined from these data. Fluency-supporting feedback might reduce filled pauses toward the human listener, thereby attenuating the difference, whereas an unresponsive listener might have elevated hesita-

tions. A minimally responsive human may violate expectations for social interaction, whereas limited responsiveness may be more consistent with expectations for a social robot (Tolksdorf et al., 2026). In this case, the difference would be smaller under naturalistic conditions. Establishing this would require a responsive human baseline alongside the matched conditions reported here, which we consider a priority for future work. Moreover, each participant met the robot in a single session, so our results reflect first encounters; shyness- and novelty-related effects have been shown to change with familiarization (Tykhonenko et al., 2026).

6 Conclusions

In sum, this pilot study examined how adults' pausing and gaze behavior during picture-based storytelling varies with the type of social partner and with temperamental shyness. In this sample, narrations directed to the robot contained fewer filled pauses than those directed to the human partner, whereas unfilled pauses, pause durations, and partner-directed gaze showed no differences. Higher shyness was associated with longer, though not more frequent, filled pauses in human-directed narration, however, a test of the difference between conditions provided no evidence for a partner-dependent association. Additionally, in the present analysis, pause rates were normalized by narration duration, but not by more nuanced linguistic output. Two speakers may produce the same number of pauses per minute, yet differ substantially in speaking rate and amount of speech. Future analyses will therefore control for pauses by words or syllables and include speaking rate. Furthermore, future work will take into consideration gaze proportions and the temporal dynamics of gaze initiations, and how partner-directed glances relate to the timing of disfluencies.

References

Henny Admoni and Brian Scassellati. 2017. [Social eye gaze in human-robot interaction: A review](#). *Journal of Human-Robot Interaction*, 6(1):25–36.

Jens B. Asendorpf. 1997. *Schüchternheits- und Geselligkeitsskalen für Erwachsene*. Humboldt-Universität, Institut für Psychologie, Berlin.

Jens B. Asendorpf and S Wilpers. 1998. Personality effects on social relationships. *Journal of Personality and Social Psychology*, 74:1531–1544.

Peter Auer and Elisabeth Zima. 2026. [Does gaze aversion index dispreference or complexity of upcoming answers to questions?](#) *Frontiers in Communication*, 11:1731835.

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2002. [Listener Responses as a Collaborative Process: The Role of Gaze](#). *Journal of Communication*, 52(3):566–580.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. [The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers](#). *Cognition*, 121(1):41–57.

Susan E. Brennan and Joy E. Hanna. 2009. [Partner-Specific Adaptation in Dialog](#). *Topics in Cognitive Science*, 1(2):274–291.

Edgar Brunner, Arne C. Bathke, and Frank Konietzschke. 2018. [Rank and pseudo-rank procedures for independent observations in factorial designs: using r and sas](#). Springer Series in Statistics. Springer International Publishing, Cham.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Herbert H. Clark and Meredyth A. Krych. 2004. [Speaking while monitoring addressees for understanding](#). *Journal of Memory and Language*, 50(1):62–81.

Martin Corley and Oliver W. Stewart. 2008. [Hesitation Disfluencies in Spontaneous Speech: The Meaning of um](#). *Language and Linguistics Compass*, 2(4):589–602.

Ö. Ece Demir-Lira, Junko Kanero, Cansu Oranç, Sümeyye Koşku, Idil Franko, Tilbe Gökşun, and Aylin C. Küntay. 2020. [L2 vocabulary teaching by social robots: the role of gestures and on-screen cues as scaffolds](#). *Frontiers in Education*, 5:599636.

Gwyneth Doherty-Sneddon and Fiona G. Phelps. 2005. Gaze aversion: A response to cognitive or social difficulty? *Memory & cognition*, 33(4):727–733.

Victor S. Ferreira. 2019. [A Mechanistic Framework for Explaining Audience Design in Language Production](#). *Annual Review of Psychology*, 70(1):29–51.

Kerstin Fischer. 2016. [Designing Speech for a Recipient: The roles of partner modeling, alignment and feedback in so-called 'simplified registers'](#), volume 270 of *Pragmatics & Beyond New Series*. John Benjamins Publishing Company, Amsterdam.

Kerstin Fischer, Kilian Foth, Katharina J. Rohlfing, and Britta Wrede. 2011. [Mindful tutors: Linguistic choice and action demonstration in speech to infants and a simulated robot](#). *Interaction Studies*, 12(1):134–161.

- Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ute Bohnacker, and Joel Walters. 2019. [MAIN: multilingual assessment instrument for narratives – revised](#). *ZAS Papers in Linguistics*, 63:20.
- Adriana Hanulíková, Nils Frederik Tolksdorf, and Sarah Kapp. 2026. [Robot speech: how variability matters for child–robot interactions](#). *Frontiers in Robotics and AI*, 12:1725423.
- Roy S. Hessels and Yu Fang. 2026. [A visual perceptual perspective on gaze in social robotics](#). *Psychonomic Bulletin & Review*, 33(4):131.
- Judith Holler and Stephen C. Levinson. 2019. [Multimodal language processing in human communication](#). *Trends in Cognitive Sciences*, 23(8):639–652.
- Rhys Jacka, Paola Raquel Peña, Sophie J Leonard, Éva Szekély, and Benjamin R. Cowan. 2025. [Talking to...uh...um...machines: the impact of disfluent speech agents on partner models and perspective taking](#). In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–7, Waterloo ON Canada. ACM.
- Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.
- Stephen C. Levinson. 2016. [Turn-taking in human communication – origins and implications for language processing](#). *Trends in Cognitive Sciences*, 20(1):6–14.
- K Noguchi, Y. R. Gel, E Brunner, and F Konietzschke. 2012. nparLD: An R software package for the non-parametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software*, 50(12):1–23.
- Zi Haur Pang, Yahui Fu, Divesh Lala, Mikey Elmers, Koji Inoue, and Tatsuya Kawahara. 2025. [Does the appearance of autonomous conversational robots affect user spoken behaviors in real-world conference interactions?](#) In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, Yokohama Japan. ACM.
- Jochen Peter and Caroline L. Van van Straten. 2024. [Social robots and children: a field in development](#). In Leopoldina Fortunati and Autumn Edwards, editors, *The De Gruyter Handbook of Robots in Society and Culture*, pages 371–388. De Gruyter.
- Sofia Serholt. 2018. [Breakdowns in children’s interactions with a robotic tutor: A longitudinal study](#). *Computers in Human Behavior*, 81:250–264.
- Vered Silber-Varod, Hamutal Kreiner, Ronen Lovett, Yossi Levi-Belz, and Noam Amir. 2016. [Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals](#). In *Speech Prosody 2016*, pages 1211–1215. ISCA.
- Linda Sosa-Hernandez, Charlotte Aitken, Savannah Heintzman, Elizabeth S. Nilsen, and Heather A. Henderson. 2023. [Temperamental shyness and children’s communicative behaviours in a goal-oriented task: A dyadic perspective](#). *Journal of Applied Developmental Psychology*, 87:101566.
- Nils F. Tolksdorf, Camilla E. Crawshaw, and Katharina J. Rohlfing. 2021a. [Comparing the effects of a different social partner \(social robot vs. human\) on children’s social referencing in interaction](#). *Frontiers in Education*, 5(569615):1–12.
- Nils F. Tolksdorf, Katharina J. Rohlfing, Angela Grimmer, and Kazuki Sekine. 2026. [Social robots as a method for cross-cultural gesture elicitation in children](#). *Frontiers in Communication*, 11:1651149.
- Nils F. Tolksdorf, Fanziska E. Viertel, Camilla E. Crawshaw, and Katharina J. Rohlfing. 2021b. [Do shy children keep more distance from a social robot? Exploring shy children’s proxemics with a social robot or a human](#). In *Interaction Design and Children*, pages 527–531, Athens Greece. ACM.
- Claude Toussaint, Philipp T Schwarz, and Markus Petermann. 2023. [Navel - a social robot with verbal and nonverbal communication skills](#). In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–4, Hamburg Germany. ACM.
- Valeriia Tykhonenko, Nils F. Tolksdorf, and Katharina J. Rohlfing. 2026. [The relation between children’s shyness and their contingent dialogical actions when reacting to a social robot’s instructions](#). *Philosophical Transactions B*, 381(1943):20240375.
- Fanziska E. Viertel and N. F. Tolksdorf. 2024. [Multimodal communicative behaviours in shy children in assessment situations and social evaluative contexts](#). In Ozturk Nesrin, editor, *Contemporary Issues in Early Childhood Education and Care*, volume 27. IntechOpen, London.
- Anna-Lisa Vollmer, Robin Read, Dries Trippas, and Tony Belpaeme. 2018. [Children conform, adults resist: A robot group induced peer pressure on normative social conformity](#). *Science Robotics*, 3(21):eaat7111.
- McLennon Wilson, Adrienne Richter Powell, Linda Sosa Hernandez, Emma Green, Claudia Labahn, and Heather Henderson. 2024. [Shyness, social engagement, and conversational response times in children’s dyadic interactions with an unfamiliar peer](#). *Social Development*, 33(3):e12734.
- Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. [Elan: a professional framework for multimodality research](#). In *Proceedings of LREC*, volume 2006, page 5th.
- B Zellner. 1994. [Pauses and the temporal structure of speech](#). In E Keller, editor, *Fundamentals of speech synthesis and speech recognition*, pages 41–62. Chichester.

Elisabeth Zima. 2020. [Gaze and recipient feedback in triadic storytelling activities](#). *Discourse Processes*, 57(9):725–748.