Development of Multimodal Turn Coordination in Conversations: Evidence for Adult-like behavior in Middle Childhood

Abhishek Agrawal, Jing Liu, Kübra Bodur, Benoit Favre, Abdellah Fourtassi

Development of Multimodal Turn Coordination in Conversations: Evidence for Adult-like behavior in Middle Childhood

Abhishek Agrawal[1] (abhishek-amit.agrawal@univ-amu.fr)
Jing Liu[2] (ecnucrystal@gmail.com)
Kubra Bodur[3] (kubra.bodur@univ-amu.fr)
Benoit Favre[1] (benoit.favre@univ-amu.fr)
Abdellah Fourtassi[1] (abdellah.fourtassi@univ-amu.fr)
[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[2]KU Leuven, Leuven, Belgium
[3]Aix-Marseille Université, LPL, Aix-en-Provence, France

Abstract

The question of how children develop multimodal coordination skills to engage in meaningful face-to-face conversations is crucial for our broader understanding of children's healthy socio-cognitive development. Here we focus on investigating the ability of school-age children to coordinate turns with their interlocutors, especially regarding when to take the floor (i.e., the main channel of the conversation) and when to provide attentive listening signals via the back channel. Using data of child-caregiver naturalistic conversations and data-driven research tools, we found that children aged 6 to 12 years old already show adult-like behavior both in terms of reacting to the relevant channel-specific cues and in terms of providing reliable, multimodal inviting cues to help their interlocutor select the most appropriate channel of the conversation.

Keywords: backchannel; turn-taking; child development; computational modeling

Introduction

To become a competent conversational partner, a child must learn to coordinate the timing and nature of their turn in the dialog. This is a complex task since the child must learn, among other things, (i) when it is a good time to take the floor and become the speaker and (ii) when it is more appropriate to provide non-intrusive feedback while remaining in the role of the listener. In more technical terms, children must learn when to use the *main channel*, i.e., taking or yielding the floor (hereafter, MC), and when to use the *back channel*, e.g., signaling attentive listening using verbal or non-verbal signals like "okay" or a head nod (hereafter BC) (Yngve, 1970). To illustrate, here is an example of a child using the MC:

- Interlocutor: Did you like your food?
- Child: Yes!
- Interlocutor: Nice! I am glad you did!

and an example of the child using the BC:

- Interlocutor: First, we are going to have lunch..
- Child: [head nod]
- Interlocutor: Then we can go for a walk!

The choice to use the MC vs. BC in a conversation is not arbitrary and requires attention to the interlocutor's inviting cues; otherwise, it can be perceived as unnatural or even disruptive (Sacks, Schegloff, & Jefferson, 1974). For example, if the speaker pauses after their sentence is grammatically complete (e.g., "I am going to the library.") accompanied by a falling intonation, this is most likely a signal that the speaker is yielding the MC. If, however, the speaker makes a slight pause while their sentence is not yet complete (e.g., "I am going to the library and.."); this is unlikely an invitation to take the floor. It is more appropriate in such a case to use the BC and provide a signal of attentive listening, allowing the speaker to continue (Ford & Thompson, 1996; Sacks et al., 1974; Cathcart, Carletta, & Klein, 2003; Skantze, 2021; Gravano & Hirschberg, 2011; Ward & Tsukahara, 2000; Duncan, 1972)

While developmental research has studied children's use of MC and BC, it has treated these two aspects of coordination separately. Work on MC has primarily focused on children's developing skills in terms of optimizing the response latency, i.e., avoiding excessive overlaps and pauses between turns (for a review, see Nguyen, Versyp, Cox, & Fusaroli, 2022). As for the BC, researchers have studied children's ability to provide and capitalize on listener feedback, but often in a context where the use of MC is not a valid option, e.g., during storytelling or while listening to an experimenter's instructions (e.g., Hess & Johnston, 1988; Peterson, 1990; Park, Gelsomini, Lee, & Breazeal, 2017).

A more accurate characterization of children's ability to engage in coordinated communication requires investigating their appropriate use of *both* the MC and BC of the conversation. The current study is a step toward addressing this question. We focus on middle childhood (6 to 12 years old) as some research has found children in this period to be still lacking in their conversational skills (Hess & Johnston, 1988; Baines & Howe, 2010; Maroni, Gnisci, & Pontecorvo, 2008), whereas others have found them to already show adult-like behavior in some specific aspects (e.g., Bodur, Nikolaus, Prévot, & Fourtassi, 2023), Middle childhood is, therefore, a good starting point to investigate the developmental status of a complex coordination phenomenon that may require relatively sophisticated socio-cognitive abilities (Devine & Hughes, 2014).

