Communicative Coordination in Child-Caregiver Interactions

Abhishek Agrawal

amU
Aix Marseille Université

NNT : 2025AIXM0261

THÈSE DE DOCTORAT

Soutenue à AMU — Aix-Marseille Université
le 07 Octobre 2025 par

Abhishek AGRAWAL

Communicative Coordination in Child-Caregiver Interactions

Discipline
Informatique

École doctorale
ED 184 Mathematiques et Informatique

Laboratoire/Partenaires de recherche
Laboratoire d'Informatique et Systèmes (LIS)
Institute of Language, Communication and the Brain (ILCB)

Composition du jury

Philippe MULLER Maître de Conférence HDR Université Paul Sabatier	Rapporteur
Mathilde FORT Maître de Conférence HDR Université Grenoble Alpes	Rapporteure
Delphine BATTISTELLI Professeur Université Paris Nanterre	Examinatrice
Marianne JOVER Professeur Aix-Marseille Université	Examinatrice
Sho TSUJI Chargé de Recherche CNRS Ecole Normale Supérieure	Examinatrice
Laurent PRÉVOT Professeur Aix-Marseille Université	Président du jury
Benoit FAVRE Professeur Aix-Marseille Université	Directeur de thèse

Invité

Abdellah FOURTASSI Maître de Conférence HDR Aix-Marseille Université	Co-directeur de thèse

LIS LABORATOIRE D'INFORMATIQUE & DES SYSTÈMES UMR 7020

ILCB Institute of Language, Communication and the Brain

Affidavit

I, undersigned, Abhishek Agrawal, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Benoit Favre and Abdellah Fourtassi, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the french national charter for Research Integrity and AMU charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Marseille, 03 July 2025

List of publications and conference participations

List of peer-reviewed publications published within the context of this thesis:

1. Agrawal Abhishek, Liu Jing, Bodur Kübra, Favre Benoit, & Fourtassi Abdellah (2023). **Development of Multimodal Turn Coordination in Conversations: Evidence for Adult-like behavior in Middle Childhood.** *Proceedings of the 45th Annual Meeting of the Cognitive Science Society.*

2. Agrawal Abhishek, Nikolaus Mitja, Favre Benoit, & Fourtassi Abdellah (2024). **Automatic Coding of Contingency in Child-Caregiver Conversations.** *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).*

3. Agrawal Abhishek, Favre Benoit, & Fourtassi Abdellah (2024). **Analysing Communicative Intent Coordination in Child-Caregiver Interactions.** *Proceedings of the 46th Annual Meeting of the Cognitive Science Society.*

4. Nikolaus Mitja, Agrawal Abhishek, Kaklamanis Petros, Warstadt Alex, & Fourtassi Abdellah (2024). **Automatic Annotation of Grammaticality in Child-Caregiver Conversations.** *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).*

5. Goumri Dhia, Agrawal Abhishek, Nikolaus Mitja, Vu Hong, Bodur Kübra, Emmar Elias, Armand Cassandre, Mazzocconi Chiara, Gupta Shreejata, Prévot Laurent, Favre Benoit, Becerra-Bonache Leonor, & Fourtassi Abdellah (2024). **CHICA: A Developmental Corpus of Child-Caregiver's Face-to-face vs. Video Call Conversations in Middle Childhood.** *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).*

6. Agrawal Abhishek, Favre Benoit, & Fourtassi Abdellah (2025). **Identifying Repair Opportunities in Child-Caregiver Interactions.** *Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue.*

7. Agrawal Abhishek, Favre Benoit, & Fourtassi Abdellah (Currently under revision). **Mapping the Communicative Landscape of Early Child-Caregiver Dialogue.** In the journal of *Cognitive Science.*

Participation at conferences and summer schools:

1. Institute of Language, Communication and the Brain (ILCB) Summer School 2022
2. Institute of Language, Communication and the Brain (ILCB) Summer School 2023
3. Institute of Language, Communication and the Brain (ILCB) Summer School 2024
4. Lisbon Machine Learning Summer School (LxMLS) 2023
5. CogSci 2023
6. CogSci 2024
7. LREC-COLING 2024
8. CORIA-TALN 2025
9. SemDial (Bialogue) 2025

Résumé et mots clés

La communication est une tâche complexe qui exige des interlocuteurs une coordination simultanée à plusieurs niveaux. Cette capacité à coordonner la communication va de pair avec la capacité de l'enfant à acquérir le langage, puisque les interactions sociales jouent un rôle crucial dans l'acquisition du langage. En outre, elle joue un rôle important dans le bien-être socio-cognitif global de l'enfant. La recherche sur le développement de la coordination communicative chez les enfants est traditionnellement basée sur des études d'observation à petite échelle ou sur des études expérimentales contrôlées en laboratoire. Nous pensons qu'en plus de ces travaux, des études quantitatives à plus grande échelle basées sur des données naturelles sont nécessaires pour une compréhension plus complète de la coordination communicative chez les enfants. Nous soutenons que l'arrivée du récent "printemps de l'intelligence artificielle" nous a fourni de nouveaux et puissants modèles d'apprentissage automatique (ML) qui peuvent être exploités pour une étude quantitative plus écologiquement valide du développement de la coordination communicative chez les enfants.

Pour tester cette théorie, nous étudions dans cette thèse la coordination communicative dans les interactions enfant-parent à trois niveaux différents, à savoir : i) la gestion du tour de rôle, ii) la cohérence du dialogue et iii) l'ancrage conversationnel à l'aide de modèles de ML. Nous utilisons les modèles de ML comme outil pour annoter automatiquement un large corpus contenant divers phénomènes de coordination, ce qui nous permet de mener une étude à grande échelle du développement de ces phénomènes dans les interactions enfant-parent. Cette approche nous a permis de délimiter le paysage communicatif des premières interactions entre l'enfant et le parent, en termes d'intentions communicatives et de cohérence. Elle a également conduit à la découverte que les parents saisissent rarement l'occasion de réparer les malentendus lors de ces interactions. Nous avons également constaté que les enfants comprennent la notion de base du tour de rôle assez tôt dans leur enfance. Nous utilisons également les modèles de ML pour modéliser informatiquement des mécanismes tels que le tour de parole dans les interactions enfant-parent. Avec cette approche, nous trouvons des indices d'un comportement de prise de tour similaire à celui des adultes chez les enfants en milieu d'enfance.

Les différents résultats issus de nos études constituent des étapes exploratoires initiales visant à montrer comment l'apprentissage automatique peut être mobilisé pour une étude à grande échelle, complète et écologiquement valide du développement de la coordination communicative chez les enfants.

Mots clés: Coordination communicative, Tour de parole, Cohérence, Connaissances partagées, Interactions enfant–parent, Apprentissage automatique

Abstract and keywords

Communication is a complex task requiring interlocutors to simultaneously coordinate on multiple levels. This ability to coordinate communication goes hand-in-hand with a child's ability to acquire language since social interactions play a crucial role in language acquisition. Furthermore, it also plays an important role in the overall socio-cognitive well-being of the child. Research on the development of communicative coordination in children has been traditionally based on small-scale observational studies or on controlled experimental studies in lab environments. We believe that in addition to these studies, more large-scale quantitative studies based on naturalistic data are required for a more comprehensive understanding of communicative coordination in children. We argue that the arrival of the recent "Artificial Intelligence (AI) Spring" has provided us with new and powerful Machine Learning (ML) models that can be leveraged for a more ecologically valid quantitative study of the development of communicative coordination in children.

To test this theory, in this thesis, we study communicative coordination in child-caregiver interactions on three different levels, namely: i) turn-taking management, ii) dialog coherence and iii) conversational grounding with the help of ML models. We utilize ML models as a tool to automatically annotate a large corpus for various coordinative phenomena thereby allowing us to conduct a large-scale bottom-up study of the development of those phenomena in child-caregiver interactions. This approach enabled us to delineate the communicative landscape of early child-caregiver interactions in terms of their communicative intents and coherence. It also led to the discovery that caregivers take the opportunity to repair misunderstandings in very limited cases during child-caregiver interactions. We also found that children understand the basic notion of turn-taking fairly early in their childhood. We also use ML models for computationally modeling mechanisms like turn-taking in child-caregiver interactions. With this approach we find evidence of adult-like turn-taking behavior in children in their middle-childhood.

The various insights obtained from our studies are the initial exploratory steps towards showcasing how ML can be leveraged for a comprehensive and ecologically valid large-scale study of the development of communicative coordination in children.

Keywords: communicative coordination, turn-taking, coherence, common ground, child-caregiver interactions, machine learning

Acknowledgements

I left the writing of this part of my thesis for the end, as it brings me bittersweet feelings: it marks the closure of a major chapter of my life (and three years!). It would be nearly impossible for me to express my gratitude and appreciation to the multitude of people who have helped me throughout my time as a doctoral candidate — the list is far too long. Please forgive me if I happen to leave anyone out.

First and foremost, I'd like to thank both Abdellah and Benoit for being such wonderful supervisors! Without your guidance and patience, I never would have been able to succeed. Abdellah, you took a chance with me and believed in me when I wasn't sure about myself, and for that I will always be deeply grateful.

I am also thankful to the entire TALEP team and my friends — Ioanna, Jules, Mitja, Maïwenn, Monica, Elie, Eliott, Alice and everyone else. I'm especially grateful to Dhia, whose help was invaluable in navigating French bureaucracy and for his help in my recovery after my accident.

A big thank you to all my friends who made my stay in this sunny and beautiful city even more enjoyable — Diya, Floor, Varun, and many others. I also owe my sanity to the countless discussions and shared Caipis with Dasha and Kübra — a thousand thanks for that.

And last but certainly not the least, I am grateful to my old friends and family for their constant support and encouragement, without which this thesis would not have been possible.

Contents

List of Figures

List of Tables

Introduction

Communication in between humans can be considered an art form in itself. It is a very complex task with a lot of nuanced parts working in concert with each other. It can almost be thought of as conducting a symphony where different tasks need to be coordinated between the interlocutors at both an interpersonal as well as intrapersonal level. For instance, at the intrapersonal level, the listener listens and interprets the utterance while planning their next utterance at the same time whereas at the interpersonal level the listener might need to indicate to the speaker that they understand the speaker's utterance (e.g., by nodding their head) or they have trouble understanding the utterance (e.g., by issuing a clarification request).

Communication isn't a one way street; it requires a coordinated effort by all the interlocutors involved in the communication to work together (H. H. Clark, 1992; H. H. Clark & Wilkes-Gibbs, 1986; Pickering & Garrod, 2004). When we say coordinated, we mean it in the literal sense as well as the sense that over time, as the conversation progresses, the interlocutors start affecting one another due to their interaction (Paxton & Dale, 2013a, 2013b).

To be able to socially interact with others, one must first be proficient or at least adequate in the requisite conversational skills. A lack or deficit in these skills can affect a person's capability of forming new and lasting relationships thereby affecting both their mental as well as physical health since humans as a species are social in nature and thrive when surrounded by a like-minded community. Furthermore, we learn by interaction (e.g., by learning from more knowledgeable people in school) and thus poor communication skills hamper our socio-cognitive development (De Felice et al., 2023; Garzaniti et al., 2011; Hazen & Black, 1989; Miczo et al., 2001; Place & Becker, 1991). A low conversational proficiency has also been shown to correlate with low popularity amongst peers and a reduced capacity to learn with others by collaborating (Black & Hazen, 1990; Murphy et al., 2014; Place & Becker, 1991).

To ensure that such a scenario doesn't occur and to mitigate its effects, it is crucial to study the development of communicative coordination in children. Once we gain better understanding of the different ages around which children start showing the capabilities of various mechanisms used to coordinate communication and can track when these capabilities mature to exhibit adult-like performance, we can potentially identify whether a child has some issues if they exhibit extremely poor capabilities compared to the norm. Furthermore, once we improve our understanding of the different complex mechanisms involved in communicative coordination and the way they work in conjunction, we can try to design specific interventions for children that lack the requisite skills to improve their capabilities to communicate effectively and not become social pariahs.

Theories of Communicative Coordination There are various theories of communicative coordination that have been proposed out of which the two primary ones are coordination as joint-action (H. H. Clark, 1996; Harris, 1996) and coordination as convergence (Giles et al., 1991; Pickering & Garrod, 2004). In the latter theory, the coordination is explained in terms of priming where the participants of the conversation automatically (often subconsciously) start aligning with each other on various levels over time. These levels can be lexical and syntactic (e.g., Ferreira & Bock, 2006; Pickering & Branigan, 1999), acoustic (e.g., Levitan & Hirschberg, 2011), motor (e.g., Robledo et al., 2021), etc. In contrast to this, the former theory explains coordination as an intentional process where the development of common ground and adapting to the other interlocutor's needs plays a central role (Brennan, 1991; Brennan & Hanna, 2009; H. H. Clark, 1996; H. H. Clark & Wilkes-Gibbs, 1986).

Levels of Communicative Coordination Considering the extensive research on conversation in the fields of pragmatics, psycholinguistics and conversation analysis (for e.g., H. H. Clark, 1996; Levinson, 1983; Pickering & Garrod, 2021; Sacks et al., 1974), Fourtassi (2023) argues that for a conversation to occur, it has to be coordinated on at least three levels (as indicated in Figure 1) These levels can be succinctly described as follows:

- **Turn-taking management:** The interlocutors involved in a conversation need to accurately determine when it is their turn to speak and their turn to listen to keep the conversation flowing with minimal pauses and interruptions.
- **Coherence:** All participants in a conversation need to contribute to it in a meaningful manner by building on top of each other's turns and the surrounding context of the conversation.
- **Conversational Grounding**: Interlocutors in a conversation need to ensure that the intents expressed throughout the conversation via their utterances are correctly understood by everyone by signalling their understanding (e.g., a head nod) or misunderstanding (e.g., by asking clarification request).

The emergence of certain aspects of coordination on each of these levels can be traced back to early childhood (E. V. Clark, 2020; Nguyen et al., 2022; Nikolaus et al., 2022). Children are able to acquire the syntax of a language and its phonology pretty early and fairly quickly in their childhood whereas the developmental trajectory for coordinating on the above mentioned levels spans much of the entire childhood period. There is evidence for this protracted timeline for turn-taking (e.g., Maroni et al., 2008), providing appropriate feedback as the listener which helps in communicative grounding (e.g., Hess & Johnston, 1988), and for coherency in conversations (e.g., Baines & Howe, 2010). Fourtassi (2023) posits that concurrent development of cognitive competencies like *Theory of Mind* (i.e., inferring other people's beliefs and mental state) and *Executive Functions* like inhibition control and working memory through the middle childhood period could be one of the reasons for this protracted timeline in the development of conversational skills (Matthews et al., 2018; Wang et al., 2016).

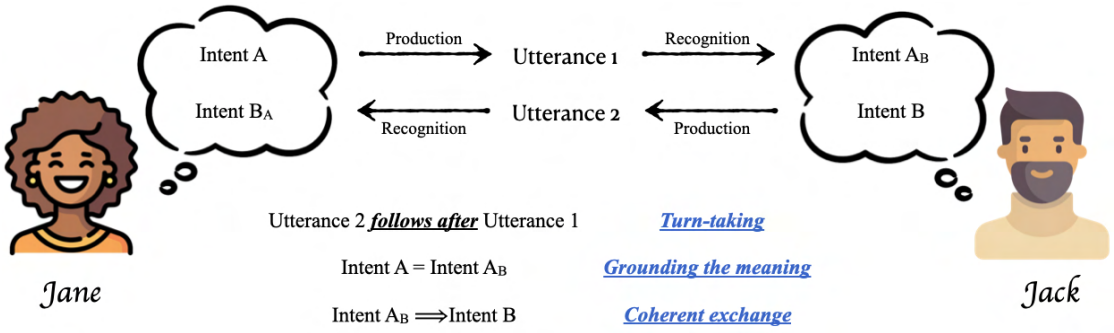

Figure 1.: Conversational skill development requires learning to coordinate across multiple levels in order to 1) master turn-taking, 2) negotiating and developing the shared belief space with the other interlocutors (also known as conversational grounding), and 3) contribute to the conversation in a coherent manner. The above figure demonstrates the multiple levels of coordination and the interaction between two people (Jane and Jack) in the following manner. Jane wants to communicate the *Intent A* to Jack which she does by producing the linguistic *Utterance 1*. Given Utterance 1, Jack tries to infer Jane's intent from it; however, since this inference is not deterministic it can possibly lead to a misunderstanding of the intent. Thus Jack here, doesn't recover Jane's true *Intent A* but his understanding of the the intent which is denoted here as *Intent* A_B. The levels of coordination can be viewed as follows: Firstly, during Jane's turn — while she continues to speak — Jack shouldn't interrupt and needs to wait for his turn (Turn-taking). Secondly, Jack's reaction (both verbal and non-verbal) to Jane's turn will ensure that the inferred intent is indeed correct, i.e., *Intent* A_B = *Intent A*; otherwise, one of them would initiate a communicative repair to clear up the misunderstanding (Grounding). Finally, after the end of Jane's turn and successful grounding of her *Intent A*, Jack can take his turn by producing *Utterance 2* which will be a linguistic expression of his *Intent B*. Note however, that this intent cannot be disconnected from Jane's previous discourse; it has to be relevant to the overall conversation, for instance by responding to a question with an answer while staying on topic (Coherence). Figure adapted from (Fourtassi, 2023).

Pitfalls of Traditional Research on Conversational Skills Traditionally, research on children's conversational skills was done primarily in a controlled, experimental setting or as an observational study. Controlled experimental studies are generally conducted in settings that lack genuine communicative coordination on multiple levels. This is true not only for non-interactive experiments but also for semi-interactive experiments where the experimenter uses a script rather than freely engaging in a conversation with the child. Thus, it is hard to judge how accurately the findings of these controlled experiments on their own truly reflect children's natural

conversational skills in the wild. Observational studies of children's conversational data focuses on fine-grained, in-context examinations in the style of conversation analysis and relies on frequency counts of observed phenomena. While these studies are able to provide us with some theoretically relevant and plausible insights, they are insufficient when it comes to quantifying the complex dynamics of multimodal and multilevel coordination. Another limitation of observational studies is that it is hard to scale them due to their reliance on manual annotations.

With the advent of modern neural networks and techniques for bolstering their performance (e.g., the attention mechanism), we believe we can leverage these computational models for investigating multimodal conversational development on large-scale, complex and naturalistic data. We embrace the two use cases proposed by Fourtassi (2023) for these neural network based machine learning (ML) models: ML as one of the tools in a researcher's toolkit to automatically scale the annotation of raw data and ML as a computational model of children's coordination.

ML as a Tool Manually annotating data is often a thankless and expensive task — in terms of labor, time and money — thereby resulting in small corpora of annotated data which can cause generalizability issues when trying to extrapolate the findings of these small corpora to larger data samples. Manual annotations are also often riddled with inconsistencies due to the fallible nature of humans and possibly due to cases of annotators' fatigue when the annotation scheme is multifaceted and highly involved as is the usual norm in social interactions. Although softwares for automatically annotating some communicative cues exist (e.g., OpenFace Baltrusaitis et al., 2016), care needs to be taken to ensure that these models can handle data from young children since they are generally trained on data from adults (see Erel et al., 2022). Thus, existing models need to first be evaluated against manual annotations by experts on children's datasets (e.g., Luchkina et al., 2025). If their performance is found to be lacking then, new tools should be developed to address any shortcomings of these off-the-shelf tools. This kind of undertaking has recently garnered the interest of developmental scientists with a marked degree of success in automating the coding of children's gaze (Erel et al., 2022, 2023), smiles (D.-E. Goumri et al., 2023), body posture (B. L. Long et al., 2022), head movements (López Pérez et al., 2017), and communicative intents (Nikolaus et al., 2022). An important consideration that needs to be highlighted with this approach is that while we can reduce the labor cost of manual annotations to some extent, we still need a *human-in-the-loop* for it to be effective. For instance, if a tool detects that the child is nodding his/her head, it is up to the expert annotator to determine the communicative significance given the conversation context (e.g., if it is a backchannel, an inadvertent gesture or a response to a polar question). Depending on the kind of analysis that is intended to be done with the automatically annotated data, another contribution that a human-in-the-loop brings is that of ethical considerations. For instance, the human can ensure that the model isn't being overly biased and that it isn't propagating any harmful stereotypes

in its predictions[1] (for e.g., Birhane, 2022; Birhane et al., 2024; Buolamwini & Gebru, 2018). An example for this is the Gender Shades study (Buolamwini & Gebru, 2018), where the authors found that commercial gender classification systems fail miserably in correctly identifying darker-skinned females as compared to lighter-skinned males.

ML as a Model Computational models have been traditionally used to test the learnability of various phenomena by providing it with the requisite ingredients necessary for said learning i.e., training data, an algorithm for learning and some assumptions (Alishahi, 2010). The benefits of using these models in a study is that they can be replicated since the algorithm is deterministic (usually), extensible by modifying the algorithm and offers accountability and transparency since you have control of all the variables in the model[2] (Cruz Blandón et al., 2023). Using ML to study a child's development is not a revolutionary new idea but an old one particularly when it comes to studying language acquisition in children (for e.g., Brent, 1997; Kelley, 1967; Langley & Carbonell, 1987). However, with the impressive advancements made in ML with the advent of deep learning and modern neural networks, Dupoux (2018) argues that we should make attempts to understand language acquisition in children by training neural networks with real sensory inputs — similar to what a child experiences in his natural environment — and evaluating on human-comparable tests. We believe this argument can be extended to use modern ML models to study children's conversational skill development as well. Researchers have previously used ML models as proxies for word learning from visual cues (Rane et al., 2023), to understand developmental changes in children's internal representations of concepts (B. Long et al., 2024), and to study the role of executive attention and the rearing environment on Effortful Control (EC) in early childhood (Musso et al., 2023). In this alternative approach we propose to make use of ML models as a quantitative model of children's conversational coordination. For instance, consider the case where a model is trained to predict when an interlocutor is going to perform a communicative move given signals from the other interlocutor. Studies have shown the feasibility of this approach when it comes to using a ML model to predict the listener's backchannel given the speaker's inviting cues as a model of coordination for backchannel signalling (J. Liu et al., 2022; Park et al., 2017).

Research questions Given the importance of conversational skills for the socio-cognitive development of children and the noticeable gap in literature (see Chapter 1) along with the lack of empirical evidence regarding the development of these skills in children, we believe it is crucial to address and shore up this gap in our understanding. We've identified some of the skills necessary to coordinate conversation from the available literature; however, we are still in the dark when it comes to our

[1] It is important to note here that this approach doesn't guarantee complete fairness and perfect accuracy in the model predictions as humans themselves are inherently flawed and carry their own social biases.

[2] This is not the case with modern neural networks, most of which are black boxes.

understanding of the development of these skills in children. Some of the challenges that researchers face in studying these phenomena were outlined above along with a potential approach to address these challenges.

Thus with the help of this thesis we will be exploring the following questions:

- Are children able to coordinate their interactions with caregivers on the following levels:
  - Are they able to take turns and provide feedback?
  - Are they coherent in conversations?
  - Can they negotiate and develop their shared mutual understanding with the caregiver?
- In terms of development, how does the use of these skills evolve in children?
  - At what age do children show adult-like maturity in deploying these skills in conversations?
- Is it possible to leverage ML models as tools for automatically annotating data related to any of the above mentioned skills?
- Could we use ML models to model any of the conversational skills from the child's perspective as well as the caregiver's perspective?
  - Can the models make effective use of multi-modal information for modeling these phenomena?

  Since we are focusing on three different levels of communicative coordination, we were only partly able to address all the above questions. From a developmental perspective, we only studied the development of coherence in children between the ages of 20-32 months. In terms of adult-like maturity in deploying conversational skills, we studied turn-taking management and backchanneling in children. In the case of utilising multi-modal information for modeling these skills, we only leveraged this information for one of the turn-taking studies.

Main contributions In this thesis, we study communicative coordination in child-caregiver interactions with the help of ML models on three levels: **(i) Turn-taking management**, **(ii) Dialog coherence** and **(iii) Conversational Grounding**.

The main contributions of this thesis are as follows:

- Theoretical contributions:
  - Evidence of adult-like behavior by children in their middle childhood in terms of taking turns and providing backchannels in a more naturalistic home setting.
  - Further support for the claim that children acquire the notion of turn-taking fairly early in their childhood.

  - Proof of the increasing developmental trajectory of the degree of coherence in children from the ages of 20-32 months.
  - A map of children's communicative intent use along with various patterns of their meaning coordination in early child-caregiver interactions.
  - Evidence of the challenging nature of the task of identifying repair opportunities in child-caregiver interactions.
  - An approach for conducting large-scale, bottom up investigations for developing ecologically sound theories of communicative coordination and its development.

- Methodological contributions:
  - Manual annotations for semantic coherence for a subset of the New England corpus (Snow et al., 1996).
  - Models trained to automatically annotate utterances for semantic coherence in child-caregiver interactions where the children are between the ages of 20-32 months.
  - Manual annotations for repair opportunities in ChiCa corpus (D. E. Goumri et al., 2024).

Through the studies conducted as a part of this thesis, one of our goals was to showcase the potential of using ML models to study the development of conversational skills in children as a means of inspiring other researchers working on the same topic to consider leveraging these models to conduct ecologically valid large scale studies.

Thesis Structure Broadly speaking, this thesis is divided into four different parts; each part corresponds to one of the levels of communicative coordination that we would like to study and the final part is dedicated to concluding this thesis.

In Chapter 1, we describe the three levels of coordination we focus on in our study of communicative coordination. We describe some of the existing studies and discuss the shortcomings in them to come up with potential avenues of research to overcome these shortcomings. Some of these avenues of research are pursued and described in detail in the following chapters.

Chapters 2 and 3 form the first part of this thesis wherein we study turn-taking in child-caregiver interactions. In Chapter 2, we take an existing Natural Language Processing (NLP) model and train it to predict whether the listener is going to take their turn, do a backchannel or continue to listen in the next time frame given the social cues from the speaker leading up to the next time frame. These cues are multimodal in nature comprising of acoustic, visual and verbal cues based on existing literature. We also compare turn-taking in child-caregiver interaction with turn-taking in adult-caregiver interaction. Chapter 3 looks at training an auto-regressive language model to predict Transition Relevant Places (TRP) in adult utterances. We then utilize this fine-tuned model to predict TRPs in caregiver utterances in a corpus of child-caregiver

interactions and analyze the relationship between TRPs and various facets of the interaction.

Chapters 4 and 5 form the second part of this thesis wherein we study how children coordinate the meaning and intent of their utterances in child-caregiver interactions. In Chapter 4, we first annotated a corpus of child-caregiver interactions for semantic coherence and then developed a machine learning based tool to automate this process. We then used this tool to annotate several English language corpora present in CHILDES (MacWhinney, 2000) and show that our findings from the much larger automatic annotations pertaining to semantic coherence and its development hold similar to the findings from the manually annotated corpus. In Chapter 5, we used the tool developed in the previous chapter and another tool to analyze multiple corpora for communicative intents and semantic coherence thereby enabling us to study linguistic coordination and its emergence in infancy.

Chapter 6 forms the third part of this thesis wherein we performed an exploratory study to identify the frequency with which caregivers seize the opportunity to repair a child's misunderstanding from amongst all possible repair opportunities in child-caregiver interactions. We also judged the capacity of popular large language models (LLMs) in identifying repair opportunities as compared to humans.

Finally, in the last part, we conclude this thesis with a discussion of all the insights that we gained from our studies and the future strands of research that our studies unlock to further enable the study of communicative coordination in child-caregiver interactions.

1. Background and Related Work

Table of contents

1.1. Turn-taking management

Turn-taking is a complex mechanism (as illustrated in Figure 1.1) of coordinating a conversation which was first described in detail in the literature surrounding conversation analysis (Ford & Thompson, 1996; LERNER, 2003; Sacks et al., 1974; Schegloff, 2000). It is the mechanism by which humans decide whose turn it is to speak, which interlocutors need to listen and who the next speaker shall be. The absence of turn-taking would result in chaotic conversations where multiple people will try to speak at the same time and the informative content of the speech would be lost in the ensuing overlapping noise and chatter.

Turn-taking isn't restricted to just a particular language or even humans for that matter; it is a universal mechanism found across various languages, cultures and even across species (Pika et al., 2018; Stivers et al., 2009). In humans, this mechanism is characterized by very short gaps in between turns (average periods of silence in between turns are approximately 200 ms) and very few overlaps in speech after which one of the interlocutors yields to the other (Stivers et al., 2009). An unstated norm of human conversation is that only one person talks at a time (overlapping speech can occur but one of the speakers usually yields the floor) although the speakers may change throughout the conversation.

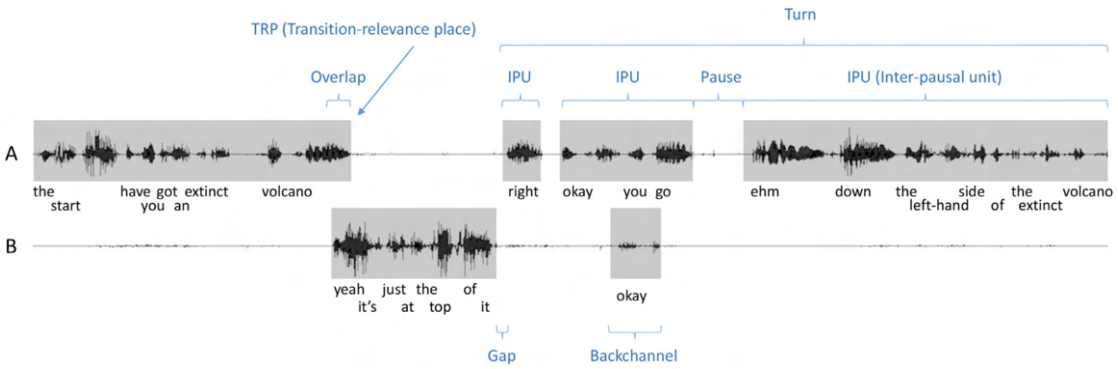

Figure 1.1.: Example of the complex nature of turn-taking as demonstrated by Skantze (2021). The annotated data is from the Map Task corpus (Anderson et al., 1991) where interlocutor A is describing some route on a map to interlocutor B. The figure shows that the gap between turns is very small or even non-existent. An example of non-existent gap is when interlocutor B starts speaking even before interlocutor A has completed their turn resulting in a small overlap. It also demonstrates how a single turn from interlocutor A can contain several Inter-Pausal Units (IPUs) with multiple pauses in between without yielding the floor. Interlocutor B also does a verbal backchannel by saying "okay" during interlocutor A's turn. A backchannel isn't considered as a separate turn or interruption.

1.1.1. Standard model of Turn-Taking

One of the first turn-taking models that was proposed and is now considered as the standard model was described by Sacks et al. (1974). It consists of two components and a set of rules which help to decide the interlocutor to whom the turn should be allocated. The first component helps to describe the various units involved in the construction of a turn (e.g. words, clauses and sentences) and is known as the Turn-Constructional component. The various units involved in constructing a turn are also termed as Turn-Constructional Units (TCUs). The completion of a TCU by a speaker constitutes what is known as a Transition-Relevant Place (TRP). TRPs indicate potential places for a change in the speaker during the conversation. The second component, also known as the Turn-Allocation component, provides various strategies for allocating turns to the next speaker (e.g. by the speaker addressing or gazing at the interlocutor, or by allocating the next turn to the speaker who starts speaking first). The rules govern the turn construction, assignment of the next turn to one of the interlocutors, and the coordination of the turn transfer to minimize overlaps and pauses between the turns. In order, the rules first give preference to the speaker specifically chosen in the prior turn, and lacking that to any other interlocutor who speaks first and lastly, allows the last speaker to continue talking.

1.1.2. Cues useful in identifying TRPs

Various studies have tried to identify the onset of TRPs during a conversation and they found a tendency for it to occur where the boundary of a syntactic unit (e.g. a complete clause ending with a question) coincides with the boundary of a prosodic unit (e.g. rising pitch for the question) (Bögels & Torreira, 2015, 2021; Ford & Thompson, 1996; Selting, 1996). Other studies have also shown evidence of semantic and pragmatic completeness in identifying TRPs (Ford & Thompson, 1996; Magyari & De Ruiter, 2012; Riest et al., 2015; Selting, 2000). Other cues that are useful for determining a switch in turn and the next speaker is the gaze of the speaker as well as bodily gestures (Auer, 2021; Kendon, 1967; Mondada, 2007; Streeck & Hartge, 1992; Zellers et al., 2019). On the other hand, averting gaze from addressee and some manual gestures are a strong cue for holding the turn (Kendrick et al., 2023). A few studies have also shown evidence of visual cues like facial expressions (e.g. smiles or frowns) preceeding the onset of speech (Holler & Levinson, 2019; Kaukomaa et al., 2013, 2014; Nota et al., 2023) which indicates to their potential of acting as visual cues in identifying TRPs. Prosody plays a contentious role in turn taking wherein several studies have shown the use of intonation, pitch, intensity and voice quality as cues for turn taking and backchanneling to varying degrees of success (De Ruiter et al., 2006; Duncan, 1972; Gravano & Hirschberg, 2011; Koiso et al., 1998; N. G. Ward, 2019).

The cues highlighted above, although helpful, are not sufficient to explain the coordination of turn-taking since the time required by a listener to comprehend the cue, process their response and start their turn would take around 600-1500 ms (Levinson & Torreira, 2015). Since the average response time by an interlocutor is

around 200 ms, researchers have concluded that humans predict when the speaker will end their turn and thus start preparing to produce their utterance during their turn (Levinson & Torreira, 2015; Sacks et al., 1974). According to De Ruiter et al. (2006), lexico-syntactic structure is necessary for predicting end of turn while intonation is neither sufficient nor necessary. Bögels and Torreira (2015) found that interlocutors use prosodic cues to determine end of turns. Overall, it seems that humans use both turn-final cues and some predictive mechanism to identify the end of turns. The debate on the redundant nature as well as the complementary nature of these cues is an ongoing one and more work needs to be done to tease the nature of these cues apart.

1.1.3. Backchannels

It often happens that a listener produces short verbal utterances like "okay", "mm hm", "yeah", etc. to indicate to the speaker their degree of interest in the conversation, their continued attention, their degree of understanding, and degree of skepticism amongst other things (N. Ward, 2004). These signals — also known as backchannels — aren't full turns but rather provide some feedback to the speaker indicating them to continue their turn (Schegloff, 2000; Yngve, 1970). Backchannel signals can also be non-verbal in nature (e.g. head nods, eye-blinks or frowning eyebrows) (Hömke et al., 2018; Lutzenberger et al., 2024). Backchannels or interjections as Dingemanse (2024) addresses them, form the glue that holds a conversation together. Experimental studies have shown the importance of backchanneling by the listener during storytelling (Bavelas et al., 2000), and backchanneling by the caregiver on children's vocabulary growth and narrative skills (Newport et al., 1977; Peterson et al., 1999; Tolins et al., 2017).

Much like turn taking cues, it is thought that speakers provide some cues to invite backchannels from the listener as evidence of their attention and understanding of the speaker's speech. These cues might be a little different from the turn taking cues. Studies have shown that backchannels are preceded by a region of low pitch in Japanese (N. Ward & Tsukahara, 2000) and higher pitch along with higher intensity in English (Gravano & Hirschberg, 2011). The speaker directing their gaze to the listener has also found to have elicited backchannel response by the listener (Bavelas et al., 2002).

1.1.4. Development of Turn-taking in Children

Turn-taking has also been studied in child — even infant — and caregiver conversations. In the early infancy stage, the turns produced by the infant are mere vocalizations without any actual language. This stage is also addressed as the proto-turn-taking stage. Studies found that when infants are around 3 months of age, they rhythmically alternated their vocalizations with that of the caregiver with an average gap of approximately 1.5 seconds in between turns and that this gap came down to approximately

800 milli-seconds as the infant grew older but was still in the pre-linguistic age range (Bateson, 1975; Beebe et al., 1988; Jasnow & Feldstein, 1986).

Other studies have shown evidence of a slowdown in the reduction of the gap and then an increase in between turn transitions upto approximately 1.5 seconds as the infant grows older into toddler-hood and starts comprehending language (Garvey & Berninger, 1981; Hilbrink et al., 2015; Stivers et al., 2018). One possible cause for the increase in this gap is that the child needs to integrate both their language production and comprehension system as well as their interactive skills (Ervin-Tripp, 1979; Hilbrink et al., 2015). Around 3 years of age, children are able to use prosody to predict turn transitions between speakers (Casillas & Frank, 2013; Keitel et al., 2013; Tice & Henetz, 2011). According to Garvey and Berninger (1981) the gap duration remained at around 1 second even for 5 year old kids. Isaacs (2013) posit that turn taking is difficult for children below the age of 5 years due to a lack of inhibitory control and the misguided belief that their turn to speak will never come. There is conflicting evidence about the period when children attain adult-like maturity when it comes to their turn-taking skills. For instance, on the one hand, there is evidence that children aren't as efficient as adults in terms of the frequency and timing patterns of turn taking until late in middle childhood (Stivers et al., 2018). On the other hand, another study with a small sample size has shown that children occasionally exhibit adult-like behavior in terms of taking turns in multi-party and dyadic conversations at around 6 years of age (Ervin-Tripp, 1979). Casillas et al. (2016) argue that children actually learn turn-taking management early in their life however its planning the response that takes time causing a delay in taking the turn. It has also been found that while conversing with children, adults tend to adapt their behavior to accommodate the delays and irrelevant turns by the children (Dunn & Shatz, 1989; Ervin-Tripp, 1979).

Although we have developmental studies focusing on children's turn-taking skills from a young age, there is lack of empirical evidence on whether children in their middle childhood are as mature as adults when it comes to turn-taking. Most studies are observational in nature considering small data samples. Furthermore, most studies have focused on children's turn-taking and backchanneling skills in isolation; however to have a complete picture of turn coordination we need to study both aspects in parallel. This situation is where we can leverage predictive ML models to model children's turn-taking behavior on a much larger data sample. Instead of relying on observational data, we can use ML models as proxies to understand whether the model can learn turn-taking behavior from data similar to what the child receives in their natural surroundings.

1.1.5. Turn-taking in dialog systems and human-machine interactions

To understand the various ML models that could be used to model turn-taking behavior, we turn to the literature in human-machine interactions (see Castillo-López et al., 2025; Skantze, 2021, for a comprehensive review). Traditional models of turn-taking in spoken dialog systems detect a turn switch between interlocutors by setting

a threshold value for the amount of silence in between turns. These models are also known as silence-based models (Skantze, 2021). If the duration of silence exceeds the threshold then the model considers the user's turn to have concluded. The issue with this workaround approach is that it is hard to determine a good threshold value. Too high a threshold value and it will seem the dialog model is too slow to respond and too low a value will make the dialog system frequently interrupt the user. Another issue with this approach is that humans tend to pause sometimes during their turn either with the same utterance or in between utterances that comprise the same turn and depending on how long the person pauses, the system might easily end up interrupting the user.

Since the previous approach was not informed by the theory on turn-taking and it was more of a heuristic approach, researchers quickly moved on from this method. Next, they tried to first detect the end of Inter-Pausal Units (IPUs) — speech segments bounded by a certain duration of silence — in speaker's speech and then made a decision based on turn-taking cues fed to a machine learning model (e.g. decision tree classifier, LSTM, etc.) to predict a turn-shift (e.g., Masumura et al., 2017; Meena et al., 2014; Sato et al., 2002). Models following this approach are also known as IPU-based models (Skantze, 2021).

In the recent past, a more continuous approach has been taken in predicting turn-taking where a prediction is made at certain time intervals (e.g., every 50 milliseconds) (e.g., Roddy et al., 2018a, 2018b) or on a token by token level (e.g., Ekstedt & Skantze, 2020). For instance, researchers have made use of recurrent neural networks like LSTMs (Long Short Term Memory networks) as a means of encoding linguistic information such as words, part of speech tags (POS tags) and even senones to predict turn-taking and backchannel signaling (Masumura et al., 2018; Roddy et al., 2018a; Skantze, 2017). Going one step further, Ekstedt and Skantze (2020) demonstrated how an auto-regressive language model like GPT-2 (Radford et al., 2019) can make effective use of the preceding context to predict turn shifts in conversation as compared to the LSTM based models.

The most recent approach in detecting turn-taking and backchannels in conversation makes use of Voice Activity Projection (VAP) models (Ekstedt & Skantze, 2022a, 2022b; Inoue et al., 2024a, 2024b, 2025). VAP models in a nutshell are a series of self-supervised models that are trained on the task of predicting the future voice activity of each interlocutor in a conversation.

One of the drawbacks of most computational models of turn taking is that they focus on usually just the acoustic or verbal content in a conversation. However, turn-taking is a multi-modal phenomenon with various multi-modal cues signaling the end of a turn or a backchannel. Sampling information from different modalities and fusing this information together so that the model can make effective use of this multi-modal information is a challenging task and an active area of research within the community.

1.2. Coherence in conversation

In any conversation, the interlocutors cooperate with each other to contribute to the conversation in a relevant fashion (H. P. Grice, 1975). One can always observe that the flow of conversation doesn't usually involve a series of disconnected utterances but there is usually some underlying theme, topic or purpose to it and each interlocutor contributes their utterance to build upon this underlying topic (P. Grice, 1991). This underlying principle forms the basis of having a coherent conversation. Thus, it falls up to every interlocutor to ensure that their utterances are pertinent to the topic being discussed in order to sustain the conversation.

Coherence has been studied by developmental researchers under the guise of many labels such as semantic contingency, coherency, and topic maintenance (among others) (Blain-Brière et al., 2014; Bloom et al., 1976; Capps et al., 1998; Dorval et al., 1984; Hale & Tager-Flusberg, 2005; Keenan & Klein, 1975; Matthews et al., 2018; Rosnay et al., 2014; Slomkowski & Dunn, 1996). Studies have shown that even children in primary school respond incoherently or in a non-contingent manner to the preceding conversational turn which indicates the complexity of this skill (Abbot-Smith et al., 2024; Dorval et al., 1984).

1.2.1. Mechanisms for being coherent

In their seminal study, Benoit (1979) describes three mechanisms that allow an interaction to be coherent: **(i) Formal coherence**, **(ii) Structural coherence**, and **(iii) Topical coherence**.

Formal coherence In this mechanism, the interlocutor repeats some form of structure or lexical terms from the previous interlocutor's turn. A repetition (partial or in its entirety) is coherent as it formally links the utterances. Consider the below example:

> Jane: *Is that a duck?*
> Jack: *Duck goes quack quack.*

In the above example, Jack repeated the term "duck" and then expanded upon the previous turn by adding to it making it coherent. Formal coherence is thus also a means to lexical entrainment from the interactive alignment perspective of communicative coordination (Pickering & Garrod, 2004).

Structural coherence One of the primary tenets in the field of conversation analysis is that adjacency pairs are the basic structural unit of conversation (Sacks, Jefferson, et al., 1995; Schegloff & Sacks, 1973). Since the two pairs in an adjacency pair are sequentially linked, there needs to be a pattern in which they occur i.e. the second pair has to follow the first pair. This sequential nature forms the basis of structural

coherence in that whenever a first part of an adjacency pair is produced, it starts a sequence where the following utterance forms the appropriate second pair of the adjacency pair. Some examples of adjacency pairs are question-answer, request-response and greeting-greeting among other pairs. Consider the following example:

Jane: *What would you like to eat?*
Jack: *Let's order some pizza.*

In the above example, the turns by both Jane and Jack together comprise a question-answer adjacency pair. Each part of the adjacency pair is produced by a different speaker and they are produced in a sequence one after the other. Since the first part of the adjacency pair is a question, it naturally produces an expectation that the following part will be an answer to said question. The example showcases how structural coherence can connect discourse.

Topical coherence The underlying principle of this mechanism is that the interlocutors build upon some conversational topic that is mutually agreed upon. Topical coherence takes into account the whole conversational discourse. Context is key when it comes to analyzing it. Consider the below example:

Jane: *Are you coming to Will's party?*
Jack: *Yeah it's going to be rad.*
Jane: *Are you bringing a date to the party?*
Jack: *Let's go grab a bite to eat.*

In the above example, the first utterance from Jack is topically coherent but the second utterance is incoherent as both the interlocutors were previously discussing Will's party when Jack suddenly talks about having some food to eat.

1.2.2. Role of communicative intents in being coherent

Whenever an interlocutor says something, they have some intent in mind which they wish to communicate towards their listener which in turn lends their utterance its particular meaning (H. P. Grice, 1957). Communicative intents are hard to analyze given their abstract nature and that they can only be inferred from the surrounding context. The context involves the communicative sequence between the interlocutors, their shared belief space as well as their shared communicative environment. Developmental researchers have cataloged various inventories of children's communicative intents in child-caregiver interactions. These inventories vary in their granularity going from coarse categories like "declaratives" or "imperative" intents (Bates et al., 1975; Dore, 1973) to more fine-grained ones like "disagree with proposition", "yes/no question", and "promise", as the child matures in their use of expressive language (Ninio et al., 1994; Snow et al., 1996). To analyze the coordination of intents we need to take into

account not just the relation of the intent of the speaker in relation to the listener but also in relation to the context of the ongoing communication. This is where the theory of adjacency pairs from the field of conversation analysis comes in handy. With the help of adjacency pairs, we can identify coordination on a coarse level of intents such as greeting-greeting, question-answer, and request-acceptance/refusal (Schegloff, 1986). Thus, communicative intents can help in analyzing if the conversation is *structurally coherent* or not.

When it comes to early child-caregiver interactions, some of these adjacency pairs like question-answer are extremely useful for analysis since they are frequently made use of in early interactions and have been known to provide a window into early conversational development (Chouinard et al., 2007; Peirolo et al., 2024; Stivers et al., 2018). However, many of the intents cataloged in the communicative inventory like "promise" or "express a wish" don't fall into any adjacency pair and so we must also take the topical coherence into account to ensure the coordination of the child's intent. Previous studies have considered the contingency of the child's expression of intent on the caregiver's previous message and the context of the conversation (Abbot-Smith et al., 2023; Bloom et al., 1976). A study also found that children seem to find introducing new abstract topics into the conversation to be rather difficult; it becomes easier if the topic is physically situated in their environment (Keenan & Klein, 1975).

1.2.3. Role of coherent interactions in child development

The notion of coherence or 'contingency' as is known in developmental parlance changes as the child ages (Reed et al., 2016). Starting as synchrony in behavior and its corresponding effect in infants and caregivers in the postnatal period (e.g., Feldman, 2015), to caregivers responding in real time to behaviors of infants around 6 months of age (e.g., Bornstein et al., 2008), to shared attention and proto-conversations when the infant is aged in between 6-12 months (e.g., Scaife & Bruner, 1975; Snow, 1977), to fluid conversations when the child is around 2 years of age (e.g., Gilkerson et al., 2017; Hirsh-Pasek et al., 2015).

Contingent interactions have been shown to play an important role in aiding language learning in children for instance by improving syntactic rule learning (Ferguson & Lew-Williams, 2016) and by improving vocabulary growth (Bornstein et al., 1999; Donnellan et al., 2020; McGillion et al., 2017). Studies also show that difficulties in having coherent interactions can lead to difficulties in making new friends (Hazen & Black, 1989; McGuinness et al., 2023; Place & Becker, 1991), difficulties in relationships (Arkowitz et al., 1975; Miczo et al., 2001), difficulties in the workplace (Garzaniti et al., 2011) and shunning by your peers (Putallaz & Gottman, 1981; Wolters et al., 2014). These difficulties in turn can affect the child's mental health and cause behavioral problems down the line (Helland et al., 2014; Ketelaars et al., 2010).

Thus, while developmental studies have shown the important link between language learning and contingent interactions with the caregiver, the development of a child's ability to be coherent hasn't received as much attention as needed. We don't have enough empirical evidence on the period around which children achieve adult-like

maturity in being coherent in their conversations; nor do we have enough evidence of how children engage in verbal coherent interactions with their caregiver and how these interactions change over time as the child ages.

1.2.4. Machine learning for evaluating coherence

One area of inquiry analyzing the coherence of conversations is in the field of dialog system evaluations which evaluates whether the response of a dialog agent in conversation with a human is coherent or not among other things. Computational models pre-dating the transformer (Vaswani et al., 2017) era of modern efficient neural networks utilized various cues like repetitions of noun phrases across turns (focus coherence), speech acts and adjacency pairs (structure coherence), and measures of turn-similarity (topic coherence) to estimate the coherence of dialog agents (e.g., Barzilay & Lapata, 2008; Cervone et al., 2018; Yi et al., 2019). With the advent of pre-trained language models that captured rich linguistic knowledge by training for hours on end on massive amounts of data, researchers turned towards these models to evaluate dialog agents and observed significant improvements in terms of similarity with human judgments as compared to the previous approaches (Mehri & Eskenazi, 2020; Mehri et al., 2022; Pang et al., 2020; Sai et al., 2020). Recent efforts undertaken by some members of the NLP community have provided tools for automatically classifying utterances in terms of the communicative intents they express thereby allowing large scale analysis of structural coherence in conversations (Kumar et al., 2018; Mezza et al., 2018; Nikolaus et al., 2022).

1.3. Conversational grounding and the development of common ground

Conversational grounding is the ability to negotiate and coordinate the shared beliefs and knowledge between the participants of a conversation (H. H. Clark, 1996; H. H. Clark & Schaefer, 1989; Stalnaker, 1978). The mutual knowledge between the interlocutors is also known as the *common ground* between them. Every conversation starts with a certain foundation of common ground that can be attributed to any past interactions between the interlocutors or barring that to some shared culture or social space (Baker et al., 1999; H. H. Clark, 1996).

Consider the below example from Apriliani and Muslim (2021) which demonstrates the relevance of common ground in understanding a conversation. The example involves the following scene (Figure 1.2) and dialog from the classic Hollywood film *Titanic* directed by James Cameron and released in 1997.

> Rose: *I love you, Jack.*
> Jack: *Don't you do that. Don't you say your goodbyes, not yet. Do you understand me?*

Figure 1.2.: Depiction of a scene from the movie *Titanic* by James Cameron. It shows Jack and Rose clinging to a floating piece of wood while drifting in the icy waters of the Atlantic Ocean.

An individual who hasn't watched the movie could be forgiven for being extremely confused after reading Jack's response to Rose in the above dialog. After all a normal response to someone saying "I love you" would most probably be "I love you too". However, if the individual had knowledge of the movie then they would have known that this dialog occurs during the scene where Rose and Jack are floating on the wooden debris in the Atlantic Ocean. They would have taken into account that Rose was close to succumbing to the cold; that her declaration of love was more likely to be a farewell and that Jack's reply was an exhortation to Rose to not give up hope yet. Thus, having access to the knowledge that Jack and Rose shared in that moment was critical to understanding their interaction.

To develop the common ground, one of the interlocutors will present some information through an utterance and the listener needs to indicate via some feedback signal that the information has been understood as the speaker intended (H. H. Clark & Schaefer, 1989). Thus, grounding is a collaborative process which requires all participants in the conversation to make an effort in grounding the information. The feedback signal can take the form of positive evidence (e.g., backchannels) or negative evidence (e.g., clarification requests) (H. H. Clark & Brennan, 1991). The degree of understanding that needs to be expressed by the listener to indicate to the speaker that

they have understood their utterance is determined by what is called the "grounding criterion" in the literature (H. H. Clark & Schaefer, 1989). The grounding criterion depends upon the task at hand and the goals of the various participants of the conversation.

Conversational grounding as a phenomenon is hard to study due to the inherently complex nature and dynamics of any conversation or dialog. The cues that interlocutors provide as evidence of grounding are not always explicit or verbal but are often implicit and even non-verbal (e.g., gestures like head nods, frowns etc.). It is hard to quantify and measure what information has been grounded and what has yet to be grounded successfully as it isn't always sequential in nature. For instance, consider the following snippet of conversation between a child and their caregiver:

Child: *What is it used for?*
Caregiver: *To travel very far, very fast.*
Child: *Can it fly?*
Caregiver: *Yes!*
Child: *Is it a hot air balloon?*

Here, one would assume that the caregiver's first utterance was successfully added to the common ground as evidenced by the child initiating a relevant next turn. However, the child's final utterance gives evidence to the contrary as a hot air balloon can't travel far or fast.

1.3.1. Common ground and language acquisition

Developmental studies have shown the importance of common ground going right from early non-verbal communication to covering the entire period of language acquisition in children (Bruner, 1985; E. V. Clark, 2015; Tomasello, 2010; Tomasello et al., 2007). Considering the importance accorded to caregivers' feedback on children's production as a means of helping them improve their language acquisition (e.g., E. V. Clark, 2018, 2020; Nikolaus & Fourtassi, 2023), there is a significant oversight of research on this topic. E. V. Clark (2020) argues that repairs which are a key mechanism for developing the common ground, are also important for language acquisition with self-repairs making up a majority of the total repairs (in terms of frequency) as compared to other-repairs.

Furthermore, despite the numerous studies considering the role of the caregiver in shaping the child's understanding in a situated context and in presence of the target (see review in Çetinçelik et al., 2021), there hasn't been a lot of empirical work done to study this phenomenon when the target is abstract or isn't in the field of view of the child or the caregiver.

1.3.2. Repairs as a means of developing common ground

One of the principle forms of developing common ground is that of *repairs* whenever there is a breakdown in communication (H. H. Clark & Krych, 2004; Dingemanse et al., 2015; Fusaroli et al., 2017; Purver et al., 2018). While talking, a speaker is always trying to look for signs of attention and understanding from their listeners. If there is a sign of misunderstanding by the listener as evidenced by some feedback signal (e.g., frown or confused look) or by the listener issuing a clarification request, then the speaker can attempt to repair this misunderstanding thereby cementing this information into their shared common ground. Studies in conversation analysis (e.g., Schegloff, 1987, 1992; Schegloff et al., 1977) point out three cornerstones of any repair mechanism: i) the role of the person initiating the repair in the conversation, if it is the speaker of the problematic utterance ('self') or if it is the listener ('other'), ii) who makes the actual repair to the utterance, if it is a 'self-repair' or 'other-repair' and iii) the position of the repair in the sequence of utterances; if it takes place in the same utterance (e.g., self-reformulation), or some later utterance in the dialog.

Other repairs can be one of three types (Dingemanse et al., 2015), namely: *open requests*, *restricted requests* and *restricted offers*. Open requests don't point out the issue in the speaker's utterance (e.g., "huh?"), restricted requests point out the exact issue that needs to be repaired (e.g., "who?"), and restricted offers provide an alternate construction for the speaker's utterance and ask for their confirmation (e.g., "you mean he was the last person to arrive?").

1.3.3. Conversational grounding in human-machine interactions

Conversational grounding is an important aspect of communicative coordination that is usually missing from most modern day dialog agents and systems designed by the field of human-machine interaction. These dialog systems are error prone and often fail to repair any misunderstandings that they might have caused in a dialog with a human user. Recently, large language models' (LLMs) limitations in conversational grounding abilities have drawn the eye of NLP researchers (Benotti & Blackburn, 2021; Chandu et al., 2021; Shaikh et al., 2024). For instance, studies show that LLMs struggle to understand when utterances are implicitly grounded (e.g., Jokinen et al., 2024) and that if conversational agents are augmented with theory-of-mind modeling, it improves their capacity to align with the speaker and helps in negotiating their common ground (e.g., Qiu et al., 2024). Cheng et al. (2024) explore the use of a multimodal transformer model to predict uncertainty in young children engaged in a counting task and finds a potential for improvement. The uncertainty of an interlocutor during a conversation is evidence of their understanding, which ties in directly to their shared common ground. Benotti and Blackburn (2021) raise an important concern that the way current LLMs and dialog models interact with humans can be misleading at times as they build false expectations of their common ground with their interlocutors. This, in turn, leads to a rise in misunderstandings,

which can be frustrating for humans attempting to converse with dialogue models.

One of the avenues of studying grounding is to consider using LLMs to evaluate grounding acts in a conversation as proposed by Traum (Traum & Allen, 1992). LLMs tend to struggle to classify grounding acts in a conversation as well as generate them and their classification capabilities are directly linked to the number of parameters of the model and the size of it's pre-training data (Mohapatra, Hassan, et al., 2024; Mohapatra, Kapadnis, et al., 2024; Shaikh et al., 2024). Various forms of grounding have been studied by considering several grounded language tasks like reference games (for e.g., Golland et al., 2010; Kennington & Schlangen, 2015; Monroe et al., 2017) and goal-oriented dialog tasks (for e.g., Das et al., 2017; De Vries et al., 2017; Haber et al., 2019; Kim et al., 2019; Narayan-Chen et al., 2019; Udagawa & Aizawa, 2019) amongst other things (see Chandu et al. (2021) for a non-exhaustive list of tasks). However, the issue with most of these tasks and the phenomena that the researchers are trying to model with these tasks is not really conversational grounding and it is restricted to either referential grounding or grounding in terms of a particular modality (e.g., visual grounding) (Chandu et al., 2021; Hakimov et al., 2025; Ilinykh et al., 2019; Jeknic et al., 2024).

The majority of the studies involved in studying grounding in LLMs for the purpose of developing better conversational agents focus on interactions of the agents with adults or they study adult-adult interactions to gain a better understanding of grounding. What is missing in the literature is how the agents ground information when interacting with children and how the common ground develops in child-caregiver interactions. Studies have focused on generating and responding to clarification requests in dialog systems (e.g., Purver, 2004; Rieser & Moore, 2005; Rodríguez & Schlangen, 2004, among others); however, the existing systems are still limited in their functionality when it comes to clarifiction requests.

Part I.

Turn-taking management

Table of contents

2. Turn Coordination in Middle Childhood

This chapter is based on the article "Development of Multimodal Turn Coordination in Conversations: Evidence for Adult-like behavior in Middle Childhood" (Agrawal et al., 2023), published in the *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

The review of studies in Section 1.1 showed that developmental researchers have primarily focused on turn taking in infancy and early childhood. However, there is little evidence on children's turn taking capabilities in middle childhood; are they still lacking in this skill or are they as capable as mature adults? In addition to this, most studies look at turn-taking and backchannel production independently from each other which we believe isn't conducive if we are trying to look at the bigger picture of the development and usage of these skills.

In this study, we jointly model the ability of the child and also the caregiver to take a turn as well as provide feedback in the form of backchannels in child-caregiver interactions. Here we use ML as a model of the child and caregiver's turn taking coordination capabilities. We train multimodal models on immediately preceeding communicative cues from one interlocutor to predict whether the other interlocutor will take their turn or provide some feedback and continue their role as a listener across various child-caregiver and caregiver-adult dyads.

We confirmed previous findings that children in their middle-childhood show adult-like behavior in terms of providing feedback as backchannels. We also extended this finding by showing evidence of adult-like turn-taking behavior in children in their middle-childhood. Based on our results, we also propose that there are specific communicative cues distinguishing taking the turn and providing a backchannel which both children and adults react to.

2.1. 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 et al., 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 (Cathcart et al., 2003; Duncan, 1972; Ford & Thompson, 1996; Gravano & Hirschberg, 2011; Sacks et al., 1974; Skantze, 2021; N. Ward & Tsukahara, 2000).

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 et al., 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; Park et al., 2017; Peterson, 1990).

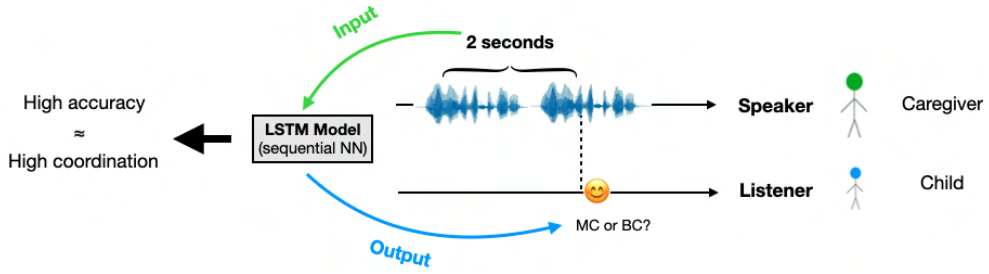

Figure 2.1.: Schematic illustration of how we characterize coordination between two interlocutors (here a child-caregiver dyad). We train a model to predict the timing of the child's conversational move (Main channel or Back channel) based on the caregiver's immediately preceding communicative signals (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.

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 (Baines & Howe, 2010; Hess & Johnston, 1988; Maroni et al., 2008), whereas others have found them to already show adult-like behavior in some specific aspects (e.g., Bodur et al., 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 J. Liu et al. (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 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 2.1).

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 two-fold: (i) the performance of a model (as illustrated in Figure 2.1) 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., Fusaroli et al., 2023a; H. Jiang et al., 2022; Misiek & Fourtassi, 2022; Snow, 1977).

2.2. 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.

2.2.1. Conversational dataset

We use the ChiCo corpus (Bodur et al., 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. The caregivers were instructed to pick a word from a pre-determined list of words whereas children had complete freedom to pick whatever word they wished to choose. 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.

[1] Using Zoom software

2.2.2. Characterization of MC and BC

2.2.2.1. 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 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.

2.2.2.2. 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 2.1. Plots for the latencies in turn-taking in the two dyads are available in the Appendix A.

2.2.3. 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).

2.2.3.1. 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) (Brunner, 1979; Duncan, 1972; Kendon, 1967; Paggio & Navarretta, 2013; Park et al., 2017). 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).

2.2.3.2. Vocal Cues

For the vocal cues, we use the features extracted by J. 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 (Goswami et al., 2020; Jain & Leekha, 2021; Morency et al., 2010; Murray et al., 2022; Ruede et al., 2017). The categories of cues we used are pitch (variation), Mel-Frequency Cepstral Coefficients (MFCC), voice quality, energy, and pausal information.

2.2.3.3. Verbal Cues

For the textual features, we relied on the Part-Of-Speech (POS) tags extracted by J. 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.

2.2.4. 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).

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

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 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).

2.2.5. 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

[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 can be found in Appendix A.

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

did a feature ablation study by considering only the set of inviting cues belonging to a particular modality, one at a time. Table 2.1 describes the size of data used (in terms of frames) in each experiment. The results for additional experiments involving cross-dyadic settings and individual feature ablation studies can be found in the Appendix A.

2.2.5.1. Experiment 1: BC vs. random non-BC

In this set of experiments, our goal was to replicate the results reported by J. 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 interlocutor'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."

2.2.5.2. 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).

2.2.5.3. 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 2.1).

Table 2.1.: 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

2.3. Results

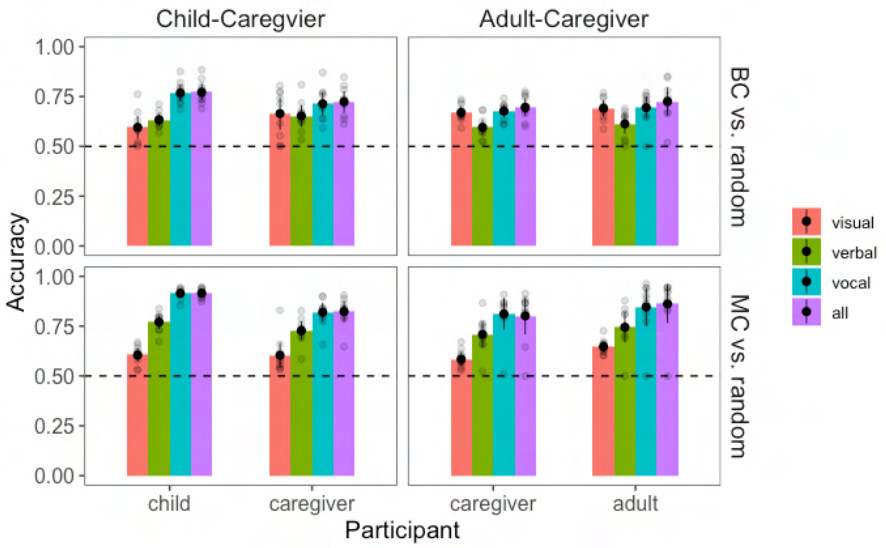

Figure 2.2.: Accuracy scores of the BC predicting models (Experiment 1, top) and the MC predicting models (Experiment 2, bottom) both in the child-caregiver condition (left) and the adult-caregiver condition (right). In each case, we show the results of the modality-specific models (i.e., the models 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.

Experiment 1 and 2: Figure 2.2 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 J. 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 2.3 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. 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.

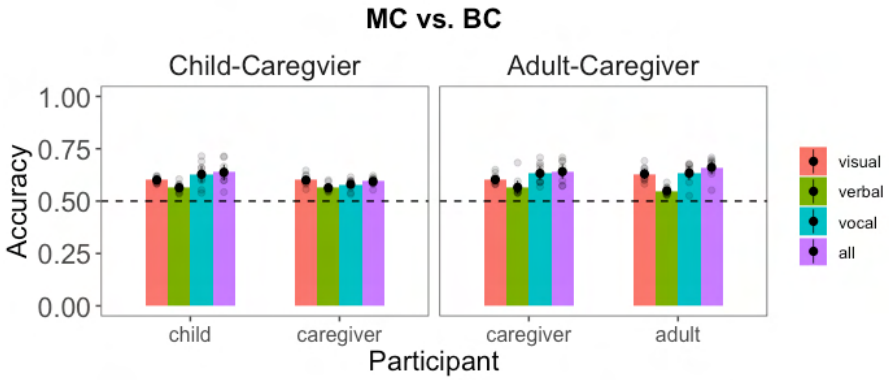

Figure 2.3.: Accuracy scores of the BC vs. MC predicting models (Experiment 3) both in the child-caregiver condition (left) and the adult-caregiver condition (right). In each case, we show the results of the modality-specific models (i.e., the models using predictors from a single modality) and the full 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.

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

2.4. Discussion

This part of the thesis 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 (Bodur et al., 2023; Hess & Johnston, 1988; Nguyen et al., 2022; Park et al., 2017; Peterson, 1990), 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 J. 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 J. 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]

2.4.1. 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, for example, 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 et al., 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).

3. Identifying TRPs in Child-Caregiver Interactions

In this chapter, we take a step back and examine one of the basic principles of turn-coordination i.e., Transition Relevant Places (TRP) in conversation (Sacks et al., 1974). TRPs indicate turn switch possibilities in the conversation and hence the first step in learning turn-taking coordination for a child would be to develop the capacity to identify these positions during the caregiver's turn. Previous studies have identified several cues that seem to help in identifying TRPs in conversation. Out of these cues, we choose the syntactic and pragmatic completeness of the utterance for our study since they are known to be strong predictors of turn-shifts (Ford & Thompson, 1996).

Here, we train an auto-regressive transformer based model (TurnGPT) (Ekstedt & Skantze, 2020) to predict TRPs in adult utterances. We use this trained model to then predict TRPs in all the caregiver utterances in the New England corpus (Snow et al., 1996). The major result in our study was that there was a significantly strong positive correlation between a TRP at the end of the caregiver's turn and the child being the speaker of the following utterance. This implies that children already between the ages of 14-32 months (the age range of the child in the New England corpus) wait for the caregiver to complete their turn before speaking.

3.1. Introduction

The ability to coordinate their turn in a conversation is one of the fundamental skills a child needs to develop in order to hold any meaningful conversation. There is an order and a system behind turn-taking otherwise conversations would appear to be chaotic and unnatural (Sacks et al., 1974). As described in the standard model of turn-taking (Sacks et al., 1974), there are certain places during a conversation — usually after the completion of a syntactic unit (e.g., phrase, clause and sentence) — where a potential turn-switch may occur. These positions are known as *Transition Relevant Places* (TRPs). The presence of a TRP during any moment of a conversation, doesn't necessarily mean that a turn-shift will occur; the speaker may choose to hold on to their turn for a while longer. Being able to recognize TRPs thus becomes an important first step in learning turn-coordination.

There is a vast literature focusing on the various cues that can be used by a listener to identify an oncoming TRP. For instance, there is the syntactic and pragmatic completeness of the utterance (Bögels & Torreira, 2015; Ford & Thompson, 1996; Selting, 1996), gaze of the speaker (Auer, 2021; Kendon, 1967), gestures (Mondada, 2007; Streeck & Hartge, 1992; Zellers et al., 2019) and prosody (Gravano & Hirschberg, 2011; N. G. Ward, 2019) among other cues. From amongst these cues, from a linguistic point of view, Ford and Thompson (1996) point out the usefulness of syntactic and pragmatic completion in identifying what they define as "Complex Transition Relevant Places" (CTRPs). According to them, an utterance is syntactically complete when it can be considered as a full clause i.e., it has a predicate or given the surrounding context of the utterance, the predicate is implied based on just the grammar and structure of the utterance.

On the other hand, their definition of pragmatic completeness simply states that the utterance needs to complete some conversational action given some conversational context. Consider the following example from Ford and Thompson (1996):

> V: *and he said we'll probably have to put an artificial knee in/ in five years/.>*
> V: *For my Dad/.>*
> C: *hmm/.>*
> V: *Because his knees is is deteriorating/ and weak/.>*

In the above example, the '/' represents points of syntactic completion and the '>' symbol represents points of pragmatic completion in the utterance.

Previous studies have leveraged part-of-speech (POS) tags (Agrawal et al., 2023; Gravano & Hirschberg, 2011; Johansson & Skantze, 2015; Meena et al., 2014; Skantze, 2017), words (Roddy et al., 2018a) or senones (Masumura et al., 2018) in predicting turn-shifts in conversation. Ekstedt and Skantze (2020) make use of an auto-regressive transformer based model (TurnGPT) which they argue takes into account the syntactic and pragmatic completeness of the utterance by taking into account the surrounding context to predict a turn-shift.

Developmental studies in the past have mainly focused on the response latency of children when considering the development of turn-taking abilities in children (see

Nguyen et al., 2022). Casillas et al. (2016) argue that children learn turn-taking early in their childhood and that it is difficulties in planning their response which causes the delays in timing their turns quickly.

The goal of this study is to look at how linguistic cues like syntactic and pragmatic completeness of an utterance can can help improve our understanding of turn-taking in child-caregiver interactions. For this purpose, we would like to analyse the presence of TRPs in child-caregiver conversations. We can observe turn-shifts in the data but TRPs are hard to observe in the data since an interlocutor doesn't necessarily switch turns at a TRP.

One of the ideas proposed in the literature is that TRPs have a probabilistic nature in that a higher probability of a TRP implies a stronger likelihood of it being an actual turn-shift (Ekstedt & Skantze, 2020). Thus, we fine-tune the TurnGPT model (Ekstedt & Skantze, 2020) on adult-adult dialogs and then use the fine-tuned model to predict TRPs in caregiver turns in a corpus of child-caregiver interactions. With this approach, we can't use the model to predict TRPs for the child turns because young children's language is quite different compared to adults and so the model would need to be specifically fine-tuned on children's conversational data for it to be accurately able to predict the TRPs. Instead, we take a different approach of comparing the occurrence of TRPs in adult-adult dialogs and child-caregiver dialogs. Furthermore, we analyse the effect of the following factors on the prediction of a TRP at the end of the utterance by a caregiver: (i) length of the utterance and (ii) age of the child. We also analyse whether a TRP at the end of the caregiver's utterance predicts an actual turn-shift in the conversation. Our aim is to analyze if a complete turn by the parent elicits a response by the child or whether the child interrupts the parent midway through their turn. Our hypothesis is that if children learn turn-taking early on in their childhood then they would wait for their caregivers to finish their turn before speaking themselves.

3.2. Methodology

3.2.1. Data

For fine-tuning our model, we make use of the DailyDialog (Li et al., 2017) and MetaLWOz (Lee et al., 2019) datasets. The DailyDialog dataset consists of everyday conversations about the daily life of the interlocutors participating in the conversation and it has about 13k dialogues in total. The MetaLWOz dataset on the other hand, contains about 38k task-oriented dialogs collected in *Wizard of Oz* style setup. Each dataset comes with clearly defined turns with the average number of turns per dialog being 7.9 and 11.4 respectively. Once the model has been fine-tuned, we perform all our analysis with the model on the New England corpus (Snow et al., 1996). The corpus contains records of conversations between $N = 52$ children and their caregivers at 14, 20 and 32 months of age.

3.2.2. TurnGPT model

Following the success of the TurnGPT model (Ekstedt & Skantze, 2020) in utilizing syntactic and pragmatic completeness of utterances to determine turn shifts, we leverage this model for our study. This model is an auto-regressive pre-trained GPT-2(base) model (Radford et al., 2019) that we fine-tune on the DailyDialog and the MetaLWOz datasets. The unique addition to the original GPT-2 model is the addition of a special *turn-shift* token (<ts>) to denote the end of a turn. Thus, each turn-shift token denotes a Transition-Relevant Place (TRP) in the turn. Upon fine-tuning, the model learns the distribution of TRPs in the data and can be used to predict if the next token in the sequence is going to be a TRP or not.

Corpus	Avg. number of TRPs/Turn	Avg. number of TRPs/Turn normalized by number of words in turn
DailyDialog	2.06	0.18
MetaLWOz	1.57	0.2
New England (14 month old)	0.69	0.21
New England (20 month old)	0.73	0.21
New England (32 month old)	0.81	0.18

Table 3.1.: The average number of TRPs predicted by our fine-tuned model on the test-sets of DailyDialog and MetaLWOz datasets. The table also shows the average number of TRPs predicted by the model for Caregiver utterances to the 14, 20 and 32 month old children in the New England corpus. For the New England corpus, we compute the average over individual utterances and not complete turns because that information isn't available in the corpus.

3.3. Results and Analysis

The results of the fine-tuned model on the test sets of the DailyDialog and MetaLWOz dataset as well as on the New England corpus are shown in Table 3.1. The average number of TRPs predicted by the model per turn after being normalized by the length of the turn is around 0.2 for both the adult-adult conversational datasets. For the New England corpus, instead of predicted TRPs for an entire turn, we predict TRPs for every utterance since the turn information isn't directly available with the corpus. We find the average number of TRPs per utterance after being normalized by the length of the utterance to be close to 0.2 across caregivers responding to 14, 20 and 32 month old children.

3.3.1. Effect of length of utterance

Intuitively, one would imagine that the length of an utterance would influence whether the end of an utterance is a TRP with larger utterances having greater odds of having a TRP at the end. To analyse this, we fit a mixed effects linear model to predict TRP at the end of an utterance as follows:

$$TRP\_at\_end \sim num\_of\_words + (1|transcript) \tag{3.1}$$

We obtain for number of words in the utterance (which indicates the length of the utterance): $\beta = 0.082$, $SE = 0.001$ and $p < 0.001$ indicating a statistically significant positive correlation with the length of the utterance.

3.3.2. Effect of age of the child

Here we analyse the effect of the age of the child on predicting TRP at the end of the caregiver's utterance. We fit another mixed effects linear model as follows:

$$TRP\_at\_end \sim age + (1|transcript) \tag{3.2}$$

We obtain for age: $\beta = 0.003$, $SE = 0.001$ and $p = 0.001$ indicating a slight positive correlation with the the age of the child.

3.3.3. Effect of TRP on actual turn switch

For our final analysis, we consider if the utterance ending with a TRP has an effect on predicting the actual turn switch where the child takes the turn. We fit another mixed effects linear model as follows:

$$next\_is\_child \sim TRP\_at\_end + (1|transcript) \tag{3.3}$$

We obtain for next is child's turn: $\beta = 0.098$, $SE = 0.005$ and $p < 0.001$ indicating a statistically significant positive correlation with the utterance ending in a TRP.

To ensure that for the next utterance is child turn analysis, the length of the utterance isn't playing a confounding factor we fit another mixed linear effects model by adding the utterance length as one of the predictors as follows:

$$next\_is\_child \sim TRP\_at\_end * num\_of\_words + (1|transcript) \tag{3.4}$$

The estimated fixed effects were as follows: TRP_at_end: $\beta = 0.131$, $SE = 0.009$, $p < 0.001$; num_of_words: $\beta = 0.026$, $SE = 0.002$, $p < 0.001$; TRP_at_end:num_of_-words: $\beta = -0.019$, $SE = 0.002$, $p < 0.001$. This shows that even after accounting for the utterance length, next_is_child is strongly correlated to a TRP at the end of utterance.

3.4. Discussion and Conclusion

This work contributes to the ongoing study of turn-taking in child-caregiver interactions by providing an automatic tool for annotating TRPs in caregiver turns in child-caregiver conversations. With our exploratory study, we show how linguistic cues like syntactic and pragmatic completeness can be used to train a model to predict TRPs in caregiver utterances. Once the TRPs have been identified, we can then study the nature of caregiver turns and factors influencing them. We see that the presence of a TRP at the end of caregiver's utterance is strongly correlated with the next utterance coming from the child indicating that children are speaking up only after complete turns by the caregivers and are not interrupting them in between. This is true for children between the ages of 14 and 32 months. We believe this finding of children waiting for the caregiver to complete their turn provides additional evidence supporting the claim that children acquire the notion of turn-taking relatively early on in life (Casillas et al., 2016).

Limitations and future work

This study was restricted to analysing the TRPs in just caregiver utterances. An extension to this study would be to fine-tune a model on the children's utterances where the turn switches are marked with a special symbol and then use it to annotate and analyse another corpus of child-caregiver interactions for children's turns. This studies also just considers the linguistic cues while identifying TRPs whereas the literature on this topic tells us that cues from other modalities like prosody and gaze also help in determining turn shifts. Another adaptation of the model could involve leveraging these additional cues to further refine the prediction of TRPs with the model. For our analysis, we also do not control for temporal latency which could be an important confounding factor since it could be the case that the last utterance in the caregiver's turn could be followed by a longer silence which is in fact what helps predict the child's next turn rather than the TRP.

Part II.

Coherence in conversation

Table of contents

4. Towards Automatic Coding of Semantic Coherence in Child-Caregiver Conversations

This chapter is based on the article "Automatic Coding of Contingency in Child-Caregiver Conversations" (Agrawal et al., 2024), published in the *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

One of the most important communicative skills children have to learn is to engage in meaningful conversations with people around them. At the heart of this learning lies the mastery of being coherent, i.e., the ability to contribute to an ongoing exchange in a relevant fashion (e.g., by staying on topic). Given its importance in the socio-cognitive development of a child, we believe it is imperative to study and obtain empirical evidence of the development of this skill in children.

Current research on this question relies on the manual annotation of a small sample of children's interactions, which limits our ability to draw general conclusions about development. Going back to our proposal of using ML as a tool to automatically scale the annotation of raw data, in this chapter we propose to mitigate the limitations of manual labor by introducing automatic tools for coherence judgments in children's early natural interactions with caregivers. Drawing inspiration from the field of dialogue systems evaluation, we built and compared several automatic classifiers. These tools generally model formal and topical coherence in the conversation (see also Section 1.2.1).

After validating the tool against manually annotated ground truth labels, we used it to automatically annotate, new and large-scale data, almost two orders of magnitude larger than our manually annotated dataset. The major theoretical result was that we were able to replicate existing results and generate new data-driven hypotheses. The broad impact of the work is to provide resources that can help the language development community study communicative development at scale, leading to more robust theories.

In the following chapter (Chapter 5), we use this tool and another ML based tool to automatically annotate all the English language based CHILDES corpora for communicative intents and semantic coherence. Based on these annotations, we identify several patterns on how children coordinate the meaning of their utterances with the caregiver in early child-caregiver interactions.

4.1. Introduction

Children's language development involves not only the acquisition of formal structures such as phonology, syntax, and vocabulary but also the learning of how to *use* this formal knowledge to communicate with people around them in day-to-day interactions. Becoming a competent conversational partner requires children to master several skills such as turn-taking (Agrawal et al., 2023; Casillas et al., 2016; Levinson, 2016), active listening (Bavelas et al., 2000; Bodur et al., 2023; J. Liu et al., 2022), communicative repair (E. V. Clark, 2020; Dingemanse & Enfield, 2024; Nikolaus et al., 2022) and interactive alignment (Chieng et al., 2024; Fusaroli et al., 2023a; Misiek & Fourtassi, 2022; Misiek et al., 2020; Pickering & Garrod, 2004).

In this chapter, we focus on a conversational behavior commonly known in the developmental literature as ***contingency*** (Abbot-Smith et al., 2023; Bloom et al., 1976; Hale & Tager-Flusberg, 2005; Keenan & Klein, 1975; Melander & Sahlström, 2009; Nadig et al., 2010; Pagmar et al., 2022; Piaget, 2005; Slomkowski & Dunn, 1996). It can be defined — broadly speaking — as the collaborative ability to contribute to a dialogue in a relevant fashion, e.g., by connecting with the topic of the ongoing exchange. It is, thus, the glue that makes conversation different from a "succession of disconnected remarks," (H. P. Grice, 1975) and "collective monologues" (Piaget, 2005).

Given that contingency is at the heart of the very definition of a conversation; similar concepts have been introduced and studied — beyond the domain of child development — in many scientific fields that deal with dialogue characterization and/or generation such as pragmatics in linguistic theories (e.g., H. P. Grice, 1975; Sperber & Wilson, 1986), Conversation Analysis in sociology (e.g., adjacency pairs Schegloff & Sacks, 1973), and dialogue evaluation in human-agent interaction (e.g., Mehri & Eskenazi, 2020).

Cognitive and social impact

Being able to provide contingent conversational turns is believed to be associated with the child's developing cognitive competencies such as Theory of Mind (the ability to infer other people's mental states such as goals, beliefs, and desires) and executive functions such as Inhibitory Control (that is, the ability to inhibit one's impulses vis-à-vis a given stimulus so as to provide a more appropriate response)(see Matthews et al., 2018, for a review). Indeed, learning how to stay on topic requires, amongst other things, the ability to *also* consider the interlocutor's perspective and to inhibit the tendency to *always* talk about one's own interests regardless of what the interlocutor is talking about.

In addition, the mastery of contingency in childhood has important social implications such as the ability to maintain friendships (Hazen & Black, 1989). For instance, peer popularity was found to be negatively correlated with children producing more non-contingent, off-topic comments in conversations with their peers (Place & Becker, 1991). More critically, research such as Garzaniti et al. (2011) and Miczo et al. (2001) suggests that many observed differences between children in terms of conversational

skills tend to persist into adulthood, with an impact on their workplace interactions and relationship satisfaction (see Abbot-Smith et al., 2023, for a review).

Towards an automatic annotation of child contingency

Given the connection of conversational contingency with children's broad socio-cognitive development and the persistence of its impact on their later well-being, it is of utmost importance to investigate this phenomenon in its earliest manifestation, i.e., in the context of child-caregiver early *natural* interactions (Pellegrini et al., 2012).

While several corpora of early child-caregiver conversations have been curated (MacWhinney, 2000), a major impediment to the study of contingency is the need for resource-intensive manual annotation. We propose that this impediment can be mitigated through partial or full automation, thanks to recent advances in language and dialogue modeling. Such tools could, in addition, make it possible to study development at a large scale; ideally allowing both an investigation of how current knowledge on the matter – typically based on small-scale studies (e.g., Bloom et al., 1976; Keenan & Klein, 1975; Piaget, 2005) – generalize to a much larger, more diverse sample of children, as well as facilitating the discovery of new insights and hypotheses using bottom-up approaches.

We turn, for inspiration, to the literature on dialogue system evaluation (e.g., evaluating the response relevance of a ChatBot in a free conversation with a human) which has made significant progress, especially since the adoption of pre-trained language models, namely transformer-based models like BERT (Devlin et al., 2019) and GPT2 (Radford et al., 2019). Earlier computational methods tended to be *feature-based*, i.e., extracting several cues and using them as estimators of contingency (as perceived by humans). Such cues included counting repetitions/distribution of certain nouns phrases across turns, the use of speech acts and adjacency pairs, contextual embeddings, and measures of turn similarity (e.g., Barzilay & Lapata, 2008; Cervone et al., 2018; Yi et al., 2019).

More recently, researchers started leveraging pre-trained language models to evaluate the contingency of a turn in the context of the dialogue history. We will call this approach *Language Model-based* (to contrast with the Feature-based approach).

Introducing pre-trained models has allowed researchers to capitalize on rich linguistic knowledge that these models had acquired from data that far exceeds the size of the typical dialogue datasets used to train feature-based methods. This addition resulted in a significant improvement, i.e., a higher similarity with human judgment – compared to previous methods (Mehri & Eskenazi, 2020; Mehri et al., 2022; Pang et al., 2020; Sai et al., 2020; Yeh et al., 2021).

The current study and related work

This work is, to the best of our knowledge, the first attempt to automatize the evaluation of contingency in early child-caregiver natural conversations. This data is

different from typical adult conversations used in most above-reviewed work on contingency evaluation (e.g., the SWITCHBOARD corpus: Godfrey et al., 1992). For instance, there is an asymmetry between young children's – rudimentary – language use abilities and the caregiver's mature conversational skills. In addition, the caregiver tends to adapt their language when they talk with children, compared to when they talk with adults. These differences in terms of conversational asymmetry, style, and context call for a dedicated investigation.

In terms of methods, while current research work – with adult data – has largely moved from a Feature-based to a Language Model-based (LM-based) approach, here we study and compare both. Indeed, it is possible that pre-trained language models fail to capture the above-mentioned specifics of child-caregiver interaction, given that these models were pre-trained on data of a very different nature. Conversely, it is possible that child-caregiver dialogues show simpler patterns that can be more adequately captured using a feature-based method. Finally, it is not impossible that neither the feature-based nor the LM-based approach provides a satisfactory account of child-caregiver dialogue contingency if, say, the overall context – which can be crucial for contingency judgment – is not very transparent in the verbal exchange.

For both the Feature-based and LM-based methods, we need a reasonable amount of hand-annotated data from child-caregiver dialogues. This annotated data is necessary for training, fine-tuning, and evaluation. There is – to our knowledge – no publicly available annotation for children's early contingency behavior. Thus, another contribution of this work is to provide such a resource, using a longitudinal corpus of children aged 20 and 32 months old (The New England Corpus, Snow et al., 1996).

The chapter is organized as follows. First, we describe how we processed and manually annotated the New England corpus. Next, we describe the various features and models we used to automatically annotate the corpus. Finally, we discuss the results of the automatic annotation and demonstrate the use of these models for a large-scale investigation of contingency within all of the English-language CHILDES corpora.

4.2. Manual Annotation

4.2.1. Corpus

We annotated contingency behavior in a subset of the New England corpus (Snow et al., 1996). This corpus consists of a longitudinal recording of $N = 52$ children at 14, 20, and then 32 months of age. The context was semi-structured free play between children and their caregivers. The corpus is transcribed and segmented into conversational turns. It is publicly accessible through the CHILDES repository (MacWhinney, 2000) using CHILDES-db R library (Sanchez et al., 2019). We picked this corpus as it covers the age range where children begin developing linguistic and (joint) attention skills that allow them to engage in increasingly extended back-and-forth conversations with the caregiver, thus offering an ideal window to study development from the earliest

stages. In addition, the corpus was manually annotated for speech act categories (using the child-adapted INCA-A scheme, Ninio et al., 1994), which we needed for our analyses.

4.2.2. Data pre-processing

After pilot annotations, it was apparent that verbal data from 14-month-olds was not intelligible enough to enable a precise study of contingency. Thus, our sample included data from children recorded when they were 20 months old and, then, when they were 32 months old.

Starting from the transcripts, we filtered out utterances that weren't intelligible or speech-related, e.g., babbling and other vocalizations. We also filtered out the utterances from the investigator of the study (keeping only utterances from the child or their caregiver). The resulting dataset included a total of 32,343 utterances out of the original size of 81,473 utterances in the New England corpus.

4.2.3. Procedure

We focused on *turn switches*, i.e., transitions in the conversation when parents or children took a turn following their interlocutor. In other words, if, say, the caregiver made several consecutive utterances, and the child did not intervene (or vice versa), we do not analyze the transition between these consecutive utterances. From a total of 12,981 turn switches across all 85 transcripts that make up the corpus, we annotated – manually – 3,898 turn switches (around 30%), from 28 transcripts that were sampled randomly from the corpus.

The sample can be broken down into 4 equivalent-size conditions as follows: 955 turn-switches of 20-months-olds responding to caregivers, 994 turns of 32-months-olds responding to caregivers, 957 turns of caregivers responding to 20-months-olds, and finally 992 turns of caregivers responding to 32-months-olds.

Two human annotators coded all these turns for contingency on a 3-point scale as `non-contingent`, `contingent`, and `ambiguous`. The annotators made their judgments based on the surrounding verbal context in the dialog. We decided to use the transcripts as our sole source of information for judging contingency as not all the transcripts in CHILDES had accompanying high-quality videos. For turn switches that were not classifiable without information from other modalities, we used the label `ambiguous`. Consider the following example:

> Caregiver: *What's in there?*
> Caregiver: *What do you think they are?*
> Child: *What's this?*
>
> — New England corpus, 32-55.cha

Here, if it can be inferred from the visual modality that the child was pointing/referring to the same thing as the caregiver, then we can consider the child's response to

be `contingent`. However, since we are only considering the verbal data, we mark the child's response as `ambiguous`.

Early attempts were used to converge on a common, systematic scheme (see Appendix B.1). Next, both annotators coded all data in batches of approximately 200 turn-switches. After every batch, they adjudicated their disagreements. All original annotations in each batch (i.e., before adjudication) were used to calculate the inter-annotation agreement. The two annotators achieved a weighted Kappa score of $\kappa = 0.728$ (using quadratic weights).

4.2.4. Results

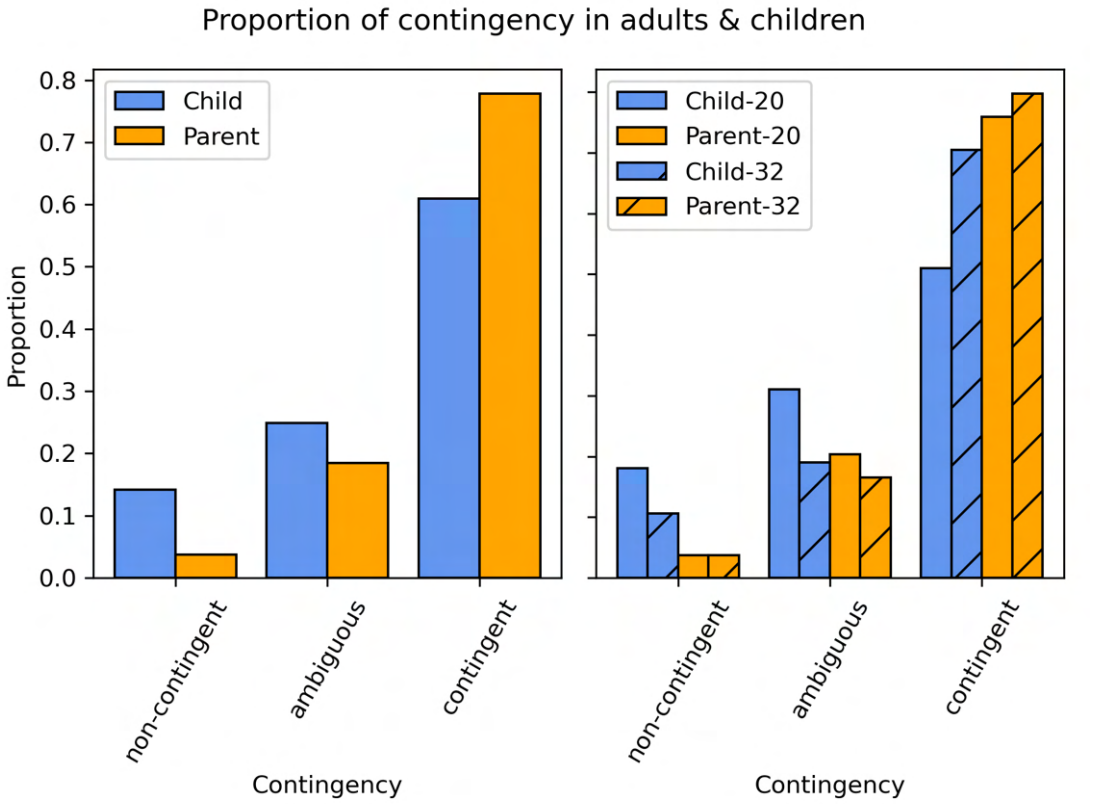

Figure 4.1.: The proportion of `contingent`, `non-contingent`, and `ambiguous` utterances spoken by children and adults in our manually annotated data. The results shown on the right-hand graph are broken down by the age of the child (20 and 32 months).

Figure 4.1 shows the results of the manual annotation of contingency for children and adults, both grouped (children vs. adults) and broken down by the age of the child (20 and 32 months). When we consider the grouped data (left panel), adults had a higher overall average proportion of contingent utterances (78% of their total turns)

compared to children (61% of their total turns). Adults also had lower ambiguous turns (compared to children) and only a very small proportion of non-contingent turns. Children's non-contingent turns represented a minority of their total, but this proportion was still noticeable: 14% of total turns.

When we look at the results broken down by the child's age (right panel), we can observe a developmental pattern. First, in terms of children's own responses, the proportion of contingent turns increases from 51% at 20 months to 70% at 32 months. Non-contingency decreased from 18% at 20 months to 11% at 32 months. Second, in terms of caregivers' responses to children, similar findings were observed: Contingency increased from 76% when talking to children at 20 months to 80% when talking to 32-months-old. Ambiguity decreased from 20% at 20 months to 17% at 32 months old (and non-contingent responses remained at floor level).

4.3. Automatic Annotation

Following recent research on dialogue system evaluation (see Mehri et al., 2022, for an overview), we define the task as labeling the contingency of a turn given a context made of several previous turns in the conversation. We test and compare two different approaches. The first is Feature-based: We extract different verbal features from the dialogue (based on previous research) and evaluate their ability to predict contingency using simple classifiers. The second approach is LM-based: We use pre-trained Language Models and test three levels of fine-tuning on our data (from broad to specific): 1) pre-training only, 2) fine-tuning with self-supervised learning on child-caregiver conversations, and 3) fine-tuning on the supervised task (contingency classification) using manual annotations.

4.3.1. Feature-based approach

We test the following features:

4.3.1.1. Speech acts

The speech act categories allow us to infer if, on a high level, the target turn is contingent. For example, we can determine that the category "Yes-no response" is contingent when following a "Yes-no question" and non-contingent when following, say, a "Greeting" (Cervone & Riccardi, 2020; Higashinaka et al., 2014; Sacks, 1967; Schegloff & Sacks, 1973). We use the Inventory of Communicative Acts - Abridged (INCA-A); the most comprehensive coding scheme to date, designed to capture children's emerging speech acts in the context of early interaction with the caregiver (Ninio et al., 1994). INCA-A has 67 different illocutionary categories, which fall into several groups such as directives, declarations, commitments, markings, statements, questions, evaluations, and other vocalizations. The New England corpus, that we use in this study, was manually annotated for INCA-A by the original authors (Snow et al., 1996).

4.3.1.2. Noun phrase repetitions

Several previous NLP studies on text coherence or dialogue contingency used repeated named entities across sentences or turns as a feature for contingency prediction (Barzilay & Lapata, 2008; Cervone & Riccardi, 2020; Cervone et al., 2018). The idea is that a turn in which the speaker refers to the same entities as the interlocutor did in a previous turn would be more contingent than one in which the speaker refers to different entities. Given that child-caregiver conversation evolves around simple daily objects or animals instead of the typical entities identified by dedicated NLP tools (e.g., famous people's names and big organizations), we decided to use a broader measure indicating the number of times any noun phrase was repeated across the context and the target turn. To identify the noun phrases, we use the English transformer-based syntactic parser from SpaCy.[1]

4.3.1.3. Semantic embeddings and similarity

Following Yi et al. (2019), we make use of sentence-level embeddings and cosine similarity as features for contingency prediction. For the embeddings, we used pre-trained Sentence Transformers (Reimers & Gurevych, 2019) to obtain the embedding of the composite {context, turn}. For cosine similarity, we first obtained separate embeddings for context and turn and then computed the cosine similarity between them. The idea behind using these features is that coherent context-turn pairs would occur closer in the representation space as opposed to non-coherent pairs since they would, for e.g., share similar semantic content.

4.3.2. Language Model-based approach

Since GPT-2, an auto-regressive transformer language model (Radford et al., 2019), was proven effective in previous research on dialogue evaluation (Mehri & Eskenazi, 2020; Pang et al., 2020), we used it as a starting point to experiment with three levels of fine-tuning on our data. Then, for comparison, we tested another – and more recently introduced – transformer-based model (i.e., DeBERTaV3, He et al., 2023) pre-trained with a different self-supervised objective function (i.e., Replaced Token Detection), compared to GPT-2 (i.e., Next-word prediction based on past context).

4.3.2.1. GPT-2

GPT-2 is a language model, built of Transformer decoder blocks (no encoder) and pre-trained on WebText: A corpus made of 8 million documents that were linked to in Reddit and received at least three upvotes (to increase the quality of training data) (Radford et al., 2019). We used the version of the model with 124 million parameters[2].

We used this model in three ways, corresponding to the three levels of fine-tuning on our data, ranging from broad to specific, as follows:

[1] link to model: https://spacy.io/models/en#en_core_web_trf
[2] link to model: https://huggingface.co/gpt2

a) GPT-2 with pre-training only First, we used the default pre-trained version of the model without any further training on our data. To estimate the contingency, we calculated the perplexity of a turn given the context, quantifying the extent to which this turn naturally follows from the preceding context. This estimation is based on the linguistic knowledge the model has gathered in pre-training.

As GPT-2 is an auto-regressive model (i.e., predicting the next word based on the *past context*), perplexity is well-defined as the exponent of the average of the negative log-likelihood. For a sequence of tokens $X = (x_1, x_2, x_3, \ldots, x_t)$, making up the composite {context, turn}, the perplexity of X is calculated as follows:[3]

$$PPL(X) = \exp\left\{-\frac{1}{t}\sum_{i=0}^{t} log\, p_\theta\,(x_i|x_{<i})\right\}$$

b) GPT-2 with self-supervised fine-tuning The second approach involved fine-tuning a pre-trained GPT-2 model on child-caregiver conversations. We used the same (self-supervised) objective function to fine-tune GPT-2 on all English-language corpora in the CHILDES repository, excluding data from the New England corpus (because it contains our test data). The fine-tuning data consisted of 4,674 transcripts from a total of N=862 children aged 26 months[4] and up to 60 months.

The model was fine-tuned for 3 epochs on the training data. After fine-tuning, we estimated the contingency by computing the perplexity of {context, turn} using the same formula as in the default version of GPT-2 above.

c) GPT-2 with supervised fine-tuning The third approach was to use the pre-trained model and fine-tune it by directly teaching it to classify whether a turn is `contingent`, `non-contingent`, or `ambiguous` (given its context) using the manual annotations.

4.3.2.2. DeBERTaV3

To compare with GPT-2, we use a more recently introduced Transformer called DeBERTaV3 (He et al., 2023); an improved variant of the DeBERTa model (He et al., 2020), which was, itself, an improved version of BERT (Devlin et al., 2019) and RoBERTa (Y. Liu et al., 2019) transformer models. The most important novelty of DeBERTaV3 is the use of a pre-training objective called Replaced Token Detection (RTD), which proved to be more data-efficient than Mask Language Modeling (MLM) used in DeBERTaV3's predecessors. This model has 304 million parameters and was pre-trained on the English Wikipedia dump, the Book Corpus (Zhu et al., 2015), OPENWEBTEXT which contains reddit content (Gokaslan & Cohen, 2019) and on the STORIES corpus (Trinh & Le, 2019) which is a subset of CommonCrawl.

[3] We compute the perplexity for the tokens in the turn only (but conditioned on the entire context).

[4] We did not include younger children to ensure we have a significant proportion of intelligible speech from children.

We fine-tuned DeBERTaV3 on the supervised task of contingency prediction of a turn (given its context) using our manual annotation.

4.3.3. Task training and Evaluation

All automatic classifications (both feature- and LM-based) were done by training on 80% of our manual annotation and testing on the remaining 20%. The task consists in learning how to associate the pair {context, turn} with one of three labels (`contingent`, `non-contingent`, or `ambiguous`). For each turn – and based on preliminary exploration – we fixed the context size for all classifiers to be the five preceding utterances. We evaluate the models with 5-fold cross-validation. Crucially, we decided to split folds using transcripts (entire conversation session) as units instead of turn-switches. The reason is to make sure there were no overlapping passages in training and the test folds regarding the context. Thus, our evaluation method is rather strict and tests the ability of the model to generalize to other conversational sessions.

For the feature-based methods, we used logistic regression classifiers,[5] testing the performance of the features both individually and in combination with each other. As for the LM-based classifiers, we had two cases: Concerning models without fine-tuning or with self-supervised fine-tuning, we used the perplexity value of {context, turn} as a feature in logistic regressions (as we did for feature-based methods). Concerning language models with supervised fine-tuning, we did not need to train further classifiers as these models were trained directly for the classification task.

For each model, we report the F-score and the Matthews Correlation Coefficient (MCC) score. While the F-score remains one of the most popular metrics, it can sometimes show misleadingly inflated results, especially with imbalanced classes as in our case. In contrast, the MCC is more reliable and has been shown to be generally unaffected by the unbalanced data issue (Chicco & Jurman, 2020)

4.3.4. Results and Analyses

The results of all classifiers are shown in Table 4.1, together with chance and majority classifiers used as baselines and human inter-annotation agreement as a top-line. The results are broken down for classifiers that were trained/tested either on children's contingency data or on adults' data.[6]

A first inspection of these results confirms that the MCC scores paint a more reliable/interpretable picture than the F-score. For instance, a simple majority classifier for adults' data has a high F1 score of 0.68, but this score only reflects the fact that the overwhelming majority of adults' turns are contingent, and not the accuracy of the classification. In contrast, the MCC score for this same majority classifier for adults is

[5] Other classifiers (e.g., random forest) were used but not reported here as their performance did not improve over the simpler logistic regression.

[6] Training/testing on each age group separately, i.e., 20 and 32 months old led to data-sparsity-related issues, in particular, noisy results with a large variance across folds. These results were not reliable enough to draw clear conclusions. The results are shown in Appendix B.2.

Classifier	Child		Adult	
	F1 score	MCC score	F1 score	MCC score
Majority classifier	0.46 ± 0.06	0.00 ± 0.00	0.68 ± 0.03	0.00 ± 0.00
Chance classifier	0.38 ± 0.02	0.03 ± 0.04	0.41 ± 0.02	0.00 ± 0.05
Speech acts	0.47 ± 0.06	0.05 ± 0.04	0.68 ± 0.03	-0.01 ± 0.02
Noun phrase reps.	0.51 ± 0.05	0.08 ± 0.04	0.17 ± 0.16	0.00 ± 0.03
Cosine similarity	0.36 ± 0.17	0.05 ± 0.11	0.55 ± 0.03	0.16 ± 0.03
Sentence transformer embedding	0.52 ± 0.07	0.10 ± 0.05	0.68 ± 0.03	0.00 ± 0.04
GPT-2 (no fine-tuning)	0.04 ± 0.02	0.00 ± 0.00	0.28 ± 0.28	0.01 ± 0.01
GPT-2 (self-supervised)	0.53 ± 0.02	0.13 ± 0.06	0.51 ± 0.19	-0.03 ± 0.02
GPT-2 (supervised)	0.62 ± 0.06	0.35 ± 0.06	0.69 ± 0.03	0.22 ± 0.08
DeBERTaV3 (supervised)	0.70 ± 0.03	0.46 ± 0.05	0.76 ± 0.03	0.41 ± 0.04
DeBERTaV3 (supervised)+ optimal context	**0.74 ± 0.01**	**0.53 ± 0.04**	**0.77 ± 0.03**	**0.42 ± 0.06**
Human score	0.82 ± 0.02	0.65 ± 0.03	0.86 ± 0.03	0.58 ± 0.07

Table 4.1.: The mean weighted F1 scores and MCC scores along with the standard deviation across a 5 fold cross-validation for children of all ages and for adults. The results for the feature based models are with a logistic regression classifier.

exactly 0, reflecting more faithfully that the classifier has not really learned anything; putting it on par with the performance of the chance classifier. Thus, in the following, we will be analyzing and discussing the results mainly in terms of the MCC scores.

The feature-based classifiers are shown for each feature (tested individually). None of the features managed to surpass a MCC score of 0.16 which is, overall, low. We also trained and tested classifiers with various combinations of features and different classifiers other than the logistic regression (results not shown here, but provided in Appendix B.2), but none of these configurations led to considerable improvement compared to individual scores.

Moving to LM-based classifiers with GPT-2, we can see that the performance increased when GPT-2 was fine-tuned in a self-supervised fashion on CHILDES (compared to GPT-2's original pre-training without any fine-tuning), but this increase was observed only for children's data. The supervised fine-tuning on the manual annotation led to the best results across both children and adults.

When comparing GPT-2 (supervised) to DeBERTaV3 (supervised), we found that DeBERTaV3 improved the results by a fairly large margin. This score was further improved (especially for children) with an optimal context size.[7] Overall, this model learned to classify children's data better than it did for adults' data, echoing a similar difference observed in terms of human agreement scores.

Note, however, that even the best-performing classifier is still lower than the human inter-annotation agreement, suggesting there is still room for improvement.

Effect of training and context size

Using DeBERTaV3 (supervised), we simulated the performance of the classifier when trained on smaller portions of the data. Figure 4.2 shows that the performance peaks when fed with around 80% of the available training data for both children and adults, indicating that the size of our manual annotation dataset, although relatively small, was sufficient for fine-tuning the language model.

Next, we tested how DeBERTaV3 (supervised) performed with different context sizes. The results are shown in Figure 4.3. We can see that a large improvement occurs by adding only 2 preceding turns as context. For children's data, performance slightly increases; peaking at a context size of around 8 preceding turns. For adults' data, however, adding context beyond 2 preceding turns does not seem to improve performance (if anything, the performance slightly decreases).

Interestingly, performance with no context at all was above zero, suggesting that some turns had intrinsic properties that correlated with their contingency status. Qualitative inspection of a few examples in the 0-context case shows that turns that were successfully classified were often short utterances or backchannels (e.g., 'yeah', 'no', 'mhm', and 'okay').

[7] The optimal context was 8 preceding turns for children and 2 preceding turns for adults, see Figure 4.3

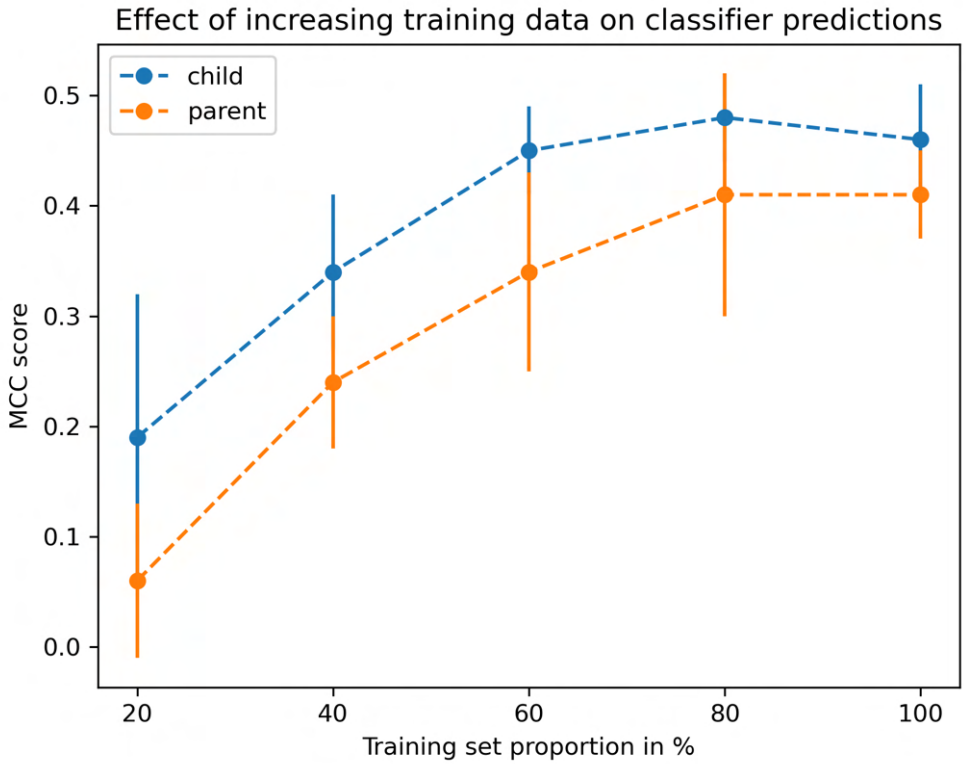

Figure 4.2.: The effect of varying fine-tuning data size (i.e., from the manual annotation data) on the performance of DeBERTaV3 (supervised). The points indicate the mean MCC score across a 5-fold cross-validation and the ranges represent the standard deviation.

4.3.5. Toward large-scale investigation

We select the best models from our training and then use them to predict the contingency for turn switches from all English-language CHILDES corpora (excluding the New England corpus) to see how the automatic annotation of the model behaves on new, large-scale data. We test the model's behavior both within and beyond the age range of the training set.

Within-range automatic annotation Since we used manually annotated data from conversations of children aged 20 and 32 months old for our training, we restricted this first exploration to all English-language turn switches in CHILDES corpora belonging to children aged 20 to 32 months (and their caregivers). Since we did 5-fold cross-validation during DebertaV3's fine-tuning (see Section 4.3.3), we ended up with 5 different classifiers, one for each fold. We ran all 5 models on this new CHILDES data and did a majority vote to get a final prediction for each {context, turn}. In this

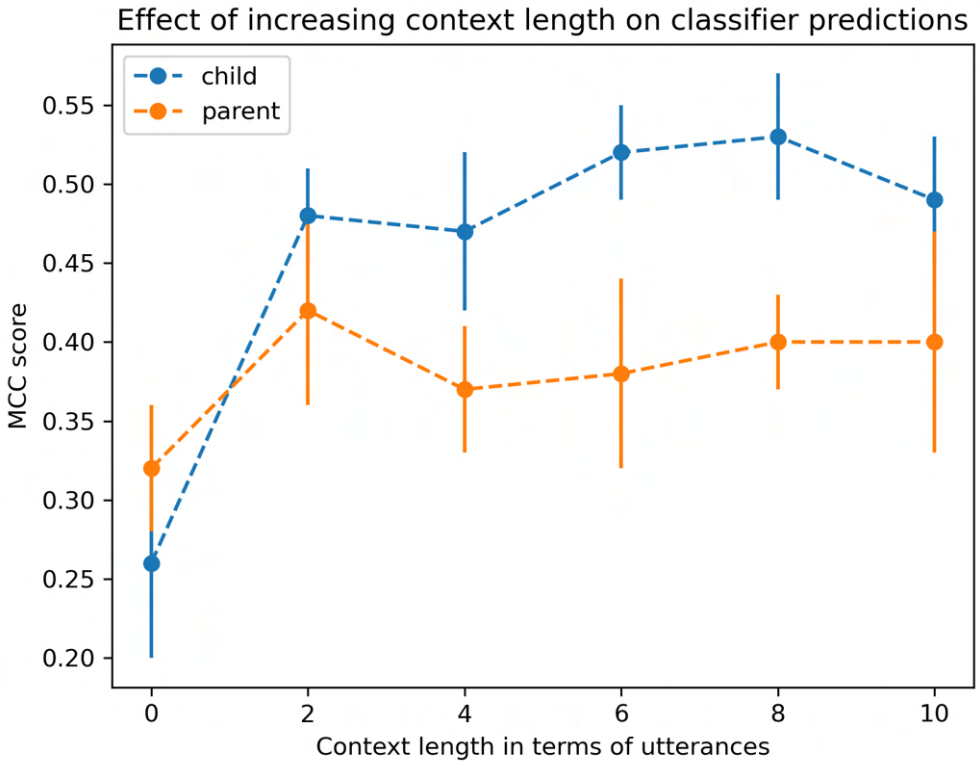

Figure 4.3.: The effect of varying context size on the performance of DeBERTaV3 (supervised). The points indicate the mean MCC score across a 5-fold cross-validation and the ranges represent the standard deviation.

manner, we automatically annotated 345,893 turn-switches for children and 345,133 turn-switches for caregivers in total, that is, two orders of magnitude larger than the manually annotated training data.

Figure 4.4 shows the results. First, the automatic annotation captures the broad developmental difference between 20-month-olds (lower contingency) and 32-month-olds (higher contingency). Thus the automatic classifier replicates the same result obtained with manual annotation (shown in Figure 4.1) using completely different corpora (that have not been seen in fine-tuning), also generalizing it at a large scale. In addition, automatic annotation reveals a new finding: There is a rather *continuous* developmental pattern in children's contingency between 20 and 32 months, although – crucially – no data from children in these intermediate ages were seen in fine-tuning. We can also see a similar (though slower) developmental pattern in parents' contingency, this slight increase appears to be due mostly to a reduction in the number of ambiguous turns, with non-contingent turns remaining largely at floor level (which is also similar to what we obtained with manual annotation).

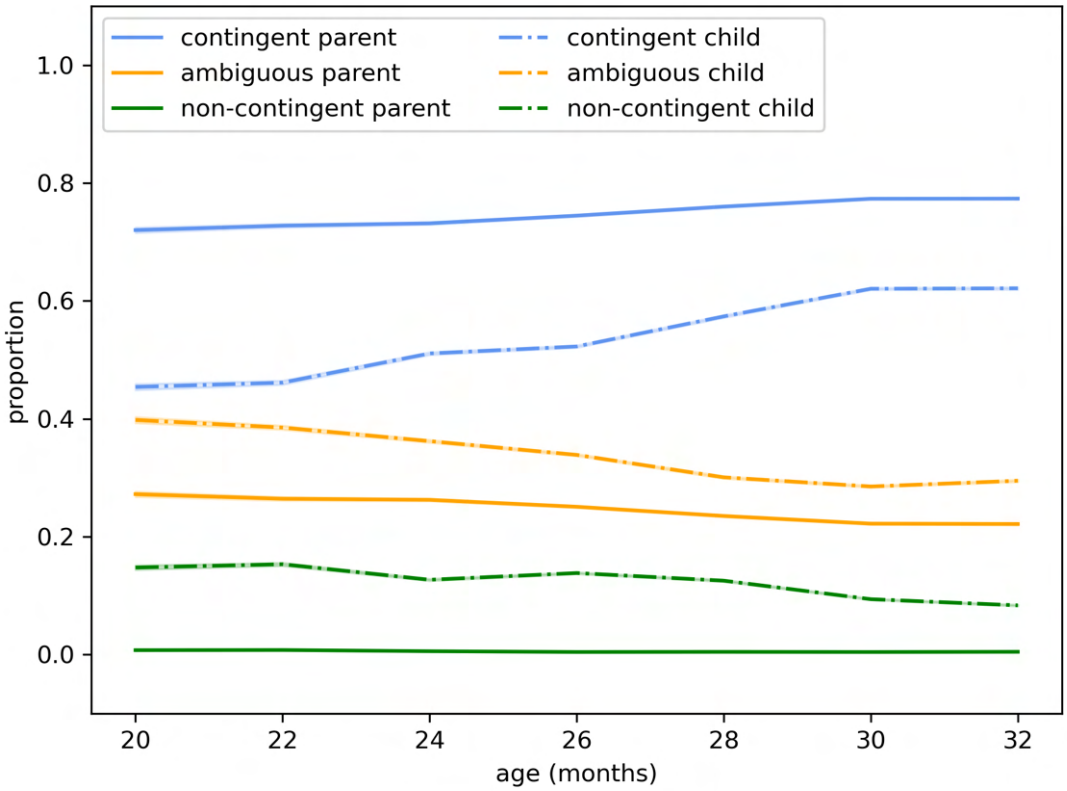

Figure 4.4.: The proportion of contingent, non-contingent, and ambiguous utterances obtained using automatic annotation of new, large-scale data within the age range of the fine-tuning set.

Beyond-range automatic annotation To investigate the extent to which our automatic classifier can be used with data beyond the age range of the fine-tuning set, we now automatically annotate conversations of children aged up to 64 months in all English-language CHILDES, following a similar procedure as above (leading to 911,143 turn-switches for children and 893,973 turn-switches for caregivers in total).

Figure 4.5 reproduces results of Figure 4.4 in the 20-32 months interval (same data) and shows the automatic annotation beyond this range, up to 64 months. The results show no increase in children's contingent turns beyond 32 to 36 months and no decrease in non-contingent turns either, a finding that is counter-intuitive and most certainly inaccurate. We conclude that the model cannot be used reliably to annotate data beyond the age range seen by the model during fine-tuning.

4.4. Conclusion

Conversational contingency plays a crucial role in children's communicative and socio-cognitive development. Understanding how this skill develops requires that

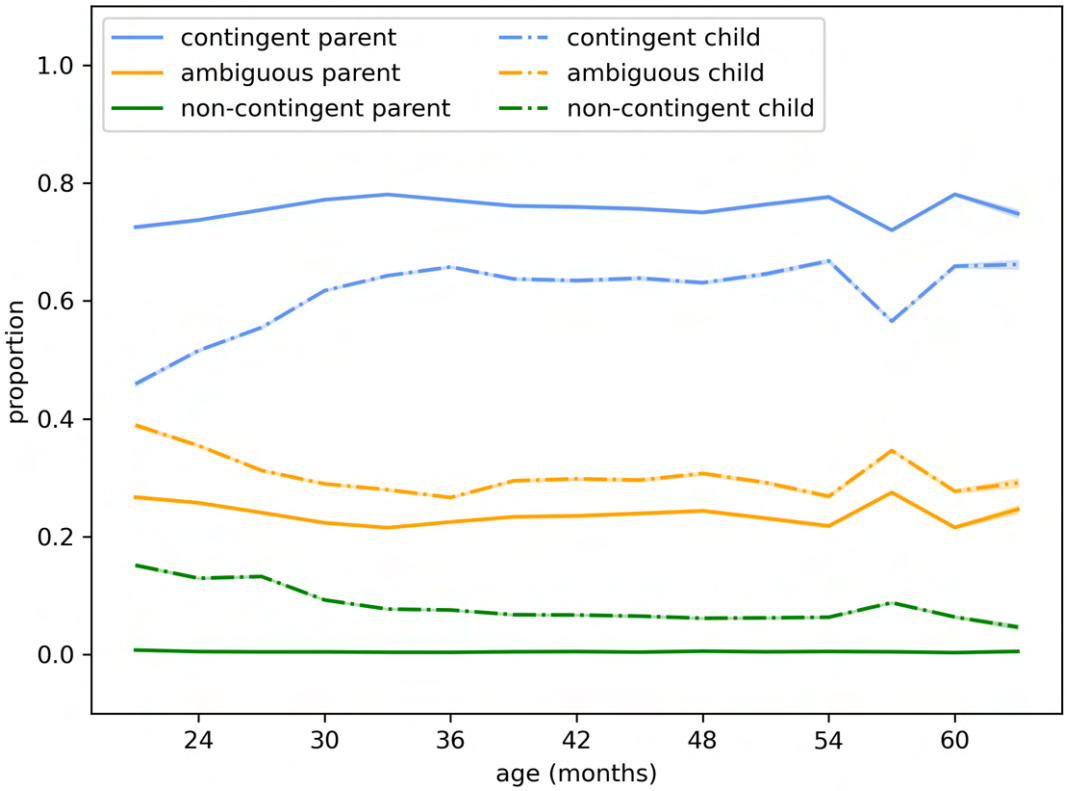

Figure 4.5.: The proportion of contingent, non-contingent, and ambiguous utterances obtained using automatic annotation of new, large-scale data beyond the age range of the fine-tuning set, and up to 64 months.

we study its earliest manifestation in child-caregiver natural interaction, soon after the child becomes able to utter intelligible speech and engage in a verbal back-and-forth with the caregiver. While several studies have investigated contingency behavior around that period and beyond (Abbot-Smith et al., 2023; Bloom et al., 1976; Keenan & Klein, 1975; Piaget, 2005), most are typically based on small-scale samples (with known limitations); due primarily to the fact that the study of this phenomenon in natural interaction requires resource-intensive manual annotations.

Here we explored the possibility of automatizing the process of child-caregiver contingency judgment, with the goal of facilitating more research into this question, e.g., by testing the generality of our current knowledge at a large scale and by allowing a bottom-up exploration of new hypotheses. We took inspiration from the field of dialogue systems evaluation to build and test various automatic classifiers. The most accurate one was based on a pre-trained language model, which we fine-tuned on a relatively small sample of data that we annotated manually. This classifier was able not only to replicate and generalize findings – obtained with human annotators – on data it had never seen but also to generate a new hypothesis about the shape of the

developmental trajectory.

Finally, this work can impact not only research on children's early conversational development but also research on the role of interaction in predicting language learning in the wild. While current methods examine the role of predictors such as children's overall linguistic input (e.g., the quantity of speech heard) or broad interactive measures like the number of turns or temporal contingency (Bergelson et al., 2023; Donnelly & Kidd, 2021; Elmlinger, Goldstein, & Casillas, 2023), the current work allows more detailed examination of the verbal content of the interaction and its semantic connectedness, facilitating empirical testing —at scale— of key proposals from interactionist theories and models of language acquisition (Bruner, 1985; E. V. Clark, 2018; Masek et al., 2021; Nelson, 2007; Nikolaus & Fourtassi, 2021, 2023; Tomasello, 2003).

4.5. Limitations

The performance of our best model was still inferior to that of human annotators (although the gap is not huge). How can we improve? The common approach is to annotate more data manually and increase the size of the fine-tuning data. However, as Figure 4.2 demonstrates, this is unlikely to improve performance as we appear to have already hit a peak. Another – and perhaps more promising way forward – is to use larg*er* language models with a lot more parameters, pre-trained on much more data than say, GPT2 or DeBERTaV3. Up until very recently, such Large Language Models (LLMs) have been closed to researchers with no possibility of fine-tuning their parameters (e.g., GPT4 OpenAI, 2023). This is changing both with the release of more open LLMs (e.g., Llama 2 Touvron et al., 2023, Mistral 7B A. Q. Jiang et al., 2023) and with improvements in machine learning techniques that allow fine-tuning of LLMs with reasonable computation resources (e.g., Dettmers et al., 2022; Houlsby et al., 2019).

Another limitation of our study is that – for practical reasons – we considered only the transcript of the conversation for our manual annotations and for training/evaluating our models. Nevertheless, the visual context can be very informative for contingency judgment in early childhood, especially in evaluating referring expressions. Adding the visual context would help resolve several instances of what we labeled as "ambiguous". This, however, will depend on the availability of curated multimodal corpora (which are rare, given the concern to protect the anonymity of children) as well as on the ability of models to learn reliably from real-life, naturalistic, and noisy multimodal scenes (which is still an open research question).

5. Exploring the Structure of Early Child-Caregiver Dialogue

This chapter is based on the article "Mapping the Communicative Landscape of Early Child-Caregiver Dialogue" (Agrawal et al., 2025a), which is currently under revision for the journal *Cognitive Science*.

From the existing literature (see Section 1.2), we know that there are a few studies that have considered the role of communicative intents in children's conversation skill development. However, they don't consider the influence of topical coherence while analyzing communicative intents for identifying the structure in early child-caregiver conversations. Moreover, most of the existing studies are conducted on small samples of data. Thus, there is a lack of quantitative studies that offer a comprehensive insight into the communicative landscape that characterizes child-caregiver dialogues.

In this chapter, we addressed this gap by utilizing the tool developed for annotating semantic coherence in Chapter 4 and a tool developed for annotating communicative intents (Nikolaus et al., 2022) to automatically annotate all the English-language corpora present in CHILDES repository (MacWhinney, 2014). Our approach combined models for communicative intent inference and models for semantic coherence evaluation, to provide a comprehensive analysis of linguistic coordination and its emergence in infancy.

Our analysis found both *focused* turn pairs i.e., turns with clear communicative intents (e.g., question-response) and *open* turn pairs (e.g., statement-statement) were common in structuring the interactions between the child and the caregiver. However, where these pairs differed significantly was in terms of their semantic coherence (e.g., open pairs were generally less coherent) and in terms of the role of the interlocutors (e.g., caregivers initiated most of the focused pairs). We identified developmental shifts in the expression of intents and we also found an interesting dissociation between the frequency and coherence of communicative intents.

This study demonstrates how we can leverage ML models as tools to establish a solid, bottom-up empirical foundation for developing ecological theories of communicative development which in the study at hand looks at coordinating interactions in terms of being coherent.

5.1. Introduction

Children's early dialogue is a window into their social, linguistic, and cognitive development (Abbot-Smith et al., 2023; E. V. Clark, 2018; Kurkul & Corriveau, 2018; Piaget, 1989; Stivers et al., 2018; Yu et al., 2019). It is a complex exercise that requires the ability to coordinate both the timing (whose turn is it to speak?) and the meaning (are we communicating about the same thing?). While a thriving line of research has focused on investigating the emergence of temporal coordination (Elmlinger, Schwade, et al., 2023; Feldman, 2007; Nguyen et al., 2022), here we focus on the emergence of *meaning coordination*, specifically as it manifests in early verbal interaction with caregivers.

Successful meaning coordination in dialogues manifests through the ability to engage with the interlocutor's communicative intents (e.g., answering a question) while maintaining the conceptual coherence of the exchange (e.g., the answer has to address the question, taking into account the broader conversational context) (H. H. Clark, 1996; H. P. Grice, 1975; Pickering & Garrod, 2021; Sacks et al., 1974; Sperber & Wilson, 2002). In contrast, unsuccessful meaning coordination results in a disconnected sequence of turns. Piaget famously coined the term "collective monologue" to describe his observation that children's early dialogues (especially with peers) lack adequate meaning coordination despite appearing to involve turn-taking (Piaget, 1926). To him, this was evidence of the preschooler's overall egocentric thinking. Nevertheless, the idea of children's egocentric speech has been intensely debated, and the nature of collective monologues—at least in its strong form as lacking genuine communication or meaning coordination—has been challenged (Bloom et al., 1976; Dorval et al., 1984; Garvey & Hogan, 1973; Keenan & Klein, 1975; Vygotsky, 2012).

Contemporary research has shown that children have rather precocious communicative skills, allowing them to interpret and engage with the interlocutor's communicative intents even before they start using verbal language. In fact, many have suggested a crucial role of these early forms of meaning coordination in the very development of verbal communication (Bates et al., 1975; Bruner, 1985; Kuhl, 2007; Ninio & Snow, 1996; Tomasello, 2009).

Despite accumulated evidence about children's early communicative skills (see Bohn & Frank, 2019, for a review), much of it has come from studies that emphasize comprehension over (spontaneous) interaction or from small-scale qualitative studies. We still lack a comprehensive and quantitative understanding of how meaning coordination emerges in early child-caregiver natural dialogues. This understanding is crucial for uncovering the challenges of translating children's precocious communicative competencies (primarily studied in the lab) into real-world verbal interactions.

The time is ripe for conducting such naturalistic studies, as many of the barriers that previously hindered progress in this area have now been significantly alleviated:

- First, large and diverse datasets of early child-caregiver interactions have been collected over the past few decades (e.g., MacWhinney, 2014). These datasets help address the challenges of variability and context- dependency that are

inherent to social interactions. The availability of such extensive data enables researchers to draw more generalizable conclusions.

- Second, recent advances in Natural Language Processing (NLP) provide automatic tools for analyzing meaning coordination in dialogues, adapting to the unique context of child-caregiver interactions. This technology enables us to scale analyses across extensive datasets in a cost-effective manner, significantly reducing the need for labor-intensive manual annotation—a longstanding bottleneck that has limited large-scale studies of early natural dialogues.

The aim of this study is to leverage these advancements to systematically map the communicative landscape in early child-caregiver dialogues, focusing on the 1 to 3 year age range, i.e., the earliest in childhood when children start engaging in relatively extended verbal interactions with caregivers, providing insights into the very emergence of meaning coordination.

How to investigate early meaning coordination in natural dialogues?

One approach proposed by Piaget and others (see Dorval et al., 1984, for detailed review) is to look in the dialogue for turns with clear communicative intents, such as asking a question or making a request. Such turns are directed toward the interlocutor to elicit a specific and immediate response: The question seeks an answer, and the request expects acceptance or rejection. We adopt a similar terminology as the one used by Dorval et al. (1984), calling these sequences "focused pairs."[1] [2] The analysis of focused pairs with clear communicative intents has been commonplace in studies of communicative development (e.g., Bates et al., 1975; C. Bergey et al., 2022; C. A. Bergey et al., 2024; Bruner, 1985; Cameron-Faulkner, 2014; Casillas & Hilbrink, 2020; Dore, 1974; Garvey, 1975; Matthews, 2014; Ninio et al., 1994; Snow et al., 1996; Zhao et al., 2024).

That said, focused pairs can miss important patterns and, therefore, are insufficient to fully account for the development of meaning coordination. For instance, many back-and-forth sequences in children's dialogues involve "open" pairs, as they do not fall into a conventional type of focused exchange (such as when the child offers an unsolicited comment on a statement made by the interlocutor). These sequences are, nonetheless, still expected to be coordinated. It is therefore crucial, especially in an investigation that seeks comprehensiveness, that we consider not only broad categories of *communicative intents* and their relations but also *the verbal content* of

[1] To be more specific, Dorval et al. (1984) used the terms "focused turns," but we used "focused pairs" instead to emphasize its interactive nature.

[2] This concept bears similarity to the idea of adjacency pairs in Conversation Analysis, (Schegloff & Sacks, 1973) though, in this literature, the focus is less on analyzing communicative intent and more on describing turn organization.

the interaction in order to determine whether the child's utterance is semantically well-coordinated with that of the interlocutor, e.g., that the child is *coherent*, maintaining the topic of conversation.

Coherence has been investigated by developmental researchers under many labels such as semantic contingency, coherency, and topic maintenance (among others) (e.g., Blain-Brière et al., 2014; Bloom et al., 1976; Capps et al., 1998; Dorval et al., 1984; Hale & Tager-Flusberg, 2005; Keenan & Klein, 1975; Matthews et al., 2018; Rosnay et al., 2014; Slomkowski & Dunn, 1996), with many researchers focusing on its potential impact on children's social relationships and on ways it can be improved via interventions (Abbot-Smith et al., 2023; Adams et al., 2012; Black & Hazen, 1990; Parsons et al., 2019; Place & Becker, 1991). However, this literature has not sufficiently interacted with the one focusing on communicative intent categories.

Outline of our research approach

As we mentioned above, the main research strategy of this study is to capitalize on advances in automatic language processing tools to map the child-caregiver communicative landscape from large-scale natural data. We use two previously developed models for characterizing child-caregiver categories of communicative intents (Nikolaus et al., 2022) and for annotating the semantic coherence of the verbal content (see Chapter 4).

Nikolaus et al. (2022) tested a variety of NLP techniques (from simple machine learning classifiers to pre-trained transformers) to label child or caregiver turns for their category of communicative intents, using a comprehensive, child-appropriate coding scheme made of 67 fine-grained communicative intent categories (INCA-A, Ninio et al., 1994) and their manual annotation on the New England corpus of child-caregiver interactions (Snow et al., 1996). The best-performing model was a Conditional Random Field, which learns to infer the most likely label sequence (i.e., the communicative intents) given a conversation.

On the other hand, in our previous study (see Chapter 4), we tested various NLP techniques in dialog systems, which seek to evaluate the semantic coherence of a given turn given the immediately preceding turn from the other interlocutor and a short history of the conversation up to that turn. When testing on manually annotated child-caregiver dialog data—also using the New England corpus (Snow et al., 1996)—we found the best performing model was a pre-trained transformer (DeBERTaV3, He et al., 2023), fine-tuned on the child-caregiver (labeled) data. The model adapts the linguistic knowledge it had acquired during pre-training on large data in order to quantify the likelihood that a given child's turn follows from a prior verbal context. Note that this model is more comprehensive than methods that capture surface relationships (like word overlap and semantic similarity) as is typically done in alignment and mimicry studies (Fernandez & Grimm, 2014; Fusaroli et al., 2023b; Misiek & Fourtassi, 2022). Here, the model has to detect if a turn naturally follows from the previous turn and the context, even if the turns do not share surface similarities, like determining that "yes" or "no" are coherent when following a polar question or that "okay" is coherent

following a request.

Combining these two models, the current study characterizes, at a large scale, the extent to which responses (whether from children or caregivers) are coordinated with the interlocutors' previous turn (i) in terms of being part of a focused vs. open exchange and (ii) in terms of being coherent.

Using this approach, we document major patterns in child-caregiver meaning coordination in the entire English-language CHILDES (MacWhinney, 2014). Since the models we used for our large-scale studies were both trained on labeled data from the New England corpus, we focused on the same age range (that is, 20 to 32 months old).[3] We systematically test how the findings in the small-scale (manually annotated) data in the New England corpus (collected from $N = 45$ children across 78 conversations) generalize to large-scale (automatically annotated) data –in the same age range – in the entire English-language CHILDES, including $N = 609$ children across 44 corpora, 2577 total conversations, and $690k$ interactive pairs.

5.2. Data and Methods

New England Corpus

The corpus that we used for our models' training is the New England corpus (Snow et al., 1996), publicly available via the CHILDES repository (MacWhinney, 2014). The original corpus contains longitudinal recordings of $N = 52$ children aged 14, 20, and 32 months interacting with their caregivers. However, we did not use data corresponding to the 14-month-old children as much of their utterances were unintelligible and, therefore, could not be properly processed by our text-based models. We were left with data from $N = 45$ children across 78 conversations.

English-language CHILDES

While New England was used for models' training, English-language CHILDES was used for large-scale investigation. We focused on the same age range as in the New England corpus (that is, 20 to 36 months) as this is the data distribution that the models were trained and evaluated on. We aggregated data across 44 corpora, involving $N = 609$ children across 2577 conversations. A full list of the corpora included is available in Appendix C.

Turn pairs

As we are interested in interactions, we focused on turn pairs in transitions, i.e., moments in the conversation where the turn switches from child to parent or vice-versa. This means we ignore consecutive utterances within the same speaker (which is mostly the case for parents), focusing only on the last utterance before the turn

[3] We excluded data from 14-month-old children; see Methods.

switches.[4] Our large-scale data includes around 690k of such turn pairs (half of which are child-to-parent and the other half are parent-to-child). These only include pairs where both the child's and caregiver's utterances are verbally intelligible (See the Limitations Section for more context on the reasons and implications for focusing on verbal data alone).

Communicative intents

The utterances in the New England corpus (Snow et al., 1996) came annotated with their respective communicative intents based on the INCA-A coding scheme (Ninio et al., 1994). The INCA-A is the most comprehensive coding scheme of its kind, containing 67 categories, including both easy and challenging intent types, thereby allowing the study of development over infancy and preschool.

The model we used for annotating communicative intents is the one introduced by Nikolaus et al. (2022). The authors used manual annotation of Snow et al. (1996) to train a variety of models, of which a Conditional Random Field (CRF) proved to be the most effective. We used an identical training procedure. For every utterance, the CRF model has access to several features such as the speaker's identity (child or caregiver), the unigrams and bigrams of the utterance along with the part of speech tags and word repetitions with the previous utterance. We used the trained model to automatically annotate all the conversational turn pairs in the English-language CHILDES corpora.

Focused pairs

For our analysis, we grouped the communicative intents into pairs of focused vs. open pairs (Dorval et al., 1984). Table 5.1 provides a list of all the focused pairs (similar to adjacency pairs in Conversation Analysis). Any pairs not present in this table were considered to be open pairs.

Semantic Coherence

The model we used for the automatic labeling of semantic coherence is the one introduced in the previous chapter (see Chapter 4). We first manually annotated a subset of the New England corpus: 4k turn pairs in transitions (out of a total of around 13k pairs in the entire corpus). For each of these pairs, the response was categorized based on its coherence with the interlocutor and the dialog context more generally. We adopted a simple 3-point scale: coherent, incoherent, or uninterpretable. We then

[4] An alternative would have been to concatenate consecutive utterances within a turn. However, this would have artificially inflated the caregiver's turn length and complexity and obscured the overall communicative intent, making it difficult to study the dynamics (e.g., focused exchange). Imagine the caregiver makes several statements (ST), ending their turn with a question (YQ). Here, the child only responds to the last utterance with an answer (AA). Considering the interlocutor's last utterance (the question) and the speaker's immediate response (the answer) allows us to study these dynamics more transparently.

Table 5.1.: A list of all the focused exchange pairs of communicative intents.

Initiating intent	Response intent
YQ (Yes/No Question)	AA (Affirmative answer to Yes/No Question)
YQ (Yes/No Question)	AN (Negative answer to Yes/No Question)
QN (Wh-question)	SA (Wh-answer)
RQ (Yes/No Suggestion)	AD (Agree to suggested act)
RQ (Yes/No Suggestion)	RD (Refuse suggested act)
RP (Request/suggest an action)	AD (Agree to suggested act)
RP (Request/suggest an action)	RD (Refuse suggested act)
MK (Social norm)	MK (Social norm)
CT (Correct wrong form)	RT (Imitate/repeat)

tested a variety of modeling techniques, the best-performing approach of which was a fine-tuned Language Model (DeBERTaV3, He et al., 2023). More specifically, the model was fine-tuned to predict the coherence of a turn given the conversational context made of five previous turns. As with communicative intents, we used the DeBERTaV3 model to automatically label all turn transitions in the English-language CHILDES corpora (see above).

5.3. Results

First, we present the outcome of the models' evaluation; then, we present results documenting the structure of child-caregiver interactions in terms of belonging to focused vs. open pairs and, next, the extent to which each type of pair is semantically coherent. Finally, we examine how both structure and coherence develop in our age range, leveraging the data-dense examination our large-scale study allows.

5.3.1. Models' Evaluation

Here, we evaluate the quality of the automatic models. We use the New England corpus (Snow et al., 1996), as it contains the ground truth annotations (i.e., manual annotations) both for INCA-A communicative intents (contributed by Snow et al., 1996) and for semantic coherence (contributed in the previous Chapter). We used these manual annotations to train the best-performing models reported for automatic communicative intent labeling (Nikolaus et al., 2022) and for semantic coherence (see Chapter 4).

We evaluated both models using 5-fold cross-validation, where each fold involved training on 80% of the data and evaluating on the remaining 20%. The results we obtained confirm the near-human-level performance of these models: the communicative intent classifier reached 72.33% accuracy, approaching the 81% accuracy of human inter-annotation agreement reported by Snow et al. (1996). The semantic

coherence model reached an F1 score of 74%, approaching the 82% F1 score of human inter-annotation agreement that we report in the previous chapter.

5.3.2. The structure of child-caregiver interaction

We start with focused pairs (e.g., question → answer or request → acceptance/refusal),[5] and the goal is to quantify the prevalence of this type of interaction in child-caregiver early dialogues. The results are shown in Figure 5.1. They indicate that around 30% of children's responses are part of a focused exchange initiated by parents. For parents, this number is below 10% of their total turns following children's.

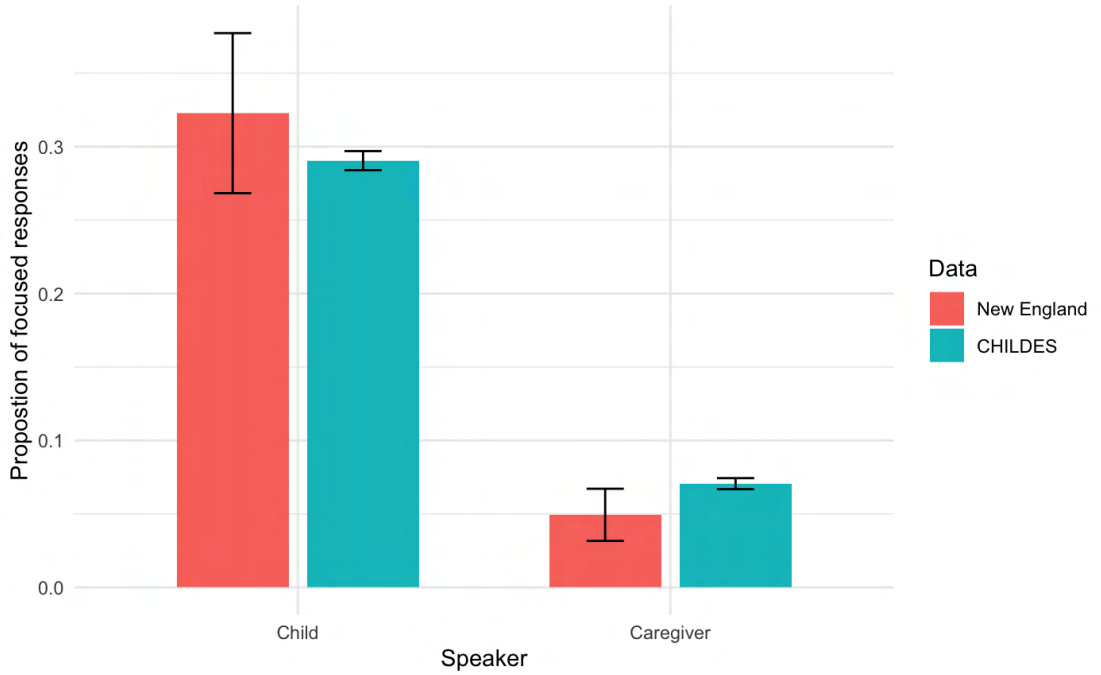

Figure 5.1.: The proportion of focused responses in adjacent pairs of turns (out of the total made up of both focused and open responses). Error bars show 95 % confidence intervals. