Title

Child-Caregiver Gaze Dynamics in Naturalistic Face-to-Face Conversations

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Goumri, Dhia Elhak
Becerra-Bonache, Leonor
Fourtassi, Abdellah

Publication Date

2024

Peer reviewed

Child-Caregiver Gaze Dynamics in Naturalistic Face-to-Face Conversations

Dhia-Elhak Goumri (dhia-elhak.goumri@univ-amu.fr)
Aix Marseille Univ, CNRS, LIS, Marseille, France

Leonor Becerra-Bonache (leonor.becerra@lis-lab.fr)
Aix Marseille Univ, CNRS, LIS, Marseille, France

Abdellah Fourtassi (abdellah.fourtassi@gmail.com)
Aix Marseille Univ, CNRS, LIS, Marseille, France

Abstract

This study examines the development of children's gaze during face-to-face conversations, following up on previous work suggesting a protracted development in attending to the interlocutor's face. Using recent mobile eye-tracking technology, we observed children interacting with their parents at home in natural settings. In contrast to previous work, we found that children, even in early middle childhood, exhibit adult-like gaze patterns toward the interlocutor. However, differences emerge in gaze allocation between speaking and listening roles, indicating that while children may focus on faces similarly to adults, their use of gaze for social signaling, such as turn-taking cues, may still be maturing. The work underscores the critical role of social context in understanding the development of non-verbal behavior in face-to-face conversation.

Keywords: Gaze; Face-to-face conversations; Middle childhood

Introduction

Gaze is one of the most important non-verbal behaviors in face-to-face conversations. It is generally understood to play two key roles in such context: information acquisition and information signaling (Argyle & Cook, 1976; Gobel, Kim, & Richardson, 2015; Risko, Richardson, & Kingstone, 2016; Hessels, 2020).

Regarding information acquisition, directing our gaze to the interlocutor's face allows us, for example, to gather their feedback on our words and adjust accordingly. For instance, a head nod generally demonstrates successful understanding, while a frown can indicate communication trouble, requiring additional effort to explain what we mean . Further, looking at the face in conversations can help the listeners understand better; for example, seeing the speaker's lips improves speech recognition, especially in noisy circumstances (Schwartz, Berthommier, & Savariaux, 2004).

As for information signaling, gaze at the interlocutor's face (vs. gaze aversion) conveys key information that supports face-to-face interaction, such as the coordination of turns (Kendrick, Holler, & Levinson, 2023). Generally, people gaze more at the interlocutor's face when listening than when speaking. This regularity provides a helpful cue for people to anticipate their interlocutor's conversational moves (Kendon, 1967; Duncan & Fiske, 2015; Freeth, Foulsham, & Kingstone, 2013; Ho, Foulsham, & Kingstone, 2015; De Lillo et al., 2021).

While much is known about adults' gaze, little has been done to study how children's gazing patterns develop, especially in an ecologically valid setting where information acquisition and signaling are both at play. This is the question we address in this paper.

In the remainder of the introduction, we first emphasize the importance of integrating the proper social context in such an investigation. Then, we review work that has measured children's gaze in naturalistic face-to-face interactions. We end the introduction by specifying the contribution of the current study.

The Role of Social Context

A large body of research has investigated the human gaze. One of the main findings has been that we have a preference for gazing at faces (over other objects) (Yarbus & Yarbus, 1967; Amso, Haas, & Markant, 2014), starting in infancy (Johnson, Dziurawiec, Ellis, & Morton, 1991; Peltola, Yrttiaho, & Leppänen, 2018). This line of research has crucial implications regarding human social biases in general. Nevertheless, since most of these studies have relied on laboratory, non-interactive settings, they are not directly translatable to the case of face-to-face conversations (Pfeiffer, Vogeley, & Schilbach, 2013; Risko et al., 2016; Bodur, Nikolaus, Prévot, & Fourtassi, 2023).