We follow the research method outlined in Liu, Nikolaus, Bodur, and Fourtassi (2022) by modeling children's coordination in a predictive fashion: We train a model capable of handling sequential/time-dependent data (here, a Long Short Term Memory recurrent neural network or LSTM) to predict *when* the child makes a specific conversational move (in our case, the use of MC vs. BC), based solely on the interlocutor's immediately preceding communicative cues, which we call hereafter "inviting cues." If the trained model makes this prediction with a higher-than-chance accuracy, it suggests the

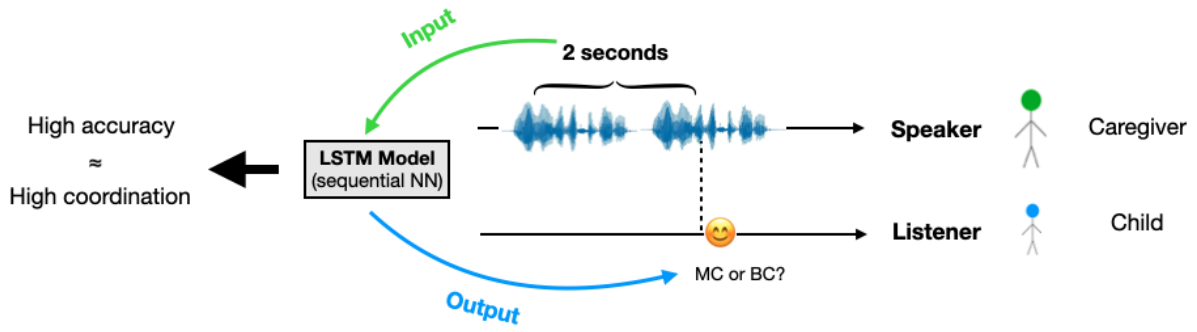

Figure 1: Schematic illustration of how we characterize coordination between two interlocutors (here a child-caregiver dyad). We train a model to predict the timing of the child's conversational move (Main channel or Back channel) based on the caregiver's immediately preceding communicative signals (for simplicity, we only illustrated the speech signal but non-verbal cues are also taken into account). The prediction accuracy of the model quantifies the extent to which both the child and caregiver have been successfully coordinating for the child to select the appropriate channel of the conversation.

child makes their moves *selectively*, based on whether or not their interlocutor had provided the relevant (inviting) cues for that specific move. The prediction accuracy of the model quantifies the extent to which both the child and interlocutor have been successfully coordinating for the child to contribute to the conversation appropriately (see Figure 1).

Using this research approach, we study the MC vs. BC coordination skills in children conversing with their caregivers. We quantify both their ability to select the most appropriate channel of conversation by reacting consistently to the caregiver's inviting cues (the child model) and, in turn, their ability to offer reliable inviting cues that the caregivers can use to pick a channel (the caregiver model). Crucially, these inviting cues can be *multimodal* and may involve changes in intonation, gaze, gesture, and/or sentence structure. The current study thus investigates children's ability to both interpret and offer such rich multimodal cues to negotiate the MC vs. the BC of the conversation with the interlocutor.

Finally, to draw conclusions about development, we compare children's skills not only to those of their adult conversational partners (i.e., the caregiver) but also to the coordination dynamics between two adults recorded in a similar conversational context. The reason we need this additional developmental "end-state" reference is twofold: (i) the performance of a model (as illustrated in Figure 1) cannot be interpreted separately for each interlocutor in a given dyad; the caregiver's model quantifies not only their ability to capitalize consistency on children's inviting cues but also the ability of children to provide these cues in a reliable fashion, and (ii) research indicates that caregivers tend to adapt to children's conversational competencies (e.g., Snow, 1977; Misiek & Fourtassi, 2022; Fusaroli, Weed, Rocca, Fein, & Naigles, 2021; Jiang, Frank, Kulkarni, & Fourtassi, 2022).

Methodology

In this section, we describe 1) the conversational dataset, 2) how we characterized the outcome measures, i.e., the MC and BC, 3) how we extracted the predictors, i.e., the inviting cues in the verbal, vocal, and visual modalities, 4) the model that uses these inviting cues to predict the outcome measures, and finally, 5) the experiments that we conducted using this model.

Conversational dataset

We use the ChiCo corpus (Bodur, Nikolaus, Kassim, Prévot, & Fourtassi, 2021). This corpus consists of video call recordings at home[1] of 10 conversations between children (aged between 6 to 12 years old) interacting with their caregivers (Child-Caregiver condition) and 10 conversations between the same caregivers interacting with other adults (Adult-Caregiver condition). To elicit a balanced exchange between children and caregivers, the conversation takes the form of an intuitive and weakly constrained game where interlocutors try to guess each other's words, giving participants the freedom to talk spontaneously. Each conversation lasted around 15 minutes, for a total of 5 hours and 49 minutes across both conditions. The setup required that interlocutors use different devices and that they communicate from different rooms (if they record from the same house) to avoid issues due to echo. The creators of the corpus took the necessary measures to ensure that BC signals were not suppressed as "background noise," by the Zoom software. For further technical details about the corpus, we refer the reader to the original paper.

Characterization of MC and BC

MC coding We segment the conversations into "turns", i.e., when an interlocutor is understood to be taking the MC. We follow research in dialog systems regarding how we define a turn and how we automatically detect it using speech technology (Skantze, 2021). A turn is defined/approximated as a stretch of speech from one interlocutor without any silence exceeding a certain amount (also known as Inter-Pausal Units, IPUs). We segmented speech into IPUs using the voice

[1] Using Zoom software

activity detector in SPPAS software (Bigi, 2015). The corpus comes with two separate audios for interlocutors (since each is recorded with a different microphone/computer), which allowed us to segment IPUs for each speaker without having to do speaker diarization or deal with speech overlap issues.

We set the minimum duration of an IPU to 150ms to be able to detect short utterances. We excluded instances of verbal BC of a similar length (using the set of BC that were already coded in the ChiCo corpus, see below). Indeed, a short segment like "yeah" can be both a response to a question, in which case it was labeled as an MC move, but it can also be a way to show attentive listening, in which case it was labeled as a BC move. We set the maximum duration of silence (within a turn) to 500ms. In addition, we set a threshold on the volume (to distinguish silence/noise from speech) to be of a minimum of 150 rms in the case of children and a minimum 200 rms for adults (this difference is to account for the fact that children tend to speak with a lower volume). Finally, we manually checked and corrected the outcome of the automatic annotations.

BC coding Instances of BC were already available in the ChiCo corpus. They were manually coded and included verbal instances such as "mmhm", "uh-huh", "okay" and non-verbal instances such as head nods and smiles. Descriptive statistics of both MC and BC instances in the corpus are shown in Table 1.

Multimodal Inviting Cues

We used vocal, visual, and verbal cues that could play a signaling role, inviting communicative moves from the interlocutor in face-to-face conversations (e.g., Holler & Levinson, 2019).

Visual Cues The visual features are manually annotated and are provided as a part of the ChiCo corpus. Most of these cues have been found in previous research to be relevant to turn-taking/MC management or BC signaling. These cues are head movements (nods & shakes), gaze, eyebrow movements (raises & frowns), mouth curves (smiles & laughs), and body posture (leaning forwards & backwards) (Duncan, 1972; Paggio & Navarretta, 2013; Kendon, 1967; Park et al., 2017; Brunner, 1979). We use one-hot encoding for the visual features, i.e., for each time frame, the visual cues were represented with a vector of ones (for cues occurring in the frame at hand) and zeros (for cues not occurring in that frame).

Vocal Cues For the vocal cues, we use the features extracted by Liu et al. (2022) for their BC study on the ChiCo corpus. These features are a subset of the eGeMAPS features (Eyben et al., 2016) a standard set of features commonly used for automatic voice annotation, including in previous work on inviting cues for MC and BC in adult-adult conversations (Murray et al., 2022; Jain & Leekha, 2021; Morency, de Kok, & Gratch, 2010; Goswami, Manuja, & Leekha, 2020; Ruede, Müller, Stüker, & Waibel, 2017). The categories of cues we used are pitch (variation), Mel-Frequency Cepstral Coefficients (MFCC), voice quality, energy, and pausal information.

