CHICA: A Developmental Corpus of Child-Caregiver's Face-to-face vs. Video Call Conversations in Middle Childhood

Dhia-Elhak Goumri[1], Abhishek Agrawal[1], Mitja Nikolaus[2], Duc Thang Vu Hong[1], Kübra Bodur[3], Elias Semmar[1], Cassandre Armand[4], Chiara Mazzocconi[3,5], Shreejata Gupta[3,4], Laurent Prévot[3], Benoit Favre[1], Leonor Becerra-Bonache[1], Abdellah Fourtassi[1]

[1]Aix Marseille Univ, CNRS, LIS, Marseille, France
[2]CerCo, CNRS, Toulouse, France
[3]Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
[4]Aix Marseille Univ, CNRS, CRPN, Marseille, France
[5]Aix Marseille Univ, INSERM, INS, Marseille, France
{first_name.last_name}@univ-amu.fr

Abstract

Existing studies of naturally occurring language-in-interaction have largely focused on the two ends of the developmental spectrum, i.e., early childhood and adulthood, leaving a gap in our knowledge about how development unfolds, especially across middle childhood. The current work contributes to filling this gap by introducing CHICA (for Child Interpersonal Communication Analysis), a developmental corpus of child-caregiver conversations *at home*, involving groups of French-speaking children aged 7, 9, and 11 years old. Each dyad was recorded twice: once in a face-to-face setting and once using computer-mediated video calls. For the face-to-face settings, we capitalized on recent advances in mobile, lightweight eye-tracking and head motion detection technology to optimize the naturalness of the recordings, allowing us to obtain both precise and ecologically valid data. Further, we mitigated the challenges of manual annotation by relying – to the extent possible – on automatic tools in speech processing and computer vision. Finally, to demonstrate the richness of this corpus for the study of child communicative development, we provide preliminary analyses comparing several measures of child-caregiver conversational dynamics across developmental age, modality, and communicative medium. We hope the current corpus will allow new discoveries into the properties and mechanisms of multimodal communicative development across middle childhood.

Keywords: Developmental Corpus, Child-Caregiver Conversations, Middle Childhood

1. Introduction

To be considered fully competent speakers of their native language, children need to develop not only theoretical knowledge of various linguistic structures (e.g., phonology, syntax, and vocabulary) but also interactive skills that would allow them to *use* language appropriately in daily, face-to-face conversations (Clark, 1996).

Researchers have identified and studied a variety of skills that play an essential role in adult conversations. These are, therefore, skills that children are supposed to learn so as to achieve adult-level mastery, including — among other things — abilities to manage turn-taking (Levinson, 2016; Casillas et al., 2016; Nguyen et al., 2022), listener's feedback (Bavelas et al., 2000; Dingemanse, 2024; Bodur et al., 2023), repair strategies (Purver, 2004; Dingemanse and Enfield, 2024; Clark, 2018), contingent/coherent responses (Grice, 1975; Keenan and Klein, 1975; Bloom et al., 1976; Abbot-Smith et al., 2023) and interactive alignment (Pickering and Garrod, 2021; Chieng et al., 2024; Fusaroli et al., 2023; Misiek et al., 2020; Misiek and Fourtassi, 2022).

From a developmental point of view, the majority of studies have focused on documenting the precursors of these skills in children's non-verbal stages or on their earliest verbal signs, often in an effort to study their potential role in helping with children's language learning (e.g., vocabulary development) in infancy and pre-school (e.g., Donnelly and Kidd, 2021; Elmlinger et al., 2023; Nguyen et al., 2022; Clark, 2018; Masek et al., 2021; Nikolaus and Fourtassi, 2023).

There is very little work on how children develop these conversational skills beyond the early years of language development and how these skills reach adult-like maturity. Indeed, the few existing studies point towards a rather protracted developmental trajectory that would span much of middle childhood and sometimes well into adolescence (Maroni et al., 2008; Baines and Howe, 2010; Hess and Johnston, 1988; Sehley and Snow, 1992).

Further, much of the socio-cognitive competencies that are generally understood to underlie conversational skills such as mentalizing (i.e., the ability to infer people's mental state) and various executive functions such as inhibitory control, working memory, and metacognition (see Matthews et al., 2018) undergo significant changes across middle childhood (e.g., Wang et al., 2016). This fact invites a much deeper investigation into how

LREC-COLING 2024, pages 3153–3164

changes in socio-cognitive skills enable new and more sophisticated conversational abilities across the same period.

Towards a naturalistic corpus

Observing conversation under the most natural conditions is a necessary step to accurately identifying and characterizing conversational phenomena (Sacks et al., 1974; Schegloff, 1991). Indeed, conversation is unique in that it involves a process of reciprocal monitoring and adaptive adjustments between interlocutors. This process is inherently spontaneous and unpredictable; it cannot be captured entirely with strictly controlled designs; in fact, doing so runs the risk of "compromising the naturally occurring constitution of talk-in-interaction" (Schegloff, 1996).

That said, well-curated developmental corpora are scarce; they constitute a notoriously challenging research endeavor. The major impediment is the need for resource-intensive manual annotation. This challenge can, however, be mitigated with recent technological advances both in the precision of mobile measurement tools (e.g., mobile eye-tracking systems and motion detection) and in machine learning tools (e.g., in speech processing and computer vision). These advances can be a game changer for the ecological study of multimodal conversational interaction and its development. The current work is an attempt to make full use of these technologies.

We sought to obtain the most naturalistic data we could. To this end, we made the following decisions in terms of context, task, and measurement device:

- **Ecological context:** We recorded conversations involving school-age children talking with their parents at their homes. Conversing with an experimenter in the lab would have been less tedious regarding data collection. It would also have provided some degree of control (i.e., same conversational partner) across children. However, it would not have been optimal from a naturalistic point of view: Social interactions are known to be highly *context-sensitive* (Dideriksen et al., 2023; Kleinke, 1986; Risko et al., 2016; Bodur et al., 2023); the way a child would talk to an experimenter (a stranger) in the lab (an unfamiliar place) might not be indicative of their spontaneous behavior under more familiar conditions where they can show more of their natural competences, namely, their conversation with parents at home.
- **Intuitive elicitation Task:** The goal is to elicit an exchange that would be as representative as possible of child-parent spontaneous conversations. Researchers have traditionally used physical prompts to elicit conversations, such as the maze game, the map task, or the spot-the-difference tasks (Anderson et al., 1993; Garrod and Anderson, 1987; Van Engen et al., 2010). However, we realized in piloting that such prompts tend to absorb children's attention, making the face-to-face multimodal interaction sub-optimal. Thus, we opted for an easy and prompt-free elicitation where the interlocutors play a loosely structured word-guessing game, switching roles whenever a word has been guessed (see also Pincus and Traum, 2016). In addition to optimizing face-to-face signaling, this game also allowed us to mitigate the *social asymmetry* effect: When the interaction is left entirely free and unstructured, our piloting showed that parents tend to orchestrate the dialog, often resulting in imbalanced exchange.
- **Light measurement device:** We were interested in capturing direct and precise measurements of eye-gaze behavior and head movement; two behaviors known for their crucial role in regulating conversations and social interactions more generally (e.g., Kendon, 1967; Hale et al., 2020) and for their relatively late development into adolescence (Hess and Johnston, 1988; De Lillo et al., 2021), making it relevant to investigate developmentally across middle childhood. We used mobile sensors consisting of an eye-tracking system, a gyroscope (measuring angular velocity), and an accelerometer (measuring linear acceleration). To maintain a high degree of naturalness, we used recent technology integrating all these sensors into one lightweight device that looks and feels like normal eyeglasses (Tonsen et al., 2020), thus making it less likely to limit/hinder the speaker and also less likely to distract the listener. This same device has proven to provide a good measure of gazing patterns in natural settings in previous research, including with young children (e.g., Schroer and Yu, 2023).

Face-to-face vs. Video calls

To better understand the specificities of face-to-face conversation, we contrast it with another popular medium of communication: Video calls. We optimize our ability to draw valid conclusions from this comparison by adopting a within-dyad design: The child and parent played the same conversational game both via video call and face-to-face. Here again, to make the conversation relatively naturalistic, the video call conversation takes place at home: The caregiver and child used different devices and they communicated from different rooms. They were instructed to act as if they were communicating from remote places.

Related corpora

Most existing multimodal corpora of child-caregiver interaction either used a third-person-view camera (most corpora in the CHILDES repository MacWhinney, 2000) or a head-mounted camera providing an egocentric view of the child (Sullivan et al., 2021). While these corpora continue to play a crucial role in the study of language development (e.g., Vong et al., 2024), they do not allow clear access to the interlocutors' facial expressions and gestures and, therefore, are not ideal for the specific study of face-to-face interaction. A notable exception is the Ecolang corpus (Shi et al., 2023), which, however, investigates child-caregiver interaction at a much younger age. The closest multimodal corpus to ours, at least in the age range, is the corpus introduced in Bodur et al. (2021). This corpus, however, was made of video calls only and was designed to compare conversational skills of school-age children to adults, and not to study *development* across middle childhood. Thus, our corpus is, to the best of our knowledge, the first developmental corpus (involving three age groups) of child-caregiver conversations, comparing *both* face-to-face and computer-mediated video calls.

The paper is organized as follows. First, we specify details regarding participants, tasks, logistics, and recording procedures. Then, we describe data processing steps involving synchronization, transcriptions, and annotation. Finally, we provide preliminary analyses comparing several measures of child-caregiver conversational dynamics across developmental age, modality, and communicative medium, showcasing the richness and potential of the corpus.

2. Recording Methods

2.1. Participants

The target sample is $N = 30$ of French-speaking child-caregiver dyads, 10 dyads per age group (around 7, around 9, and around 11 years old). Procuring datasets of this nature presents considerable challenges, primarily due to the difficulty in recruiting volunteers willing to undertake recordings with their children in their home environments. The aim is to strike a balance between obtaining a feasible sample size and ensuring rich and ecological intra-individual data.[1]

N	Av. Age	Parents
5 (F=2)	7;3 (+/- 3.3 months)	F = 3
5 (F=3)	9;5 (+/- 3.4 months)	F = 3
5 (F=2)	11;3 (+/- 4.1 months)	F = 2

Table 1: The distribution of our 15 dyads across age groups, children's gender, and parent's gender. The children's average age is given in years;months (+/- average deviation from the mean).

In the current manuscript, we report processing steps, annotation, and preliminary analyses from half of this target sample (15 dyads), amounting to about 4 hours of face-to-face conversations and 4 hours of video calls. See demographic information in Table 1).

2.2. Task

Each dyad plays a word-guessing game in which one of the participants thinks of a word and the other tries to find it by asking all sorts of questions (and not just yes-no questions). To make the task less rigid, each dyad was told to take the freedom to ask and give hints as they deemed necessary. After a word had been guessed, the interlocutors switched roles. The parents were told that they could stop the game after 10 to 15 minutes and as long as both had guessed a similar number of words (to keep the conversation balanced).

2.3. Logistics and Equipment

The video call recording step:

The logistics required from the parents were:

- Two devices: either two computers or a computer and a tablet/smartphone. If the family had only one computer, we recommended that the child use it (to optimize recording stability), while the parent uses the tablet/smartphone (put in a stable position). Both devices should be equipped with a functioning microphone and a camera.
- A high-speed internet connection.
- The Zoom software should be installed and tested on both devices.

[1] Our previous research shows that samples of this magnitude, especially when using the task (described next), provide rich intra-individual data that can be adequate for a wide range of analyses and modeling tasks (Agrawal et al., 2023; Liu et al., 2022; Bodur et al., 2023; Goumri et al., 2023; Mazzocconi et al., 2023).

- Two rooms from which a video call can be done. These rooms need to be distant enough so that the child and parent can hear each other only via the video call (and not through the walls).

The face-to-face recording step: The logistics required from the family were only a room and two chairs.

Additionally, the researcher brought with them the following equipment:

- 2 wearable devices integrated with eye-tracking and head movement detection ("pupil invisible" Tonsen et al., 2020). The device is lightweight (< 50g). Its size and shape are very similar to typical eyeglasses (144mm x 48mm X 160mm).
- 2 smartphones Samsung A135 with a 8160 x 6120px camera resolution and a microphone.
- 2 tripods to hold the smartphones while recording a fixed, frontal view of each interlocutor during the conversation.

The main characteristics of the sensors integrated into the Pupil invisible device are:

- Internal Eye Cameras (filming the left and right eyes) sampled at 200Hz.
- Gyroscope and Accelerometer (to track head movement) sampled at 200Hz.
- External Scene Camera, sampled at 30Hz with 1088 x 1080px resolution and a field of view of 82°x82°.
- A microphone is integrated into the scene camera component.

2.4. Procedure

Interested families filled out an online form. The form linked to documents that explained the procedure in detail (including the required hardware and software) as well as to the documents related to the consent and data protection forms. Parents had to agree to the procedure and give their consent before they could move ahead with the registration (see also the Ethics section below).

Data collection was done in two steps: 1) video-call recording, and 2) face-to-face recording,[2] as follows (see also Figure 1):

Video call recording: Parents booked an online appointment with the researcher, via Zoom. During the online appointment, the researcher explained the procedure for the video recording step, which is also done using Zoom. He made sure that a) both the child and parents were well positioned (fully visible on the screen), b) that there were no sound issues or echoes (in case the child and parent's devices were not distant enough from each other), and c) that both have checked "hide self-view" and pinned the other interlocutor's window on full-screen mode (so the child only sees the parent and vice versa).

The researcher remained in the Zoom conversation while making sure he had disappeared completely from view. This was achieved by the researcher turning his camera (and microphone) off and asking both participants to check "hide non-video participants." Once the participants are ready, the researcher starts the recording and the dyad starts playing the word-guessing game. After they are done, the researcher stops the recording and turns on his camera so that he can reappear in view. He then congratulates the child and organizes the next step (face-to-face recording session) with the parents.

Face-to-face recording: At the end of the Zoom recording, the researcher and parent convened for a future in-person appointment at the family's home. The recording procedure during the appointment was as follows. The researcher first makes sure the lighting and chair arrangement are adequate. Once the interlocutors are seated, the researcher installs the tripods (with recording smartphones) behind each interlocutor, verifying that they capture a clear, frontal view of the other interlocutor. The researcher helps the participants wear the Pupil Invisible device, and if necessary, uses dedicated head straps (by the same manufacturer) to tighten it, especially for children. The researcher calibrated the device before each use: This was done by asking each participant to fixate on different objects in the house (without moving their head) and then adjusting the gaze marker to match the target object.[3] This process was facilitated by software allowing real-time streaming – on a dedicated smartphone – of the gaze data, overlaid on the egocentric view of the participant (see Figure 2).

Note that the eyeglasses needed to be cable-connected to a smartphone for all computation and storage of the recorded data. While this choice from the manufacturer can be understood as making the device itself less cumbersome, the cable adds some constraints on movement. In our case,

[2]For practical reasons, these two steps were always in this order.

[3]In most cases this device did not need adjustment as it has been manufactured to adapt to each participant without calibration. We still performed this action before each use.

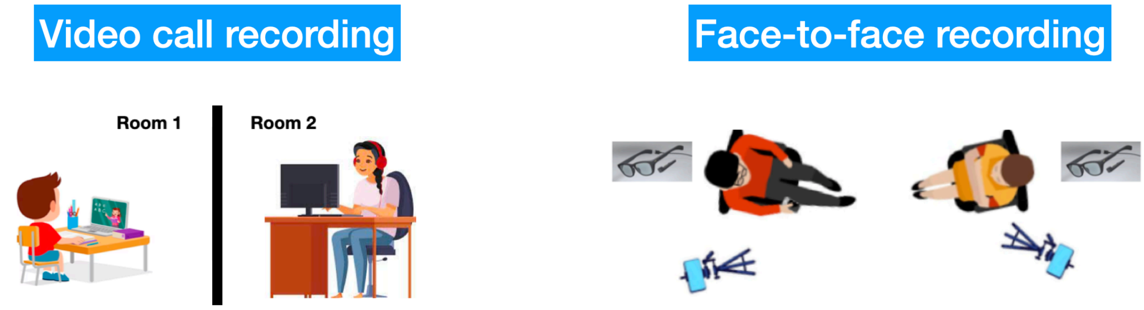

Figure 1: The recording procedure involved 1) video-call recording where the child and caregiver communicated from different rooms at home, and 2) face-to-face recording at home using mobile eye-gaze and head movement detection, in addition to fixed frontal view using cameras on tripods.