In particular, all non-interactive settings that have been used (from observing face-like stimuli to watching videos of others interacting) focus on the information acquisition function of gaze, ignoring any effect that can result from the information-signaling awareness. However, it has been established that whether we know that our gaze is available to others has significant consequences on how much we look at them (Laidlaw, Foulsham, Kuhn, & Kingstone, 2011; Gobel et al., 2015; Cañigueral & Hamilton, 2019). Crucially, this effect is modulated by the nature of the relationships between the interlocutors (e.g., familiar person vs. stranger), the context of their interaction (e.g., cooperative vs. collaborative), and the role in the interaction (e.g., listening vs. speaking) (Kleinke, 1986; Dalmaso, Castelli, & Galfano, 2020).

Measuring Children's Gaze in the Wild

Recent technological advances have facilitated the study of gaze in face-to-face interactions and across increasingly ecologically valid settings and contexts (Pfeiffer et al., 2013;

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Ho et al., 2015; Risko et al., 2016; Dalmaso et al., 2020; Hessels, 2020). These advances include mobile eye-tracking systems and, more recently, combining these systems with robust automatic detection of social content, namely faces, from videos recorded in unconstrained contexts (Deng, Guo, Ververas, Kotsia, & Zafeiriou, 2020; Varela, Towler, Kemp, & White, 2023).

Yet, studies that use these advances to study children's gaze in face-to-face natural interaction are still lacking. There are a few exceptions, however. For instance, Schroer and Yu (2023) studied the interaction of infants aged 12 to 26 months with their caregivers while playing together with objects. While the use of mobile eye-tracking at such a young age is impressive, the study is designed to investigate optimal interactive conditions for word learning rather than the investigation of face-to-face conversation per se. It is, therefore, not directly related to our question.

More relevant to the current work is a study by De Lillo et al. (2021), who used mobile eye-tracking to monitor gaze at faces (vs. other objects) in face-to-face conversations and while navigating a natural environment. One major finding was that adolescents attended less to faces (compared to young adults) in both tasks. The reasons behind this age-related difference, especially in face-to-face conversation, were unclear. While De Lillo et al. (2021) argued that observed differences could indicate reduced cognitive control in adolescence, they also provided alternative explanations related to the social context. In particular, the experimenter – the interlocutor during face-to-face conversations – was also a stranger, which could have played in disfavor of children because of known effects of age difference and social rank asymmetry on gaze (Kleinke, 1986; Gobel et al., 2015).

The current study

As we saw above, attention to the interlocutor's face is heavily context-dependent. Studying children's gaze in face-to-face conversations can be tricky as it is not always easy to disentangle developmental differences from contextual factors. The main focus of the current work is, thus, on mitigating the latter's effects to characterize the former better.

To this end, we put children in a more favorable social context than in previous studies. First, instead of inviting children to the lab, an unfamiliar environment where they may feel shy, we observed their behavior at home, where they feel more comfortable and in control (see aslo Bodur, Nikolaus, Kassim, Prévot, & Fourtassi, 2021; Shi, Gu, & Vigliocco, 2022). Second, children did not converse with an experimenter, a stranger who would also be older and more intimidating. Instead, they talked to one of their parents, i.e., someone with whom they have high interpersonal intimacy (Argyle & Dean, 1965).

Finally, while De Lillo et al. (2021) studied adolescents, we observed a younger age group in middle childhood (7 to 11 years old). Suppose there are differences in adolescents vs. adults (due to the development of cognitive or social skills). In that case, we should be able to observe these differences even more clearly in this younger age group.

Figure 1: Eye-tracking glasses were calibrated before each use by asking participants to fixate on an object in the house. The gaze marker (the red circle) was then adjusted to match the target object using a real-time camera and gaze signal streaming.

We investigated two questions. The first concerns children's general attention to the interlocutor's face in the conversation. More specifically, we ask i) whether school-age children as young as seven years of age look consistently and preferentially to the face of their interlocutor and ii) whether children show reduced attention to faces compared to adults (replicating findings in De Lillo et al. (2021) with adolescents).

The second question concerns children's use of gaze for effective information-signaling in face-to-face conversation. In particular, we test i) whether and how children's gaze is modulated by their role in the conversation as speakers vs. listeners, and ii) whether this modulation – if there is one – is as strong (and thus, provides as effective of a signal to the interlocutor) as the one typically shown by adults (Freeth et al., 2013; De Lillo et al., 2021).

Methodology

Corpus Collection

Participants We collected face-to-face data from $N = 30$ French-speaking subjects (15 dyads). The dyads were made of children interacting with one of their parents. The data included three age groups: children around 7 (hereafter "Younger"), 9 ("Middle"), and 11 ("Older"), with five dyads per age group. The genders of both children and parents were balanced. More details can be found in Goumri et al. (2024).

Conversational task The dyads play a weakly structured word-guessing game in which one person chooses a word, and the other attempts to find it by asking different questions (not just yes-or-no questions). Upon correctly guessing a word, the players switch roles. To foster flexibility, we encouraged participants to provide suggestions and ask questions as they see fit. They were also free to digress or com-

Average age	Av. conversation duration	Speaker	Av. no. words	Av. no. segments	Av. seg. duration (s)
7;3 (+/- 3.3 months)	15min 32sec	child	331	142	2.46
		parent	664	202	2.44
9;5 (+/- 3.4 months)	16min 32sec	child	446	189	2.26
		parent	660	211	2.53
11;3 (+/- 4.1 months)	14min 13sec	child	408	154	2.07
		parent	690	195	2.33

Table 1: Demographics of the participants and key conversation metrics

ment on each other's performance, the goal being to elicit a spontaneous and balanced conversation. The game ends after roughly 15 minutes, provided each player has guessed a similar number of words (typically around three words each).

Equipment The recordings were done at each family's house. The child and the caregiver were seated in the same room on chairs facing each other. The researcher brought two mobile eye-tracking devices, one for each participant.

The eye-tracking device (Pupil invisible) is manufactured by pupil-labs (Tonsen, Baumann, & Dierkes, 2020). It is lightweight (less than 50 grams) and shares a similar size and shape to conventional eyeglasses. Each device is equipped with the following:

- External Scene Camera, attached to the eyeglasses' left temple, operates at $30\,Hz$ and features a 1088×1080 pixels resolution, with a field of view measuring 82×82.
- Internal Eye Cameras capturing footage of the left and right eyes, sampled at a rate of $200\,Hz$. Each eye camera is coupled with an IR LED to ensure adequate illumination.
- Integrated microphones.

Procedure Initially, the researcher ensured that the lighting and chair arrangement met the minimal requirement. Participants were then assisted in wearing the Pupil Invisible device. For children, we used an additional head strap to tighten the eyeglasses. Before each use, the researcher calibrated the device by instructing participants to focus on various objects in the surroundings and adjusting the gaze marker to align with the target object (see Figure 1).

Before starting the game, the researcher initiated a clap to serve as an audio-visual marker for later synchronization. Subsequently, when possible, the researcher withdrew to a corner or a different room, conveying engagement in a separate work-related activity to minimize interference or the perception of being observed by a third party.

Data Processing

Audio data We used WhisperX (Radford et al., 2023; Bain, Huh, Han, & Zisserman, 2023) for audio processing. We extracted speech segments, defined as stretches of speech by one speaker, segmented using Voice Activity Detection. Within each segment, words were recognized, transcribed, and forced-aligned with the speech signal.

An important processing step for our study is speech diarization, that is, determining whether the child or the parent uttered a given speech segment. This is crucial since one of our goals is to determine the gazing behavior while speaking (vs. while listening).

Despite each speaker employing a separate microphone, this information was unclear (due to the relative spatial proximity of the child and caregiver, both microphones picked up speech from both interlocutors). WhisperX incorporates an automatic diarization module based on "*pyannote*" (Bredin et al., 2020). However, it did not yield satisfactory results on our data. Thus, we resorted to full manual labeling of speakers for each speech segment. Table 1 provides summary statistics of the corpus based on this processing.

Gaze on Face estimation The device estimates the gaze data using an end-to-end deep learning algorithm. Gaze data is then projected into the 2D-pixel space of the scene camera (see Tonsen et al. (2020) for further technical details and Figure 1 for an illustration). To detect the interlocutor's face in the scene videos, RetinaFace was used; it is a state-of-the-art algorithm for automatic face detection in the wild (Deng et al., 2020). It outputs a bounding box for the face.

A crucial technical step for our study is estimating when the gaze is on the face. To this end, we used and compared two automated methods. The first simply measures when the projected gaze coordinates lie within the bounding box of the detected face.[1] The second method proceeds bottom-up and relies on the distribution of gaze in the conversations. Fixations at the interlocutor's face tend to form a cluster in the visual space. We used this cluster to derive a threshold for each participant. [2]

[1]Note that, upon manual investigation, we realized the original box was too narrow to cover the effective face area attended to. For example, if a participant happens to adjust the device on their head, this can sometimes cause their gaze data to be projected slightly outside the box. Doubling the area of the face's bonding box was sufficient to solve such issues.

[2]We will explain this method in more detail in the Results section, as it requires background related to our first exploratory analysis.

Data Synchronisation We captured data streams at different levels (video, audio, and gaze). The time synchronization of these streams is crucial as one of our primary scientific goals is to investigate how gazing at the face correlates with talking vs. speaking in the conversation. First, we aligned the scene videos from both participants using the audio marker provided by the researcher's clap at the beginning of each conversation. Each gaze data was synchronized with the corresponding video, accounting for differences in sampling rate. Finally, the timestamps of the recording device allowed synchronization with the audio data.

Results

We report three findings. The first concerns an early preference for gazing at the interlocutor's face in face-to-face conversation. The second concerns how gaze is modulated by the role in the conversation (i.e., listening vs. speaking). The third describes how gazing effects manifest in short vs. long speech turns.

Preferential gaze at the face

We examine the spatial distribution of gaze in the conversation as follows. For each time frame t, the gaze is characterized with coordinates $(x_G(t), y_G(t))$ (G for gaze), indicating the point in the scene the participant is looking at. However, the scene changes when the participant moves their head; it is, thus, essential to anchor the gaze in a known reference point. We naturally used the interlocutor's face as our reference. The face was also detected dynamically: For each time frame t, RetinaFace outputs a face bounding box whose center is characterized with the coordinates $(x_F(t), y_F(t))$ (F for face). Using these two dynamic points, we characterize the spatial location of gaze relative to the interlocutor's face using the Euclidean Distance:

$$d(t) = \sqrt{(x_G(t) - x_F(t))^2 + (y_G(t) - y_F(t))^2}$$

Figure 2 shows the distribution of gaze distances $d(t)$ over entire conversations, averaged across all participants. For both children and adults, we can see a peak in the distribution around 0 (i.e., around the interlocutor's face box center), demonstrating that the face is the location where the gaze is the most highly concentrated.

In Figure 3, we show the exact data for each participant separately. In all these individual distributions, we similarly observe a clear peak around the face of the interlocutor, including in the youngest age group. This data shows that preferential gaze at the interlocutor's face in conversations is robust, systematic, and well-established starting from early middle childhood.

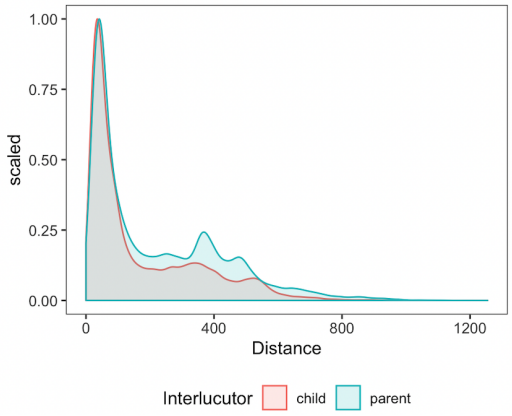

Figure 2: The kernel density estimate (i.e., a smoothed version of a histogram for continuous data) shows the scaled distribution of gaze as a function of the distance (in pixels) from the center of the interlocutor's face box. Here, we show the average distribution across all participants.

In many of these graphs, we observe a second (and sometimes a third) peak at distances larger than zero. These secondary peaks correspond to objects in the surroundings that participants also attended to regularly besides the face (e.g., see Figure 1).

Gazing while talking vs. listening

We investigated whether participants' gaze was modulated by their role in the conversation (i.e., talking vs. listening). To this end, for each speech segment in a given conversation, we calculated the proportion of time (during this segment) that was spent gazing at the interlocutor's face. For the same segment, we calculated two values: One for the speaker (the person uttering the segment) and one for the listener (the person listening to this segment).

We categorized gaze data as looking at the face vs. looking away, comparing two methods (see also Methods section). The first method we used considers the face box detected automatically as the Area Of Interest (AOI). The second method we used considers a different AOI for each participant based on the variance of the first peak in their continuous data (shown in Figure 3). Both methods led to very similar conclusions. The results we report in the following are based on the second method, which we deemed more precise as it accounts for each participant's idiosyncracies regarding face gazing (the results using the other method will be provided in supplementary material).

Figure 4 shows the results of this analysis. For presentation clarity, this Figure is based on participants' data averaged across all speech segments in conversations (but a more detailed, segment-level analysis will be shown in the following sub-section). We observe that i) children spent a longer time gazing at the interlocutor's face compared to adults, ii) both children and adults gazed longer at the interlocutor's face when listening than when talking, and iii) The difference in listening vs. talking was larger in adults (these quali-

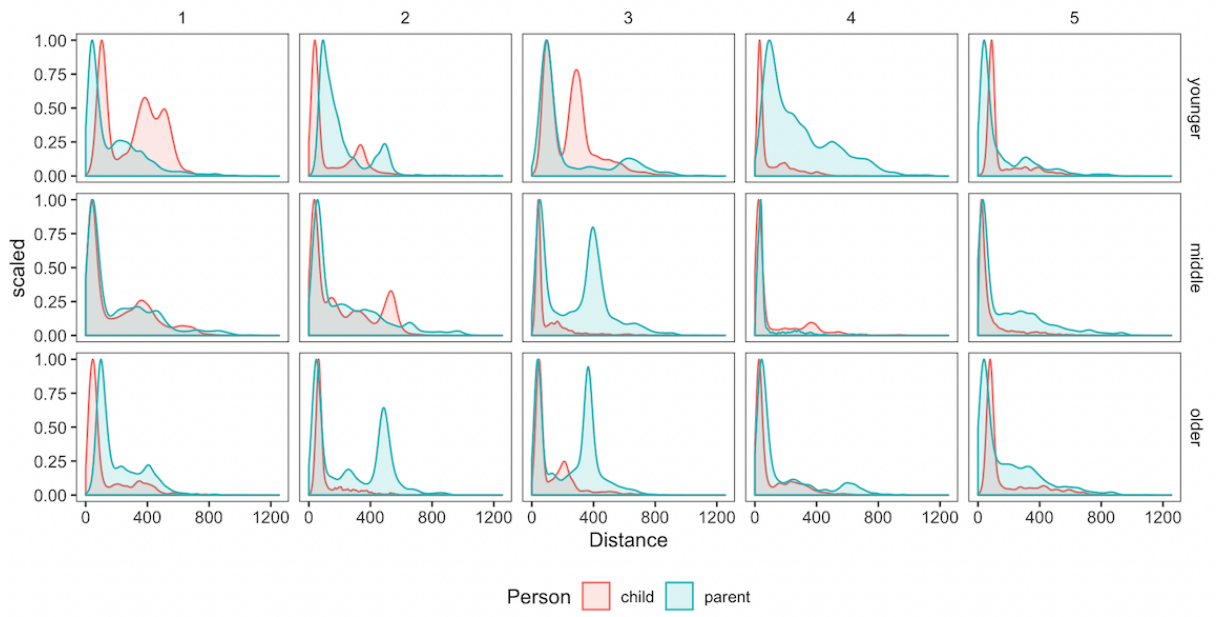

Figure 3: The kernel density estimate (i.e., a smoothed version of a histogram for continuous data) showing the scaled distribution of gaze as a function of the distance (in pixels) from the center of the interlocutor's face box.

tative observations were corroborated with statistical testing, below).

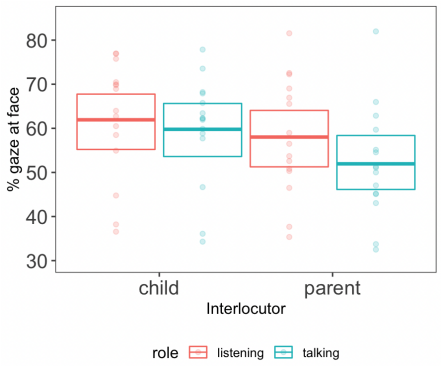

Figure 4: The proportion of time spent gazing at the face. Each dot corresponds to a given participant, representing the average percentage across all segments in their conversation. The bar and box represent the mean and confidence intervals over these averages.

The Influence of Speech Segment Duration

As a follow-up analysis, we investigate how much the segment's duration influences gaze. First, by examining the distribution of segments' duration (Figure 5), we see that a significant portion of segments is very short (less than 1 second) or relatively long (longer than 3 seconds). However, the majority of segments are between 1 and 2 seconds. These numbers align with turn-taking literature in adults, where the average and median turn duration have been reported to be between 1 and 2 seconds (Levinson & Torreira, 2015).

Figure 6 shows the gaze results at the segment level. More specifically, it shows the proportion of time spent gazing at the interlocutor's face as a function of segment duration (up to five seconds). We observe that longer segments lead to more significant differences in listening vs. talking in children and adults.

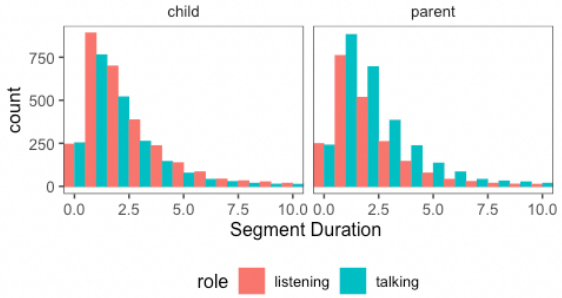

Figure 5: The histogram of segments' duration. Note that the segments children listen to are the same as the ones that parents produce (and vice versa). Hence, the left and right panels contain the same data but with reversed colors.

Statistical Testing

Statistical testing corroborated all the above-made qualitative observations based on Figure 4 and 6. We fit a mixed-effects linear model predicting the proportion of time spent at the interlocutor's face for each conversational turn, using as predictors the *Age group* (Younger, Middle, and Older), *Identity* (child vs. parent), *Role* in the speech segment (listening vs. talking), and the *Duration* of the segment. Given the depen-

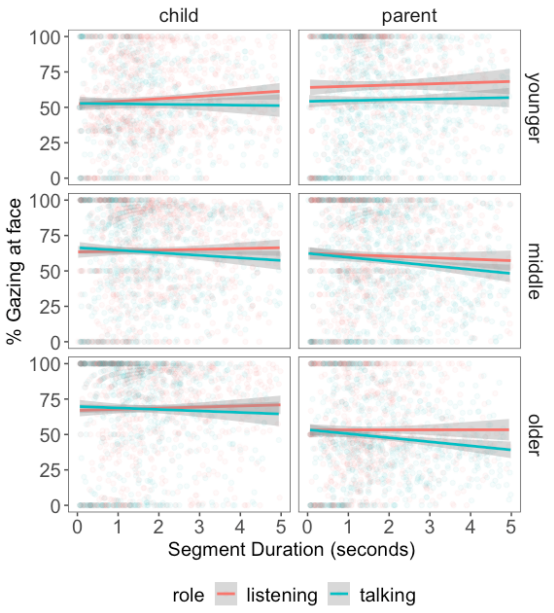

Figure 6: The proportion of time spent gazing at the face as a function of segment duration. Each dot corresponds to a segment in a conversation, showing its duration (x-axis) and the percentage of this segment spent gazing at the interlocutor's face (y-axis). The lines and their envelopes represent the linear fits and confidence intervals.

dencies we observed, we also tested the interaction of *Role* by *Duration* and *Role* by *Identity*. The model was specified as follows:

```
gaze_at_face ~ Age + Role*Identity* +
Role*duration + (1 | dyad)
```

First, we observed a robust effect of *Identity* (child vs. adult) on gaze whereby children gazed more at faces: β = 5.387 ($SE = 0.92$, $p < 0.001$). The interaction of *Role* by *Identity* was significant β = 2.88 ($SE = 1.30$, $p < 0.05$), as was the interaction of *Role* by turn *Duration* β = 0.58 ($SE = 0.29$, $p < 0.05$), confirming that the difference in listening vs. talking was larger in adults and with longer segments, respectively. Finally, *Age* did not affect gaze β = 0.99 ($SE = 2.68$, $p = 0.71$).

We followed up on the *Role* by *Identity* interaction by fitting simpler models examining each *Identity* (child or parent) separately. For children, the effect was significant $\beta = -2.06$ ($SE = 0.88$, $p < 0.05$) and, indeed, smaller than in adults: β = 5.32 ($SE = 0.89$, $p < 0.001$). Finally, we found no effect of Age group on gaze in either children or adults.

Discussion

How do children develop in terms of their ability to use gaze in face-to-face conversation? Previous research has found evidence for an apparent developmental delay in attending to the interlocutor's face, well into adolescence (De Lillo et al., 2021). While this finding can be associated with immature Executive Functions and Theory of Mind (Humphrey & Dumontheil, 2016), here we investigate if it can also reflect the social context where the measurement is done. Indeed, we know that how much one looks at the interlocutor's face depends not only on information acquisition/processing skills but also on social-signaling awareness in context (Kleinke, 1986; Laidlaw et al., 2011; Gobel et al., 2015; Risko et al., 2016; Cañigueral & Hamilton, 2019; Dalmaso et al., 2020).

Using recent mobile eye-tracking and automatic processing methods of data in the wild (Varela et al., 2023), we observed children's gaze in a natural and familiar context: They conversed with one of their parents at home, playing a weakly structured word-guessing game that afforded a spontaneous conversation. The results strongly supported the hypothesis: Children, including those in early middle childhood, did not show a noticeable difference compared to adults regarding how much they looked at the interlocutor's face. If anything, we found that children gazed more (Figures 2 and 4).

Interestingly, however, while children were not delayed compared to adults in terms of *total* amount of time spent gazing at the interlocutor's face, they did show a difference in terms of *how* they allocated gaze depending on their role in the conversation, i.e., in listening vs. speaking. Previous observational and eye-tracking methods showed that interlocutors tend to gaze more while listening than speaking, providing a helpful cue to the interlocutor regarding turn-taking management (Kendon, 1967). Here, we found that adults' gazing patterns provided a clearer cue than children's (Figure 4 and 6), although whether this cue had an actual effect on turn-taking dynamics is unclear. Further work is needed to investigate this interpretation, including by considering the fine-grained temporal dynamics (Ho et al., 2015; Liu, Nikolaus, Bodur, & Fourtassi, 2022; Agrawal, Liu, Bodur, Favre, & Fourtassi, 2023).

Limitations The present study has certain limitations. For instance, measuring children in their natural "habitat" introduced variability (despite our best efforts for mitigation), e.g., in terms of lighting conditions, the distance between interlocutors, and the degree of sound noise and visual distraction in the background. Another limiting factor is our use of speech segments to characterize moments of listening vs. speaking. While this choice is motivated by the goal of measuring gaze in spontaneous, daily conversations where turns are short and fast (Levinson, 2016), it may have introduced some imprecision, especially for estimating gaze within very short turns (e.g., yes/no answers) (see Figure 6). Such variability/imprecision may have conspired to underestimate some effects, especially in listening vs. speaking. That said, it does not explain the difference we observed between children vs. adults since interlocutors within a dyad were observed under similar conditions.

Acknowledgement

This work was carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002). It has benefited from support from the French government (France2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix Marseille University (A*MIDEX). Furthermore, this study was also supported by the ANR MACOMIC (ANR-21-CE28-0005-01) grant.

References

Agrawal, A., Liu, J., Bodur, K., Favre, B., & Fourtassi, A. (2023). Development of multimodal turn coordination in conversations: Evidence for adult-like behavior in middle childhood. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45). Retrieved from https://hal.science/hal-04411367/document

Amso, D., Haas, S., & Markant, J. (2014). An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PloS one*, *9*(1), e85701.

Argyle, M., & Cook, M. (1976). Gaze and mutual gaze.

Argyle, M., & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, 289–304.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Bodur, K., Nikolaus, M., Kassim, F., Prévot, L., & Fourtassi, A. (2021). Chico: A multimodal corpus for the study of child conversation. In *Companion publication of the 2021 international conference on multimodal interaction* (pp. 158–163). doi: https://doi.org/10.1145/3461615.3485399

Bodur, K., Nikolaus, M., Prévot, L., & Fourtassi, A. (2023). Using video calls to study children's conversational development: The case of backchannel signaling. *Frontiers in Computer Science*, *5*. doi: https://doi.org/10.3389/fcomp.2023.1088752

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... Gill, M.-P. (2020). Pyannote. audio: neural building blocks for speaker diarization. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7124–7128).

Cañigueral, R., & Hamilton, A. F. d. C. (2019). Being watched: Effects of an audience on eye gaze and prosocial behaviour. *Acta psychologica*, *195*, 50–63.

Dalmaso, M., Castelli, L., & Galfano, G. (2020). Social modulators of gaze-mediated orienting of attention: A review. *Psychonomic Bulletin & Review*, *27*, 833–855.

De Lillo, M., Foley, R., Fysh, M. C., Stimson, A., Bradford, E. E., Woodrow-Hill, C., & Ferguson, H. J. (2021). Tracking developmental differences in real-world social attention across adolescence, young adulthood and older adulthood. *Nature human behaviour*, *5*(10), 1381–1390.

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5203–5212).

Duncan, S., & Fiske, D. W. (2015). *Face-to-face interaction: Research, methods, and theory*. Routledge.

Freeth, M., Foulsham, T., & Kingstone, A. (2013). What affects social attention? social presence, eye contact and autistic traits. *PloS one*, *8*(1), e53286.

Gobel, M. S., Kim, H. S., & Richardson, D. C. (2015). The dual function of social gaze. *Cognition*, *136*, 359–364.

Goumri, D., Agrawal, A., Nikolaus, M., Vu, H., Bodur, K., Semmar, E., ... Fourtassi, A. (2024). Chica: A developmental corpus of child-caregiver's face-to-face vs. video call conversations in middle childhood. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation.*

Hessels, R. S. (2020). How does gaze to faces support face-to-face interaction? a review and perspective. *Psychonomic bulletin & review*, *27*(5), 856–881.

Ho, S., Foulsham, T., & Kingstone, A. (2015). Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS one*, *10*(8), e0136905.

Humphrey, G., & Dumontheil, I. (2016). Development of risk-taking, perspective-taking, and inhibitory control during adolescence. *Developmental Neuropsychology*, *41*(1-2), 59–76.

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, *40*(1-2), 1–19.

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

Kendrick, K. H., Holler, J., & Levinson, S. C. (2023, April). Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *378*(1875), 20210473.

Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological bulletin*, *100*(1), 78.

Laidlaw, K. E., Foulsham, T., Kuhn, G., & Kingstone, A. (2011). Potential social interactions are important to social attention. *Proceedings of the National Academy of Sciences*, *108*(14), 5548–5553.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences*, *20*(1), 6–14.

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

Liu, J., Nikolaus, M., Bodur, K., & Fourtassi, A. (2022). Predicting backchannel signaling in child-caregiver multimodal conversations. In *Companion publication of the 2022 international conference on multimodal interaction* (p. 196–200). Retrieved from https://doi.org/10.1145/3536220.3563372

Peltola, M. J., Yrttiaho, S., & Leppänen, J. M. (2018). Infants' attention bias to faces as an early marker of social

development. *Developmental science*, *21*(6), e12687.

Pfeiffer, U. J., Vogeley, K., & Schilbach, L. (2013). From gaze cueing to dual eye-tracking: novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews*, *37*(10), 2516–2528.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science*, *25*(1), 70–74.

Schroer, S. E., & Yu, C. (2023). Looking is not enough: Multimodal attention supports the real-time learning of new words. *Developmental Science*, *26*(2), e13290.

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69–B78.

Shi, J., Gu, Y., & Vigliocco, G. (2022). Prosodic modulations in child-directed language and their impact on word learning. *Developmental Science*, e13357. doi: https://doi.org/10.1111/desc.13357

Tonsen, M., Baumann, C. K., & Dierkes, K. (2020). A high-level description and performance evaluation of pupil invisible. *arXiv preprint arXiv:2009.00508*.

Varela, V. P., Towler, A., Kemp, R. I., & White, D. (2023). Looking at faces in the wild. *Scientific Reports*, *13*(1), 783.

Yarbus, A. L., & Yarbus, A. L. (1967). Eye movements during perception of complex objects. *Eye movements and vision*, 171–211.