Verbal Cues For the textual features, we relied on the Part-Of-Speech (POS) tags extracted by Liu et al. (2022). We use these features to represent the morpho-syntactic cues (e.g., indicating whether a sentence is complete). We know from previous research that interlocutors can use morpho-syntactic cues for coordinating both BC and MC (Cathcart et al., 2003; Ford & Thompson, 1996). We had a total of 17 POS tags and we used a one-hot encoding to signal the presence or absence of each POS tag for each time frame.

LSTM Model

The model should take as input inviting cues from one interlocutor to predict the channel of the conversation selected by the other. For all our experiments (see below), we make use of a recurrent neural network known as Long Short-Term Memory (hereafter LSTM) (Hochreiter & Schmidhuber, 1997). We use this modeling architecture because of its ability to capture sequential input. This feature is crucial for learning and testing many important inviting cues that are sequential in nature, such as the utterance structure and some vocal features (e.g., rising vs. falling intonation). Following previous work (e.g., Jain & Leekha, 2021), the model is fed a sequence of 40 time-frames of 50ms each (that is, a 2-second-long context window[2]) where each frame contains information about the value (or presence/absence) of all the cues considered. The context window immediately precedes the target move, and the goal of the model is to guess the identity of this move, i.e., MC or BC (or nothing), depending on the experiment (see Experiments).

For each target conversational move, we predicted its early few frames, more precisely, the first 4 frames (while moving the context window input accordingly). This is done for each frame independently and without seeing the values of the preceding frames (remember, the model only "sees" the other interlocutor). Predicting more than one frame makes the model more robust to noise. At the same time, we do not predict frames much further into the target conversational move in order not to trivialize the task. To illustrate, imagine the move to be predicted is an MC and that our target participant is now taking the floor for a few seconds while the interlocutor is completely silent. Training the model to predict MC frames at this point will make it – trivially – associate the prediction of MC with silence (as the predictive 2-second context window will be mostly "empty"). If we restrict the prediction to just the first few frames of the move, the model would be forced to learn the cues used by the target interlocutor to *initiate* their move.

The LSTM has several hyperparameters (such as the number of hidden dimensions, neural layers, dropout, learning rate, batch size, etc.). We tuned these hyperparameters using Ray Tune (Liaw et al., 2018). The hyperparameters have

[2]We experimented with larger context window sizes, but this led to lower model performance.

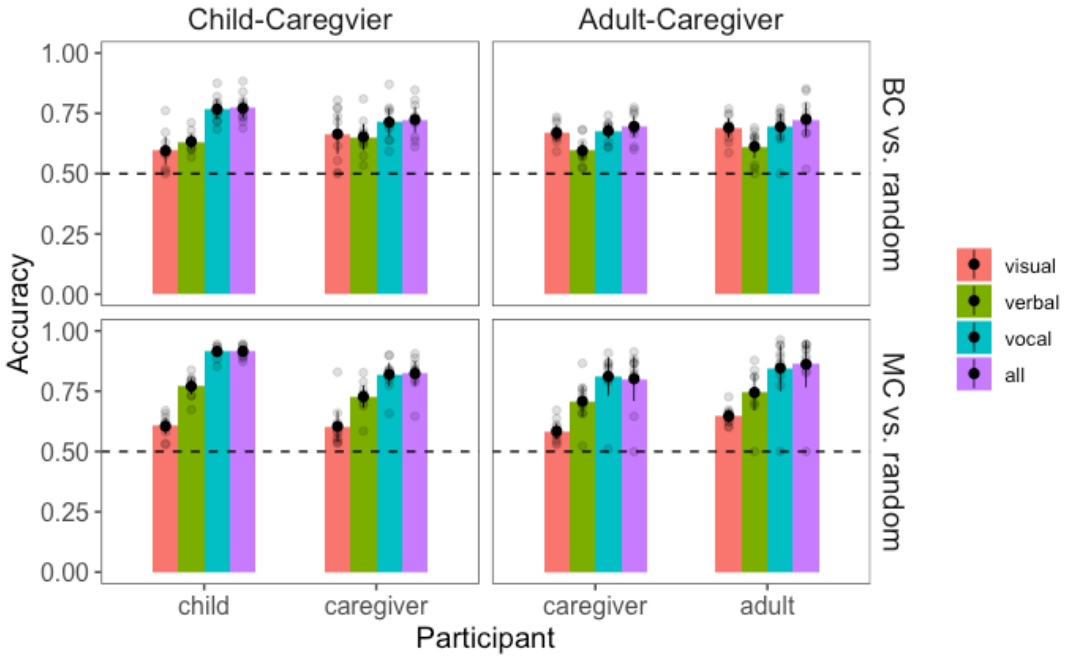

Figure 2: Accuracy scores of the BC predicting models (Experiment 1, top) and the MC predicting models (Experiment 2, bottom) both in the child-caregiver condition (left) and the adult-caregiver condition (right). In each case, we show the results of the modality-specific models (i.e., the models using predictors from a single modality) and the full model (where cues from all modalities are combined). We show the accuracy scores obtained with different training/testing configurations according to the LOOCV cross-validation (here represented with dots) in addition to the mean and 95 % confidence intervals over these scores. The dotted line represents the chance level.

been tuned for each of the three Experiments below. Further, the tuning was done with respect to both children's data and adult data[3]. Finally, the hyperparameters were tuned with respect to the model that uses inviting cues from all modalities. For each experiment, the same hyperparameters are used to train models across all 4 groups of participants (child and caregiver in the first condition and adult and caregiver in the second) and for single-modality models.[4]

Model training and evaluation The conversational data is heavily imbalanced with respect to our target moves (i.e., MC or BC), as the speech signal contains many more frames containing neither a BC signal nor a MC switch between interlocutors. To obtain interpretable accuracy scores, we train and test the models to discriminate between our target frames and a sample of an *equal* number of random frames in each conversation.[5] As for model evaluation, and to test the ability of our models to generalize across participants, we use the Leave-One-Out Cross-Validation technique (hereafter, LOOCV). If we take the child model as an example, LOOCV means that we train the model on all children except one, and then we test it on the child that was left out in training. This procedure is repeated with all training/testing configurations (here we have 10 children, which means we have 10 possible configurations and 10 accuracy scores evaluating each model).

Experiments

We had three sets of experiments. Each experiment was conducted on all groups of participants. Further, for each experiment and each group of participants, we did a feature ablation study by considering only the set of inviting cues belonging to a particular modality, one at a time. Table 1 describes the size of data used (in terms of frames) in each experiment.

Experiment 1: BC vs. random non-BC In this set of experiments, our goal was to replicate the results reported by Liu et al. (2022) on the same corpus regarding the prediction of BC moves (which they did separately from MC). We trained the model to use inviting cues from the speaker to identify instances of the listener's BC. The model had to distinguish BC instances from an equivalent number of random non-BC frames in each conversation. In this random sample, we did not consider frames from inside the target interlocu-

[3]We found almost no changes in the results across these two sets of hyperparameters, so we only report the results using the first.

[4]The details of the hyperparameters as well as all materials and code necessary to reproduce the results can be found at `https://osf.io/jv6y2/?view_only=3bdd2749410142dbbfbd2ad3ec97c54c`

[5]Except in the case of Experiment 3 (as we describe in the subsection below).

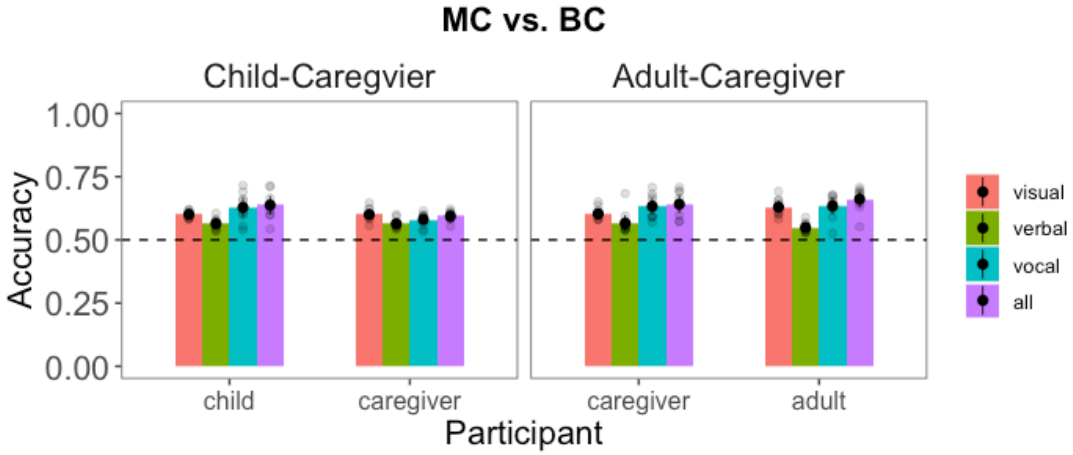

Figure 3: Accuracy scores of the BC vs. MC predicting models (Experiment 3) both in the child-caregiver condition (left) and the adult-caregiver condition (right). In each case, we show the results of the modality-specific models (i.e., the models using predictors from a single modality) and the full model (where cues from all modalities are combined). We show the accuracy scores obtained with different training/testing configurations according to the LOOCV cross-validation (here represented with dots) as well as the mean and 95 % confidence intervals over these scores. The dotted line represents the chance level.

tor's turns (while the other interlocutor is silent), as this could trivialize the task by making the model learn to associate BC move with trivial features in the inviting cues such as "no silence."

Experiment 2: MC vs. random non-MC In this set of experiments, we test the prediction of MC moves (separately from BC). The procedure was similar to Experiment 1. In the random non-MC sample, we did not consider frames from inside the turn (for the same reason as above).

Experiment 3: BC vs. MC While Experiments 1 and 2 tested the prediction of BC and MC independently from each other, Experiment 3 dealt with both. Crucially, here we did not test the ability of the models to identify BC or MC signals from a random sample of frames but to tease these two signals apart. We trained and tested the models on an equal sample of BC and MC (see Table 1).

Table 1: The number of BC, MC, and/or random samples used in our experiments per interlocutor in each condition.

Interlocutor	Experiment 1		Experiment 2		Experiment 3	
	BC	Rand.	MC	Rand.	BC	MC
Child	1836	1836	5191	5191	1836	1836
Caregiver/C	1640	1640	6802	6802	1640	1640
Adult	2736	2736	5321	5321	2736	2736
Caregiver/A	2532	2532	6340	6340	2532	2532

Results

Experiment 1 and 2: Figure 2 shows the scores for the predictability of BC moves on the one hand (top) and of MC moves on the other (bottom) made by one interlocutor (i.e., the outcome measures), given the immediately preceding 2-second window of multimodal cues from the *other* interlocutor (i.e., the predictors). We show the results for predictors in a single modality ("visual", "vocal", or "verbal") and for inviting cues from all modalities combined ("all").

We report two main findings. The first is that the overall predictability of both MC and BC moves (i.e., "all") is well above chance across all groups of interlocutors in both conditions. This finding suggests that interlocutors provide consistent, informative cues to invite MC and BC moves *and* – when on the receiving end – they capitalize on these cues to make the corresponding move. Crucial to our research goals, this was observed in both children and adults alike, thus replicating the results reported in Liu et al. (2022) for the case of BC and extending them to the case of MC as well. The second finding concerns the predictive power of single modalities: We found that all three modalities, when considered alone, allowed for an above-chance prediction[6] of both BC and MC moves. That said, cues in the vocal modality were, overall, the most informative, especially in the case of MC. Here again, this finding was observed in both children and adults.

Experiment 3: Figure 3 shows the scores quantifying the ability of predictors from one interlocutor to distinguish when the other interlocutor is making a BC move or an MC move.

[6]As can be deduced from the fact the %95 confidence intervals do not cross the chance threshold of 0.5

We report two main findings. The first is that the scores for the combined cues (i.e., "all") are above chance, suggesting that interlocutors do not only provide – and capitalize on – consistent cues to invite MC and BC moves (as reported in Experiments 1 and 2 above), they *also* provide and capitalize on cues that are *distinctive* to MC vs. BC moves, allowing interlocutors (both children and adults) to coordinate in terms of which conversational channel is more appropriate to use at a specific time. The second finding concerns the role of specific modalities. Each modality contained predictive cues, allowing the distinction of BC from MC moves. In contrast to Experiments 1 and 2, where the cues from the vocal modality were predominant, this was no longer the case here. In particular, the visual modality seems to bring, overall, as helpful cues as the vocal modality does.

Discussion

This paper studied an essential dimension of children's conversational coordination: How they coordinate the use of the main vs. the back channel of the conversation with their interlocutors. While previous work in the developmental literature has studied aspects of both main-channel and back-channel coordination (Nguyen et al., 2022; Hess & Johnston, 1988; Peterson, 1990; Park et al., 2017; Bodur et al., 2023), the current is – to the best of our knowledge – the first to study both phenomena jointly, especially in a face-to-face setting. The goal is to better characterize the complexity of the challenge that children face, i.e., learning how to coordinate across several dimensions *simultaneously* and learning this coordination not only with words but also via *multimodal* signaling.

To capture children's spontaneous use of their communicative skills in real life, we relied on a corpus of dialogs where children conversed freely with their caregivers at home while playing an intuitive word-guessing game. Such naturalistic data come with a methodological challenge: Unlike in-lab, controlled experiments where communicative signals (or the elicitation of these signals) are pre-designated *top-down* by the experimenter, here we need to rely on sophisticated computational tools that allow a *bottom-up* study of how interlocutors negotiate their contribution to the conversation via complex multimodal signaling mechanisms (that cannot all be anticipated a priori by an experimenter).

Thus, following Liu et al. (2022), we borrowed techniques from the literature on dialog systems (e.g., Skantze, 2021) to provide a quantitative account of children's coordination skills in a naturalistic context. This approach was fruitful as our study resulted in several findings. Consider first the results from the "child models" (compared to the "adult models"): (a) We replicated the results from Liu et al. (2022), confirming that, by middle childhood, children show adult-like behavior in terms of their high responsiveness to interlocutors' inviting cues to provide BC signals (Experiment 1), (b) we extended this finding and showed that children are also on par with adults in their consistency in reacting to inviting cues to take the MC (Experiment 2), and (c) we found children to be as capable as adults in *selectively* reacting to the inviting cues specific to BC vs. MC (Experiment 3). If we look at results from the "caregiver model" in the child-caregiver condition, we found that children also showed similar consistency (to adults) in terms of *providing* relevant, inviting cues for the caregiver to capitalize on.

The overall accuracy scores for Experiment 3 were lower compared to those obtained in Experiments 1 and 2. This reflects the fact that the task in Experiment 3 is much harder: The models did not only have to predict instances of BC and MC signals but to differentiate these two signals, whose inviting cues may overlap. This is also apparent regarding the role of modalities. In particular, the vocal modality played a rather dominant role in predicting MC and – to some extent – BC, but this role diminished when the models needed to tease MC and BC apart (and we observe an opposite pattern for the visual modality). This could be due to the fact that both BC and MC share some similar vocal inviting cues (e.g., they can both be invited by pauses) while they may diverge slightly in terms of visual cues (e.g., pausing while looking away invites BC but pausing while looking at the interlocutor invites MC). More research is needed for a finer-grained examination of these findings.[7]

Limitations and future work

The current work, like any data-driven modeling study of naturalistic data, remains mainly correlational. The *causality* of the conclusions we draw from it should thus be taken with a grain of salt (pending further confirmatory work). For example, while we found that models mimicking children's behavior (given similar contextual input) performed similarly to the models mimicking adults, this finding does not entail with certainty that children and adults use exactly the same coordination mechanisms. Take, e.g., the result that all modalities were predictive of children's BC vs MC moves. This could be due to caregivers systematically providing multimodal signals in a redundant fashion, and not necessarily to children capitalizing on all these modalities.

Another limitation of the current study is its reliance on video-call data as an approximation of face-to-face conversations. While this data acquisition method allows for naturalistic recording (it takes place at home instead of the unfamiliar context of a lab), it also involves introducing a medium (i.e., a screen) and is subject to time lag issues (Boland, Fonseca, Mermelstein, & Williamson, 2022). While our conclusions remain valid in this specific context, more research is required to precisely quantify the potential effect that online video call systems might have on conversational coordination as opposed to direct face-to-face communications.

[7] Here, we could not apply off-the-shelf interpretability algorithms such as SHAP (Lundberg & Lee, 2017) due to their presupposition of feature independence (a condition that is not met in our data).

Acknowledgements

This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011013886). 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).

References

Baines, E., & Howe, C. (2010). Discourse topic management and discussion skills in middle childhood: The effects of age and task. *First Language*, *30*(3-4), 508–534. doi: 10.1177/0142723710370538

Bigi, B. (2015). Sppas-multi-lingual approaches to the automatic annotation of speech. *The Phonetician. Journal of the International Society of Phonetic Sciences*, *111*(ISSN: 0741-6164), 54–69.

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* (p. 158–163). New York, NY, USA: Association for Computing Machinery. doi: 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*.

Boland, J. E., Fonseca, P., Mermelstein, I., & Williamson, M. (2022). Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General*, *151*, 1272-1282. doi: 10.1037/xge0001150

Brunner, L. J. (1979). Smiles can be back channels. *Journal of Personality and Social Psychology*, *37*, 728-734. doi: 10.1037/0022-3514.37.5.728

Cathcart, N., Carletta, J., & Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the tenth conference on european chapter of the association for computational linguistics-volume 1* (pp. 51–58).