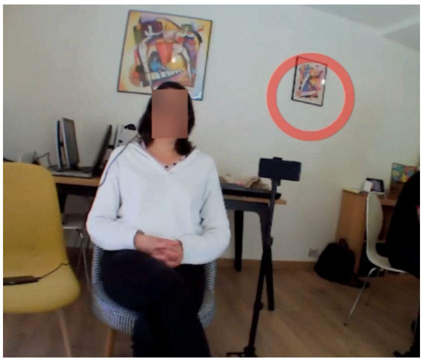

Figure 2: Eye-tracking glasses were calibrated before each use by asking participants to fixate on an object in the house. The gaze marker (the red circle) was then adjusted to match the target object, using real-time streaming of the camera and gaze signal.

however, both speakers were in a sitting position and the cable was long enough to allow freedom of head movements.

Finally, and right before the start of the game, the researcher makes a clap to provide an audio-visual marker to help with later synchronization. The researcher then retreated to a corner (or a different room when possible), telling the participants he would be busy with a different work-related activity – in order to minimize interference or the feeling of being observed by a third person.

3. Data Processing and Annotation

3.1. Synchronization

The Zoom data required no synchronization. As for the face-to-face data, we had – for each dyad – 4 streams of audio-visual recordings, 2 for each participant: (a) the frontal-view recording via the smartphone's camera and (b) the egocentric-view recording via the Pupil invisible device. These sources were synchronized pairwise based on the clap marker. This process was further checked by manually reviewing the resulting videos and verifying that both audio sources were properly synchronized.

3.2. Transcription, Diarization, and Forced Alignment

After piloting a handful of automatic transcription tools in the French Language (e.g., Kaldi, Speech Brain, and Coqui), the software Whisper (Radford et al., 2023) provided the most promising results on our data, including for children. Radford et al. reported a Word Error Rate (WER) of 8.3% for French on the Fleurs dataset, a score that is generally considered good quality.

That said, Whisper has some limitations. In particular, while it produces timestamps, these are for long segments of speech (corresponding more to conversational turn segments). Ideally, we would need a finer alignment at least at the word level, allowing a more precise analysis of how the verbal, vocal, and visual components of speech interact as the utterance unfolds in time. To this end, we used an augmented version of the software called WhisperX by Bain et al. (2023); it provides word-level timestamps using forced phoneme alignment.

The way we did the transcription differed between the video calls and face-to-face conditions. For video calls, Zoom allows recording in a separate audio channel for each speaker and, therefore, bypasses the need for speakers' diarization (i.e., who is speaking and when; a classification that is crucial for accurate transcription of dialog, especially when there is speech overlap between interlocutors). A manual investigation of the transcription confirmed their high quality and low error rate.[4]

[4]Note, however, that while Whisper is good at capturing the semantic content, its transcription systematically

As for the face-to-face recordings, and unlike the zoom data, it was not easy to reliably isolate speech from each interlocutor and create two separate channels. Even though each speaker used a separate microphone (integrated into the eyeglasses), both microphones picked up speech from both interlocutors. The reasons are (i) the child and caregiver generally sat close to each other and (ii) the parents spoke generally louder than the children did. Therefore, we needed a process of diarization. While WhisperX has a module for automatic diarization (based on "pyannote," Bredin et al., 2020), it did not perform satisfactorily on our data; so we resorted to full manual labeling of speakers for each transcribed turn.

The overall transcription of face-to-face data, using only one channel for both interlocutors, was – as one would expect – not as high-quality compared to video-call recordings with two separate channels. In addition, some turns were missing, such as short responses (e.g., "yes" and "no"), especially when overlapping with the interlocutor's speech. Thus, the automatic transcription of face-to-face recording was entirely corrected by hand while watching the recording, including by adding any missing turns.

3.3. Non-verbal behavior

3.3.1. Continuous data

The face-to-face data provides time series for both gaze and head movement using the sensors integrated into the eyeglasses.

- Regarding gaze data, the device has two internal eye cameras, filming the left and right eyes. The device uses a pre-trained machine-learning algorithm to map eye data to a 2D projection on the egocentric field of view (i.e., filmed by the scene camera). The final outcome is a video showing the view of the person, on top of which, the coordinates of their current gaze, fixation patterns, and blinks.
- As for head movement, the gyroscope and accelerometer are used to derive movement properties of roll and pitch (indicating head nods and head shakes, respectively) in addition to translational acceleration.

An important technical question one could ask here concerns the way fixation detection is made during head movement. According to Pupil lab, the device explicitly compensates for the vestibulo-ocular reflex, thus maintaining a stable rendering of the fixation even *during* head movements.

The video-call data: We extracted time-continuous measures of gaze and head nods in an *indirect* fashion. Indirect because the measures were not the outcome of physical sensors worn by interlocutors (as in the case of face-to-face data); but rather, extracted from the videos using computer vision algorithms, namely OpenFace (Baltrusaitis et al., 2018). We extracted gaze coordinates as well as head pose coordinates. To facilitate the ability for future research to detect nods and shakes, we used the raw head coordinates to compute rotation velocities along the x-axis (roll) and y-axis (pitch) across a sliding, fixed time window of 20 frames.

