ChiCo: A Multimodal Corpus for the Study of Child Conversation

KÜBRA BODUR, MITJA NIKOLAUS, FATIMA KASSIM, LAURENT PRÉVOT, and ABDELLAH FOURTASSI, Aix-Marseille Université, CNRS, LIS, LPL, ILCB, France

The study of how children develop their conversational skills is an important scientific frontier at the crossroad of social, cognitive, and linguistic development with important applications in health, education, and child-oriented AI. While recent advances in machine learning techniques allow us to develop formal theories of conversational development in real-life contexts, progress has been slowed down by the lack of corpora that both approximate naturalistic interaction and provide clear access to children's non-verbal behavior in face-to-face conversations. This work is an effort to fill this gap. We introduce ChiCo (for Child Conversation), a corpus we built using an online video chat system. Using a weakly structured task (a word-guessing game), we recorded 20 conversations involving either children in middle childhood (i.e., 6 to 12 years old) interacting with their caregivers (condition of interest) or the same caregivers interacting with other adults (a control condition), resulting in 40 individual recordings. Our annotation of these videos has shown that the frequency of children's use of gaze, gesture and facial expressions mirrors that of adults. Future modeling research can capitalize on this rich behavioral data to study how both verbal and non-verbal cues contribute to the development of conversational coordination.

CCS Concepts: • **Applied computing** → **Psychology**.

Additional Key Words and Phrases: corpus, multimodal, conversation, non-verbal communication, child development

ACM Reference Format:
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 (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3461615.3485399

INTRODUCTION

Conversation is a ubiquitous and important activity in our lives. Cognitive scientists consider it as a hallmark of human cognition as it relies on a sophisticated ability for coordination and shared attention [13, 27] while prominent computer scientists have described it as the ultimate test for Artificial Intelligence [28]. Its role for healthy development is also crucial: When conversational skills are not well developed, they can negatively impact our ability to learn from others and to maintain relationships [20]. Thus, the scientific study of how conversational skills *develop* in childhood is of utmost importance to understand what makes human cognition so special, to design better child-oriented conversational AI, and to allow more targeted and efficient clinical interventions (e.g., for individuals with autism).

Conversation involves a variety of skills such as turn-taking management, negotiating shared understanding with the interlocutor, and the ability for a coherent/contingent exchange [6, 23]. We know little about how these skills develop in face-to-face conversations, mainly because the latter has two characteristics that has made it difficult to study using traditional research methods in developmental psychology. First, conversation is inherently spontaneous, i.e., it cannot be scripted, making it crucial to study in its *natural* environment. Second, conversation involves collaborative *multimodal* signaling, e.g. gesture, eye gaze, facial expressions, as well as intonation and linguistic content [14]. While controlled in-lab studies have generally fallen short of the ecological validity (that is, the first characteristic), observational studies

Manuscript submitted to ACM

in the wild (typically in the context of child-caregiver interaction) have fallen short of accounting for the multimodal dynamics, especially the role of non-verbal communicative abilities that are crucial for a coordinated conversation.

Recent technological advances in Natural Language Processing and Computer Vision allow us, in theory, to go beyond the limitations of traditional research methods as they provide tools to study complex multimodal dynamics while scaling up to more naturalistic data. Nevertheless, work in this direction has been slowed down by the lack of annotated corpora of child conversations that provide clear access to the interlocutors' facial expression and communicative gestures. The current work is a step towards filling this gap.

The current work

The novelty of the current effort compared to previous work based on available child-caregiver conversational data (e.g., [15]) is threefold:

1) We focus on middle childhood (i.e., 6 to 12 years old), an age range that has received little attention compared to the preschool period, although middle childhood witnesses important developmental changes in conversational skills, e.g., in turn-taking management [9], conversational grounding (e.g., backchannels, [10]), and the ability to engage in coherent/contingent exchange [3, 8]. Another — perhaps more pragmatic — reason we focus on middle childhood is that children in this period have already mastered much of formal language (e.g, phonology and syntax), allowing us to minimize the interference of language processing- and production-related issues with the measurement of "pure" conversational skills.

2) We aim at capturing the *multimodal* aspects of conversational development, especially facial expressions, gaze, and gestures which are crucial to face-to-face conversations. While much of previous work focused on the verbal or vocal signal [5, 24, 30], a comprehensive study of conversational development requires that we also investigate how children learn to coordinate with the interlocutor using visual cues. This learning includes both understanding the interlocutor's visual signal in order to plan a move (e.g., engaging in conversational repair after noticing that the interlocutors has frowned) and producing contingent visual cues (e.g., head nodding or smiling as an accommodating backchannel signal).