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child development*, *85*(5), 1777–1794.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*, 283–292. (Place: US Publisher: American Psychological Association) doi: 10.1037/h0033031

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... Truong, K. P. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190-202. doi: 10.1109/TAFFC.2015.2457417

Ford, C., & Thompson, S. (1996). Interactional units in conversation: Syntactic, intonational and pragmatic resources. *Interaction and grammar*(13), 134.

Fusaroli, R., Weed, E., Rocca, R., Fein, D., & Naigles, L. (2021, Nov). *Caregiver linguistic alignment to autistic and typically developing children.* PsyArXiv. Retrieved from psyarxiv.com/ysjec doi: 10.31234/osf.io/ysjec

Goswami, M., Manuja, M., & Leekha, M. (2020). *Towards social & engaging peer learning: Predicting backchanneling and disengagement in children.* arXiv. doi: 10.48550/ARXIV.2007.11346

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, *25*(3), 601-634. doi: https://doi.org/10.1016/j.csl.2010.10.003

Hess, L. J., & Johnston, J. R. (1988). Acquisition of back channel listener responses to adequate messages. *Discourse Processes*, *11*(3), 319-335. doi: 10.1080/01638538809544706

Hochreiter, S., & Schmidhuber, J. (1997, nov). Long short-term memory. *Neural Comput.*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639–652.

Jain, V., & Leekha, M. (2021, 5). Exploring semi-supervised learning for predicting listener backchannels. In *Conference on human factors in computing systems - proceedings.* doi: 10.1145/3411764.3445449

Jiang, H., Frank, M. C., Kulkarni, V., & Fourtassi, A. (2022). Exploring patterns of stability and change in caregivers' word usage across early childhood. *Cognitive Science*, *46*(7), e13177.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*, 22-63. doi: https://doi.org/10.1016/0001-6918(67)90005-4

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., & Stoica, I. (2018). Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

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). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3536220.3563372

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Maroni, B., Gnisci, A., & Pontecorvo, C. (2008). Turn-taking in classroom interactions: Overlapping, interruptions and pauses in primary school. *European journal of psychology of education*, *23*, 59–76.

Misiek, T., & Fourtassi, A. (2022, August). Caregivers exag-

gerate their lexical alignment to young children across several cultures. In *Proceedings of the 26th workshop on the semantics and pragmatics of dialogue - full papers.* Dublin, Ireland: SEMDIAL. Retrieved from http://semdial.org/anthology/Z22-Misiek_semdial_0005.pdf

Morency, L. P., de Kok, I., & Gratch, J. (2010, 1). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, *20*, 70-84. doi: 10.1007/s10458-009-9092-y

Murray, M., Walker, N., Nanavati, A., Alves-Oliveira, P., Filippov, N., Sauppe, A., ... Cakmak, M. (2022). Learning backchanneling behaviors for a social robot via data augmentation from human-human conversations. In *Conference on robot learning* (pp. 513–525).

Nguyen, V., Versyp, O., Cox, C., & Fusaroli, R. (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. doi: https://doi.org/10.1111/cdev.13754

Paggio, P., & Navarretta, C. (2013, Mar 01). Head movements, facial expressions and feedback in conversations: empirical evidence from danish multimodal data. *Journal on Multimodal User Interfaces*, *7*(1), 29-37. doi: 10.1007/s12193-012-0105-9

Park, H. W., Gelsomini, M., Lee, J. J., & Breazeal, C. (2017). Telling stories to robots: The effect of backchanneling on a child's storytelling. In *2017 12th acm/ieee international conference on human-robot interaction (hri* (p. 100-108).

Peterson, C. (1990). The who, when and where of early narratives. *Journal of child language*, *17*(2), 433–455.

Ruede, R., Müller, M., Stüker, S., & Waibel, A. (2017). Enhancing backchannel prediction using word embeddings. In *Interspeech* (pp. 879–883).

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735.

Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, *67*, 101178. doi: https://doi.org/10.1016/j.csl.2020.101178

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, *4*(1), 1–22. doi: 10.1017/S0305000900000453

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, *32*(8), 1177–1207.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago linguistics society, 6th meeting, 1970* (pp. 567–578).