3.3.2. Categorical characterization of non-verbal behavior

The above – continuous – non-verbal data are sufficient for the study of several important aspects of multimodal conversational dynamics, thanks to time series analysis (e.g., Hale et al., 2020). In addition to the continuous characterization, a more categorical analysis can be useful in studying some non-verbal behaviors, especially the ones for which clear time boundaries can be defined (i.e., determining when the behavior begins and when it ends). One example of such behavior is "gazing at interlocutor" vs. "averting gaze;" a categorical signal that plays an important role in regulating multimodal conversational dynamics (Kendon, 1967). Our first step was to attempt and extract these categories fully automatically from both face-to-face data and Video-call data using computer vision tools. However, we realized that manual annotation/checking was still necessary, especially for the video call data. We proceeded as follows.

For face-to-face data, this process consisted of two steps. First, the face of the interlocutor was detected in the egocentric video using RetinaFace, a state-of-the-art computer vision algorithm for face detection (Deng et al., 2020). Second, the gaze coordinate data (overlaid on the pixel space of the egocentric video) were used to determine when gaze fixations intersected with the box.[5]

filtered out several forms of disfluency, e.g., "uh" and "um" (see Radford et al., 2023). Such units can be important when analyzing dialog; their study in our corpus would require inserting them back into the transcript by hand and/or by using specific detection techniques.

[5]Note that we had to double the size of the area of interest (face box) from the original size detected by RetinaFace in order to account for the following two sources of variability. First, the distance between interlocutors varied slightly across dyads, leading to the face appearing slightly smaller or larger in the interlocutor's view. The gaze precision being the same in the video pixel space (see Figure 2), it would tend to capture fewer gaze

For Video call data: There has been research to use recent advances in computer vision tools to categorize looking behavior in video calls (e.g., Erel et al., 2023). Such methods perform relatively well for specific age groups (e.g., infants), using (semi-)experimental settings that constrain and minimize variability between participants. When similar (or even more advanced) computer vision tools are used for naturalistic, unconstrained settings such as ours, they tend to provide good results on average; they are less reliable at the level of a single recording, given the high between-subject variability in naturalistic/unconstrained settings (e.g., Goumri et al., 2023).

Thus, for gaze categorization in video calls, we resorted to full manual annotation. We used the coding scheme defined by Bodur et al. (2023) whose annotated corpus involved children in the same age group as ours. Thus, we categorized gaze into "looking at screen" (a proxy for "gazing at interlocutor") vs. "looking away." (a proxy for "averting gaze").[6] A human annotator first trained on 80 % of the videos and manual annotation of Bodur et al. (2023). Then our annotator used 20% of the remaining video of the CHiCO corpus to estimate inter-rater reliability. Since this required not only assessing agreement on identification (whether a gaze was detected) but also agreement on *segmentation* (start-time and end-time boundaries), we cannot use standard measures like Cohen Kappa. Instead, we used the Staccato algorithm implemented in ELAN (Lücking et al., 2011), which is more adapted to time-related data. We ran the analysis with 1000 Monte Carlo Simulations, a granularity for annotation length of 10, and α= 0.05. The agreement score (known as the degree of organization) was 0.66. After reaching this relatively good agreement score, our annotator then coded all videos in *our* corpus.

fixations for the more distant, smaller-appearing faces. The second source of variability is that, for a few participants, we detected a slight, but systematic miscalibration, leading to their gaze fixation being projected slightly beside the (original) face box. Doubling the size of the area of interest allowed us to better control these artifacts.

[6]Note that, this distinction does not map exactly to "gazing at interlocutor" vs. "averting gaze" in face-to-face conversations, mainly because the position of the webcam is not aligned with the face. However, this is a constraint inherent to most current commercial video call software and people have had to adapt to it. In fact, one of the many goals of the current corpus is precisely to allow future research to investigate how such constraint may influence gaze dynamics in online conversation compared to face-to-face.

4. Analyses

In this section, we provide preliminary analyses to demonstrate the richness of the corpus for studying child-caregiver conversational dynamics across interlocutor (child or caregiver), developmental age in middle childhood (7, 9, 11 years old), modality (e.g., verbal and non-verbal), and medium of communication (i.e., face-to-face vs. video call). The goal is not to provide a thorough scientific study or test specific hypotheses, but rather, to summarize the main characteristics of our data using, mainly, high-level measures.

4.1. Methods

For the verbal modality, we quantified: 1) the number of words uttered, 2) the number of turns taken, 3) the time duration of a word, and 4) the time duration of a turn. For the first measure, we used the – manually corrected – automatic transcription. For the second and fourth measures, we used the turn boundaries segmented automatically with Whisper software. For the third measure, we used estimates for each word's timestamps obtained via the forced alignment module of WhisperX (see Section 3.2).

For the non-verbal modality, we quantified the proportion of "gazing at the interlocutor" vs. "averting gaze" in the conversation, following the methodology outlined in Section 3.3.2, for face-to-face and video calls.

Note that all these measures can a priori be affected by various sources of variability – that are not of a developmental or social nature – as is always the case in naturalistic, largely unconstrained data. Such sources include, e.g., varying conversation lengths, varying pauses within a conversation due to unpredictable events, differences between participants' hardware or software (for video calls), and differences in annotation methods (e.g., gaze annotation using eye-tracking in face-to-face vs. manual coding in video calls). To control for such factors, all measures are calculated relative to the interlocutor in each conversation. More precisely, to obtain a normalized measure for interlocutor A, we simply divide the original estimate of interlocutor A by the sum of this estimate and the estimate from interlocutor B, i.e.,

$$measure_A(normalized) = \frac{measure_A}{measure_A + measure_B}$$

The assumption is that external sources of variability would affect both interlocutors similarly. Thus, with this normalization, our aim is to tap directly into the *dyadic* dynamics and how these dynamics are potentially influenced by developmental age, modality, and medium of communication.

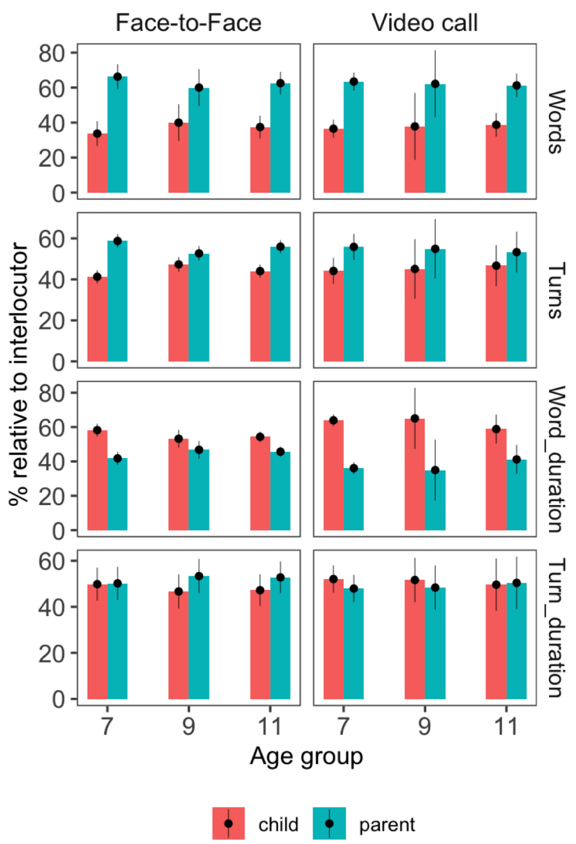

Figure 3: For each measure, we show a percentage; a normalization relative to the interlocutor in the same conversation. Results are broken down by interlocutor (child vs. parent), age (7, 9, or 11 years old), and communicative medium (face-to-face vs. Video call). Dots and ranges indicate the average and 95% confidence intervals.

4.2. Results and Discussion

Figure 3 shows the results for the verbal modality. We can make several observations. First, regarding the total number of words uttered in a conversation, children produced much fewer words; almost half the number of words produced by their parents. Second, regarding the total number of turns in a conversation, children also took fewer turns than the parents – although the difference is not as large as in the case of words. As for the duration of a single turn, children generally took as much time as their parents. Finally, we observe a reliable pattern whereby children's word utterance took more time compared to their parents, suggesting that children generally spoke at a slower rate.

Note that these verbal measures are not totally independent, e.g., the combination of the finding that children produced much fewer words with the finding that the duration of their turns was comparable to that of their adult interlocutors (and the fact that the number of turns itself was only slightly smaller) logically predicts that children would speak at a slower pace, a prediction that was indeed confirmed in by the results of the word duration measure.

Figure 4 shows the results for the verbal modality, more specifically, the proportion of "gazing at the interlocutor" vs. "averting gaze", a binary variable that we further normalized relative to the interlocutor. Figure 4 shows largely similar behavior in children and adults. One can note a slightly higher tendency to gazing at interlocutors for parents in video calls, and a reverse pattern in face-to-face. However, given that the differences are rather small (and the variability large), these patterns are not highly reliable.

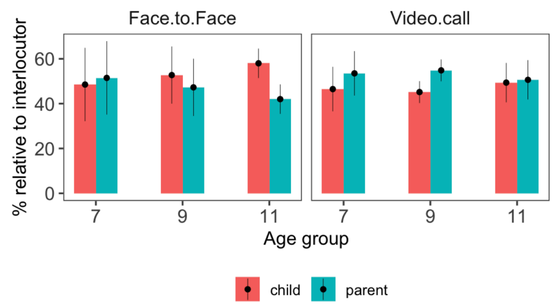

Figure 4: The proportion of gaze at interlocutor (normalized relative to the interlocutor). Dots and ranges indicate the average and 95% confidence intervals.

One important observation, for both verbal and non-verbal measures, is that we did not find any noticeable *developmental change* across our age groups. This, of course, does not mean there is no change in children's conversational skills more generally. Rather, it is likely the consequence of our using of measures that are high-level, underspecified averages, reflecting broad aspects of child-caregiver dynamics that may not undergo change across middle childhood.[7]

Another important observation is the fact that all findings were strikingly similar both in face-to-face and video call settings. This was unequivocally the case for the verbal modality. For the non-verbal modality, although – as we noted earlier – one could notice a slight difference, this difference is not large nor systematic. The overall similarity demonstrates that our high-level measures – reflecting broad distribution/duration in verbal and non-verbal signals – are quite robust across mediums of communication, providing useful baselines

[7] That said, if the "stable" patterns we report here are further validated in experimental, confirmatory studies, they would represent novel and interesting findings.

for future, finer-level comparative analysis in our corpus.

5. Conclusion

This work introduces, to the best of our knowledge, the first developmental corpus of child-caregiver conversations; comparing face-to-face and computer-mediated interactions in middle childhood. A major goal of ours was to capture – to the extent possible – naturally occurring (multimodal) language-in-interactions, therefore we made all recordings at home, using an intuitive – prompt free – elicitation task.

We mitigated the challenges of cost-intensive manual curation of naturalistic corpora by a) capitalizing on recent advances in miniature sensors to automatically detect gaze and head motion while minimizing interference with the spontaneity of the interaction, and b) making use of advances in the automatic coding of both verbal and non-verbal signals.

We provided preliminary analyses (based on half the target sample size that was fully processed) measuring high-level properties such as the distribution/duration of several verbal and non-verbal behaviors. While the scientific significance of the analyses' results should not be overstated (as each of the findings we report would require a more thorough, dedicated investigation), they do demonstrate the richness of the corpus, allowing comparisons by age, modality, and medium of communication. They also provide first steps – and baselines – for future investigations into the intricacies of multimodal communicative development.

6. Extra space for ethical considerations and limitations

6.1. Limitations

This project started with the promise/hope that recent technology in both measurement techniques and automatic processing tools would significantly facilitate the study of child communicative development in ecologically valid contexts. In the end, this technological promise was only partially fulfilled. While some traditional challenges have been largely eased with automatic tools both in the verbal (e.g., speech transcription) and the non-verbal domains (measurement of gaze), others still constitute a bottleneck, especially the ones related to the detection and categorization of head and hand gestures. These are fundamental aspects of face-to-face communication, without which our understanding of development cannot be complete (Bavelas et al., 1992; McNeill, 1992; Kendon, 1994). Unless there is a technological breakthrough, curating these signals in a largely unconstrained and natural context – as ours – will still require major investment in human expertise and annotation.

6.2. Ethical considerations

The data was collected with the approval of our university's Ethics Committee and was registered by the Data Protection Officer before starting the project. None of our recording or measuring devices involve any type of clinical intervention and are fully non-invasive. Minors' consent was formally given by their guardians/caregivers. Caregivers informed children in age-appropriate language. They could stop the recording at any moment without having to provide a reason.

All steps regarding corpus storage and sharing are in strict compliance with the local laws of the country as well as the European regulations (GDPR), to ensure full protection of children's anonymity. For instance, transcripts, annotations, and any derived data will be anonymized before sharing. Private access to raw videos will be possible by other researchers via means that are compliant with GDPR and the laws of the country.

7. Acknowledgements

This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), 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. Finally, we thank all families that have volunteered to participate in data collection.

8. Bibliographical References

Kirsten Abbot-Smith, Julie Dockrell, Alexandra Sturrock, Danielle Matthews, and Charlotte Wilson. 2023. Topic maintenance in social conversation: What children need to learn and evidence this can be taught. *First Language*.

Abhishek Agrawal, Jing Liu, Kübra Bodur, Benoit Favre, and Abdellah Fourtassi. 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*, volume 45.

Anne Anderson, Henry S. Thompson, , Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The HCRC map task corpus: Natural dialogue for speech recognition. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, page 25–30, USA. Association for Computational Linguistics.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Ed Baines and Christine Howe. 2010. Discourse topic management and discussion skills in middle childhood: The effects of age and task. *First Language*, 30(3-4):508–534.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Janet B Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941.

Janet Beavin Bavelas, Nicole Chovil, Douglas A. Lawrie, and Allan Wade. 1992. Interactive gestures. *Interactive gestures. Discourse processes*, 15(4):469–489.

Lois Bloom, Lorraine Rocissano, and Lois Hood. 1976. Adult-child discourse: Developmental interaction between information processing and linguistic knowledge. *Cognitive Psychology*, 8(4):521–552.

Kübra Bodur, Mitja Nikolaus, Fatima Kassim, Laurent Prévot, and Abdellah Fourtassi. 2021. Chico: A multimodal corpus for the study of child conversation. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 158–163.

Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. 2023. Using video calls to study children's conversational development: The case of backchannel signaling. *Frontiers in Computer Science*, 5.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

Marisa Casillas, Susan C Bobb, and Eve V Clark. 2016. Turn-taking, timing, and planning in early language acquisition. *J. Child Lang.*, 43(6):1310–1337.

Adriana Chee Jing Chieng, Camille J. Wynn, Tze Peng Wong, Tyson S Barrett, and Stephanie A. Borrie. 2024. Lexical alignment is pervasive across contexts in non-weird adult–child interactions. *Cognitive Science*, 48(3):e13417.

Eve V Clark. 2018. Conversation and language acquisition: A pragmatic approach. *Language Learning and Development*, 14(3):170–185.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, New York, NY, US.

Martina De Lillo, Rebecca Foley, Matthew C Fysh, Aimée Stimson, Elisabeth EF Bradford, Camilla Woodrow-Hill, and Heather J Ferguson. 2021. Tracking developmental differences in real-world social attention across adolescence, young adulthood and older adulthood. *Nature human behaviour*, 5(10):1381–1390.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212.

Christina Dideriksen, Morten H Christiansen, Kristian Tylén, Mark Dingemanse, and Riccardo Fusaroli. 2023. Quantifying the interplay of conversational devices in building mutual understanding. *Journal of Experimental Psychology: General*, 152(3):864.

Mark Dingemanse. 2024. Interjections at the heart of language. *Annual Review of Linguistics*, 10:257–277.

Mark Dingemanse and N J Enfield. 2024. Interactive repair and the foundations of language. *Trends Cogn. Sci.*, 28(1):30–42.

Seamus Donnelly and Evan Kidd. 2021. The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*, 92(2):609–625.

Steven L. Elmlinger, Michael H. Goldstein, and Marisa Casillas. 2023. Immature vocalizations simplify the speech of tseltal mayan and u.s. caregivers. *Topics in Cognitive Science*, 15(2):315–328.

Yotam Erel, Katherine Adams Shannon, Junyi Chu, Kim Scott, Melissa Kline Struhl, Peng Cao, Xincheng Tan, Peter Hart, Gal Raz, Sabrina Piccolo, et al. 2023. icatcher+: Robust and automated annotation of infants' and young children's gaze behavior from videos collected in laboratory, field, and online studies. *Advances in Methods and Practices in Psychological Science*, 6(2):25152459221147250.

Riccardo Fusaroli, Ethan Weed, Roberta Rocca, Deborah Fein, and Letitia Naigles. 2023. Caregiver linguistic alignment to autistic and typically developing children: A natural language processing approach illuminates the interactive components of language development. *Cognition*, 236:105422.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Dhia-Elhak Goumri, Thomas Janssoone, Leonor Becerra-Bonache, and Abdellah Fourtassi. 2023. Automatic detection of gaze and smile in children's video calls. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, page 383–388.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Joanna Hale, Jamie A Ward, Francesco Buccheri, Dominic Oliver, and Antonia F de C Hamilton. 2020. Are you on my wavelength? interpersonal coordination in dyadic conversations. *Journal of nonverbal behavior*, 44:63–83.

Lucille Hess and Judith Johnston. 1988. Acquisition of backchannel listener responses to adequate messages. *Communication Sciences and Disorders Faculty Publications*, 11:319–335.

Elinor Ochs Keenan and Ewan Klein. 1975. Coherency in children's discourse. *Journal of Psycholinguistic Research*, 4(4):365–380.

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

Adam Kendon. 1994. Do gestures communicate? a review. *Research on language and social interaction*, 27(3):175–200.

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

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

Jing Liu, Mitja Nikolaus, Kübra Bodur, and Abdellah Fourtassi. 2022. Predicting backchannel signaling in child-caregiver multimodal conversations. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, ICMI '22 Companion, page 196–200.

Andy Lücking, Sebastian Ptock, and Kirsten Bergmann. 2011. Assessing agreement on segmentations by means of staccato, the segmentation agreement calculator according to thomann. In *International gesture workshop*, pages 129–138. Springer.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Barbara Maroni, Augusto Gnisci, and Clotilde Pontecorvo. 2008. Turn-taking in classroom interactions: Overlapping, interruptions and pauses in primary school. *European journal of psychology of education*, 23:59–76.

Lillian R Masek, Brianna TM McMillan, Sarah J Paterson, Catherine S Tamis-LeMonda, Roberta Michnick Golinkoff, and Kathy Hirsh-Pasek. 2021. Where language meets attention: How contingent interactions promote learning. *Developmental Review*, 60:100961.

Danielle Matthews, Hannah Biney, and Kirsten Abbot-Smith. 2018. Individual differences in children's pragmatic ability: A review of associations with formal language, social cognition, and executive functions. *Language Learning and Development*, 14(3):186–223.

Chiara Mazzocconi, Benjamin O'Brien, Kevin El Haddad, Kübra Bodur, and Abdellah Fourtassi. 2023. Differences between mimicking and non-mimicking laughter in child-caregiver conversation: A distributional and acoustic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

David McNeill. 1992. *Hand and mind: What gestures reveal about thought.* University of Chicago press.

Thomas Misiek, Benoit Favre, and Abdellah Fourtassi. 2020. Development of Multi-level Linguistic Alignment in Child-adult Conversations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 54–58, Online. Association for Computational Linguistics.

Thomas Misiek and Abdellah Fourtassi. 2022. Caregivers exaggerate their lexical alignment to young children across several cultures. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.

Vivian Nguyen, Otto Versyp, Christopher Cox, and Riccardo Fusaroli. 2022. A systematic review and bayesian meta-analysis of the development of turn taking in adult–child vocal interactions. *Child Development*, 93(4):1181–1200.

Mitja Nikolaus and Abdellah Fourtassi. 2023. Communicative feedback in language acquisition. *New Ideas in Psychology*, 68:100985.

Martin J. Pickering and Simon Garrod. 2021. *Understanding Dialogue: Language Use and Social Interaction*, 1 edition. Cambridge University Press.

Eli Pincus and David Traum. 2016. Towards automatic identification of effective clues for team word-guessing games. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2741–2747, Portorož, Slovenia. European Language Resources Association (ELRA).

Matthew Richard John Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Evan F Risko, Daniel C Richardson, and Alan Kingstone. 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.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Emanuel Schegloff. 1991. Conversation analysis and socially shared cognition. *Socially shared cognition*.

Emanuel A Schegloff. 1996. Issues of relevance for discourse analysis: Contingency in action, interaction and co-participant context. In *Computational and conversational discourse: Burning issues—An interdisciplinary account*, pages 3–35. Springer.

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

Sara Sehley and Catherine Snow. 1992. The conversational skills of school-aged children. *Social Development*, 1(1):18–35.

Jinyu Shi, Yan Gu, and Gabriella Vigliocco. 2023. Prosodic modulations in child-directed language and their impact on word learning. *Developmental Science*, 26(4):e13357.

Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C Frank. 2021. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, 5:20–29.

Marc Tonsen, Chris Kay Baumann, and Kai Dierkes. 2020. A high-level description and performance evaluation of pupil invisible. *arXiv preprint arXiv:2009.00508*.

Kristin J Van Engen, Melissa Baese-Berk, Rachel E Baker, Arim Choi, Midam Kim, and Ann R Bradlow. 2010. The wildcat corpus of native-and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and speech*, 53(4):510–540.

Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511.

Zhenlin Wang, Rory T Devine, Keri K Wong, and Claire Hughes. 2016. Theory of mind and executive function during middle childhood across cultures. *Journal of Experimental Child Psychology*, 149:6–22.