3) We use a novel conversational task (a word-guessing game) that is very intuitive and weakly structured, allowing conversations to flow spontaneously. The goal being to approximate a conversational activity that children could do with their caregivers at home, eliciting as much as possible the children's natural conversational skills. Unlike other — more classic — semi-structured tasks used in the study of adult-adult conversations (e.g., the map task, [2]), the word-guessing game does not require looking at a prompt (e.g., a map) thus optimizing children's non-verbal signaling behavior while interacting with the caregiver.

The paper is organized into three parts: Corpus, Annotation, and Analysis. In the first part, we present the methodology for data collection. In the second, we present the scheme we used to annotate the video recording for several — potentially communicative — visual features as well as the inter-rater reliability for each feature. In the last part, we present analyses (based on our annotation) comparing the frequency of non-verbal cues in child-caregiver conversations to that of adult-adult conversations.

1 THE CORPUS

The Corpus consists of video recordings where dyads of interlocutors engage in weakly structured conversations (a word-guessing game). Each dyad of child-caregiver was paired with a "control dyad" involving the same caregiver

and another adult. In what follows, we provide information about the recording setup, the task we used to elicit conversational data, the details of the recording procedure, and the participants.

1.1 Recording setup

We chose to do the recording via an online video chat system (Zoom). The primary reason for this choice is that this setting provides clearer data of facial expressions and head gestures (than, say, a third-person-view camera), making it easier to annotate manually and, perhaps, automatize these annotations using computer vision algorithms. Other advantages of such a setting include facilitating the recruitment process (especially during the Covid-19 pandemic), increasing the ecological validity of the data thanks to recording from the children's home (instead of the less familiar environment of a lab), and providing ideal visual training data for child-oriented conversational AI that have access to similar input through a webcam.

The setup requires that the caregiver and child use different devices (e.g., two laptops or a laptop and a tablet) and that they communicate from different rooms (if they record from the same house) in order to avoid issues due to echo. We also required that the caregiver wore a headset microphone during the recording for better audio quality.[1]

1.2 Tasks

The task consists of playing a guessing game in which one of the participants thinks of a word and the other tries to find it by asking questions. After a word has been guessed, the interlocutors alternate their roles. The task ends after about 10 minutes. The caregivers were provided with a list of words to use during the interaction with the child whereas the children were free to choose the words they wanted. Detailed task instructions can be found at https://osf.io/5ngwc/.

1.3 Procedure

We recorded a three-way call involving 1) the experimenter (doing the recording), 2) the caregiver, and 3) either the child (in the condition of interest) or another adult (in the control condition). The experimenter was muted during the entire interaction and used a black profile picture in order not to distract the participants. The experimenter was able to record both interlocutors by pinning their profiles side-by-side on her local machine (Figure 1). Furthermore, the participants were instructed to hide self-view and to pin only their interlocutors.

The procedure consists of three stages (all were included in the recordings). First, the caregiver explains the task to the child, then the pair does the task for around 10 minutes, and lastly, the caregiver initiates a free conversation with the child (or adult) to chat about how the task went (about 5 minutes).

1.4 Participants

We recorded 20 conversations. Among these conversations, 10 involved children and their caregivers (the condition of interest) and 10 involved the caregivers with other adults (the control condition). In total, we collected 40 individual videos (i.e., of individual interlocutors). Each video lasted around 15 minutes for a total of 5 hours and 49 minutes across both conditions. The children were aged 6 to 12 years old (M=8.5, SD=1.37). All 10 children were native French speakers, 5 of them were reported to be bilinguals, also speaking English/Portuguese/Spanish/Czech in addition to French. Children did not have any communicative or developmental disorders expect for one child who had a mild form of autism. All of the 9 caregivers were one of the parents and one parent participated twice as he had two kids.

[1]If neither of the interlocutors wears a headphone, the video chat system tends to automatically cut the sound of one speaker when the speakers happen to talk over each other, which is an undesirable feature for the purpose of our data collection.

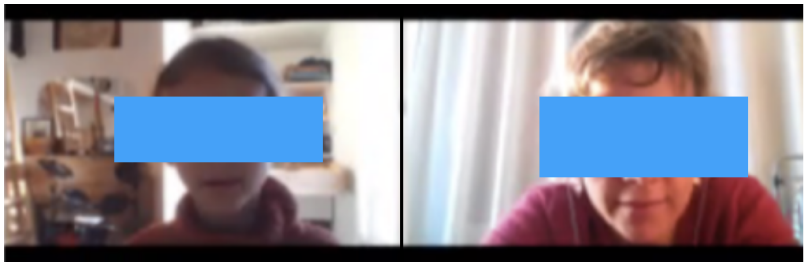

Fig. 1. A snapshot of one of the recording sessions involving a child and her caregiver communicating through an online video chat system.

2 ANNOTATIONS

We manually annotated the entire 40 individual recordings in our corpus for gaze, facial expressions and gestures of the participant. The annotation was performed using a custom developed template on ELAN [31].

2.1 Annotated features

The annotated features were: gaze direction (looking at interlocutor vs. looking away), head movements (head nod and head shake), eyebrow movements (frowning and eyebrow raising), mouth displays (smiles and laughter) and posture change (leaning froward or backward). In what follows, we elaborate on each of these dimensions.

2.1.1 Gaze Direction. Previous studies [11] suggest that eye gaze can be used to control the flow of conversation, e.g., by regulating the turn-taking behavior: Speakers tend to look away from their partners when they begin talking and look back at them when they are about to finish their conversational turn. Gaze direction can also serve as a predictor of gestural feedback [19], signaling to the listener that the speaker is waiting for feedback. Thus, we annotated the occasions where the target participant was looking directly at the screen, which we took as an indication that the gaze was directed at the interlocutor.

2.1.2 Head Movements. Head movements have various functions in human face-to-face communication such as giving feedback to — and eliciting feedback from — the interlocutor [21, 25]. We annotated two head movements that may play different communicative roles: head nods and head shakes.

2.1.3 Eyebrow Displays. Eyebrow movements are among the most commonly employed facial expressions, they can communicate surprise, anger and confusion. In our annotation scheme, we have two categories for eyebrow movements: raised (lifted eyebrows) or frown (the contraction of eyebrows and movement towards the nose).

2.1.4 Mouth Displays. Smiles and laughs are a common source of back-channel communication between participants. It can, e.g., perpetuate thoughts of being understood [4] or signal attention when combined with a head nod [7]. We annotated two categories of mouth display: Smile and laughter.

2.1.5 Posture. Posture also plays a communicative role. In particular, leaning forward can indicate attention and/or a positive feedback [22] and leaning backward may occur as a turn yielding signal [1] . Here we annotated posture by taking the starting point as the neutral position and any movement from there was tagged either forward or backward.

2.2 Annotation method and inter-rater reliability

The annotation involved: 1) Detecting that a target non-verbal event has occurred during the time course of the conversation, 2) Once the event was detected, tagging its time interval (i.e., when it began and ended) and, 3) tagging its category (e.g., Smile, head nod, or leaning backward). [2]

The entire corpus was annotated by the first author. In order to estimate inter-rater reliability, a subset of 8 recordings (i.e., 20% of the data) including 4 children and 4 for adults were annotated independently by the third author in a second time. Estimating the degree of agreement between annotators for time-dependent events is not an easy task as it requires comparing not only agreement on the classification (whether an event is a laugh or a smile) but also agreement on the time segmentation of this event.

Recently, [18] introduced a holistic measure, γ, that takes into account both classification and segmentation, accounting for some phenomena that are not well captured by other existing methods such as Krippendorff's α_u [12]. Such phenomena include when segments overlap in time (e.g., when an interlocutor nods and smiles at the same time) which occur frequently in face-to-face data such as ours. Thus, it is the γ measure that we report in the current study using the Python implementation by [26]. In a nutshell, this method estimates agreement by finding the optimal alignment in time between segments obtained from different annotators. The alignment minimizes both dissimilarity in time correspondence between the segments as well as dissimilarity in their categorization. The optimal alignment is characterized by a disorder value: δ. The measure γ takes into account chance by sampling N random annotations and computing their average disorder: δ_{random}. Finally, the chance-adjusted γ measure is computed as: $\gamma = 1 - \frac{\delta}{\delta_{random}}$. The closer the value to 1, the stronger the agreement between the annotators.[3]

While γ allows us to obtain a global score for the annotation. We were also interested in the finer-grained analysis of agreement comparing segmentation vs. categorization for each annotated feature. To obtain scores for categorization only, we use the γ_{cat} measure introduced by [17] which is computed after one finds the optimal alignment using the original γ as described above. To analyze agreement in segmentation, we computed γ for each category separately, thus eliminating all ambiguities in terms of categorization and letting γ deal with agreement in segmentation only.

Regarding agreement in our data, we found an average $\gamma_{av} = 0.56[0.49, 0.66]$ for children and an average $\gamma_{av} = 0.65[0.59, 0.78]$ for adults, where ranges correspond to the lowest and highest γ obtained in the four videos we double-annotated in each age group. How to interpret these numbers? Since γ is a new measure, it has not been thoroughly benchmarked with data similar to ours, as compared to more known measures (which, however, are less adequate for our data) such as Cohen's Kappa or Krippendorff's alpha. That said, γ is generally more conservative than other existing measures [18]. In addition, γ tends to (over-)penalize several aspects of disagreement in segmentation (see below). Thus, we can consider that the global scores obtained above reflect good overall agreement.

For finer analyses, Table 1 shows the detailed scores for each feature. We found that the scores for categorization were very high across all features and in both age groups, indicating that our non-verbal features are highly distinguishable from each other. The scores for segmentation, though overall quite decent (they are overall larger than 0.5 for all features), are much lower compared to categorization. Segmentation is harder because there are several ways disagreement can occur and for which it is penalized by γ. Part of this disagreement is quite relevant and should indeed be penalized such as when only one annotator detects a given event at a given time or when there is a mismatch between annotators in

[2]Note that similar visual events may vary in their precise communicative functions. For example, a head nod can be used either as specific feedback (e.g., as a non-verbal answer to a yes-no question), or as a generic feedback, indicating attention/understanding without necessarily signaling agreement. Such distinctions are beyond the scope of the current study. Here we annotated the above-described features regardless of their precise communicative role in the conversation.

[3]N is chosen so that γ has a default precision level of 2%, we did not change this default precision in the current work.

Table 1. Average gamma scores quantifying inter-rater reliability between two annotators using 20% of the corpus. Ranges indicate lowest and largest gamma in the videos annotated in each age group.

	Children		Adults	
Features	Categorization	Segmentation	Categorization	Segmentation
Gaze	0.93 [0.85, 0.99]	0.68 [0.63, 0.73]	0.98 [0.94, 1.00]	0.76 [0.61, 0.88]
Mouth_Smile	0.84 [0.66, 1.00]	0.55 [0.32, 0.75]	0.96 [0.94, 1.00]	0.58 [0.42, 0.70]
Mouth_Laugh	0.81 [0.58, 1.00]	0.67 [0.49, 0.86]	0.99 [0.94, 1.00]	0.79 [0.64, 0.87]
Head_Shake	0.99 [0.94, 1.00]	0.69 [0.39, 0.89]	0.94 [0.87, 1.00]	0.71 [0.48, 0.83]
Head_Nod	0.86 [0.65, 1.00]	0.57 [0.47, 0.78]	1.00 [1.10, 1.00]	0.57 [0.46, 0.68]
Posture_Forward	0.81 [0.67, 1.00]	0.50 [0.33, 0.80]	0.90 [0.79, 1.00]	0.63 [0.49, 0.88]
Posture_Backward	0.86 [0.74, 0.94]	0.52 [0.33, 0.68]	0.94 [0.83, 1.00]	0.67 [0.46, 0.91]
Eyebrow_Raised	0.82 [0.77, 0.94]	0.50 [0.43, 0.56]	0.92 [0.88, 0.97]	0.66 [0.57, 0.77]
Eyebrow_Frown	0.79 [0.71, 0.86]	0.52 [0.37, 0.68]	0.66 [0.47, 0.77]	0.49 [0.45, 0.53]

terms of the start and/or the end of the event. However, some aspects of disagreement need not necessarily be penalized such as when one annotator considers that an event is better characterized as a long continuous segment and the other annotator considers, instead, that it is better characterized as a sequence of smaller segments (e.g., a long continuous segment involving multiple consecutive nods *versus* several consecutive but discrete segments representing, each, a single nod).

We found some features to be less ambiguous (to human annotators) than others, especially concerning their segmentation. For example, gaze switch, laughter, and head shake were among the least ambiguous for both children and adults. Posture change and eyebrow movement were the most ambiguous in children. Smiles and head nods had an intermediate level of ambiguity in both age groups. Finally, we noted that the agreement scores were overall higher for adults compared to children, indicating that adult's non-verbal behavior is generally less ambiguous than children's.

3 ANALYSIS

Based on our annotation, we quantified non-verbal communication in child-caregiver multimodal conversation compared to the control condition of adult-caregiver conversation. Figure 2 shows the average number of a target non-verbal behavior (e.g., gaze switch, head nod, or smile) per minute in both conditions and for each speaker. We can observe that the frequency distribution is strikingly similar across all speakers. This finding suggests that non-verbal behavior data is quite balanced across children and adults. Thus, our corpus provides a good basis for future analysis of *how* these cues are used and combined by children to manage conversations and for comparison with how adults' behave in similar conversational contexts.

Another important observation is that non-verbal features are not equally frequent. For example, participants switch gaze to — or away from — interlocutor more than 5 time per minute on average. Other relatively frequent gestures include smiles and eyebrow raising, which occur around twice per minute on average. Most other gestures are less frequent, occurring on average once (or less) per minute including head nods, frowns, and posture change.

CONCLUSION

In this paper, we introduced a corpus of child-caregiver conversation (paired with a "control" corpus of adult-caregiver conversations). We used an online video chat platform to collect the data, allowing us to have a clear access to several

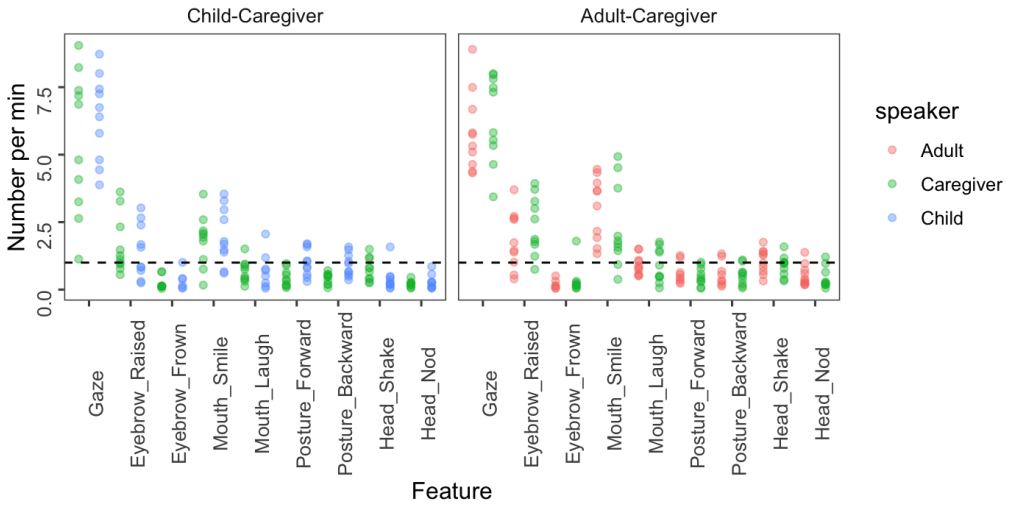

Fig. 2. The frequency of use of non-verbal features for each participant in our corpus. The frequency was normalized by the total length (in minutes) of the conversation. The black dotted line represents a frequency of 1 per minute; data above this line can be understood as indicating relatively frequent behavior.

visual features such as gaze, facial expressions, head movement, and posture change. We used a new semi-structured task (a word-guessing game) which aimed both at approximating natural/spontaneous talk and at optimizing multimodal communicative behavior.

Our hand annotation of the corpus revealed that children in middle childhood use non-verbal cues in conversation almost as frequently as adults do. Future research can, thus, capitalize on this rich behavioral data to study how verbal and non-verbal cues contribute the development of conversational skills.

Another goal of the corpus is to contribute to improving child-oriented conversational AI for applications in, e.g., education and health [16, 29]. Such applications often need to automatically detect relevant non-verbal behavior in children. The current work suggests that these behaviors are not equally clearly detectable even by human annotators as witnessed by the heterogeneity in terms the segmentation's inter-rater reliability scores. We speculate that the features that were more ambiguous to humans (e.g., eyebrow movement) would be the more challenging to (automatically) detect in conversational AI systems.

In future work, we plan to annotate the corpus for multimodal cues beyond visual features, including vocal features (e.g., intonation), verbal content (i.e., transcription) and dialog acts. This would provide us with more complete data to study how children develop their ability to integrate and use various sources of information from different modalities to manage conversations. Another direction for future research is to expand data collection to a larger, more representative sample of children (including from different cultures) capitalizing on our cost-effective procedure based on the online video chat system.

REFERENCES

[1] Jens Allwood, Loredana Cerrato, Laila Dybkjaer, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2005. The MUMIN multimodal coding scheme. *NorFA yearbook* 2005 (2005), 129–157.

[2] 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* (Princeton, New Jersey) *(HLT '93)*. Association for Computational Linguistics, USA, 25–30. https://doi.org/10.3115/1075671.1075677

[3] 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 (2010), 508–534. https://doi.org/10.1177/0142723710370538

[4] Lawrence J Brunner. 1979. Smiles can be back channels. *Journal of Personality and Social Psychology* 37, 5 (1979), 728.

[5] Eve V. Clark. 2018. Conversation and Language Acquisition: A Pragmatic Approach. *Language Learning and Development* 14, 3 (2018), 170–185. https://doi.org/10.1080/15475441.2017.1340843

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

[7] Allen T Dittmann and Lynn G Llewellyn. 1968. Relationship between vocalizations and head nods as listener responses. *Journal of personality and social psychology* 9, 1 (1968), 79.

[8] Bruce Dorval and Carol O. Eckerman. 1984. Developmental trends in the quality of conversation achieved by small groups of acquainted peers. *Monographs of the Society for Research in Child Development* 49, 2 (1984), 1–72.

[9] C. Garvey and G. Berninger. 1981. Timing and turn taking in children's conversations 1. *Discourse Processes* 4 (1981), 27–57.

[10] Lucille Hess and Judith Johnston. 1988. Acquisition of Back Channel Listener Responses to Adequate Messages. *Communication Sciences and Disorders Faculty Publications* 11 (July 1988), 319–335. https://doi.org/10.1080/01638538809544706

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

[12] Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity* 50, 6 (2016), 2347–2364.

[13] Kevin Laland and Amanda Seed. 2021. Understanding Human Cognitive Uniqueness. *Annual Review of Psychology* 72, 1 (2021), 689–716. https://doi.org/10.1146/annurev-psych-062220-051256

[14] Stephen C Levinson and Judith Holler. 2014. The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369, 1651 (2014), 20130302.

[15] Brian MacWhinney. 2000. *The CHILDES project: The database*. Vol. 2. Psychology Press, US.

[16] Michael Madaio, Kun Peng, Amy Ogan, and Justine Cassell. 2018. A climate of support: a process-oriented analysis of the impact of rapport on peer tutoring. International Society of the Learning Sciences, Inc.[ISLS].

[17] Yann Mathet. 2017. The Agreement Measure γ cat a Complement to γ Focused on Categorization of a Continuum. *Computational Linguistics* 43, 3 (2017), 661–681.

[18] Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics* 41, 3 (2015), 437–479.

[19] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20, 1 (2010), 70–84.

[20] Suzanne Murphy, Dorothy Faulkner, and Laura Farley. 2013. The Behaviour of Young Children with Social Communication Disorders During Dyadic Interaction with Peers. *Journal of abnormal child psychology* 42 (June 2013). https://doi.org/10.1007/s10802-013-9772-6

[21] Patrizia Paggio and Costanza Navarretta. 2013. Head movements, facial expressions and feedback in conversations: empirical evidence from Danish multimodal data. *Journal on Multimodal User Interfaces* 7, 1 (2013), 29–37.

[22] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. 2017. Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling Proceedings of the 2017 ACM. In *IEEE International Conference on Human-Robot Interaction. ACM*. IEEE, Singapore, 2308–2314.

[23] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50, 4 (1974), 696–735. http://www.jstor.org/stable/412243

[24] Catherine E Snow. 1977. The development of conversation between mothers and babies. *Journal of child language* 4, 1 (1977), 1–22.

[25] Loredana Sundberg Cerrato. 2007. *Investigating communicative feedback phenomena across languages and modalities*. Ph. D. Dissertation. KTH.

[26] Hadrien Titeux and Rachid Riad. 2021. pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in Python. *Journal of Open Source Software* 6, 62 (2021), 2989. https://doi.org/10.21105/joss.02989

[27] Michael Tomasello. 1999. *The Cultural Origins of Human Cognition*. Harvard University Press, US.

[28] Alan Mathison Turing. 1950. I.—Computing Machinery and Intelligence. *Mind* LIX, 236 (October 1950), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[29] Wayne Ward, Ron Cole, Daniel Bolaños, Cindy Buchenroth-Martin, Edward Svirsky, and Tim Weston. 2013. My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology* 105, 4 (2013), 1115.

[30] Anne S Warlaumont, Jeffrey A Richards, Jill Gilkerson, and D Kimbrough Oller. 2014. A social feedback loop for speech development and its reduction in autism. *Psychological science* 25, 7 (2014), 1314–1324.

[31] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1556–1559.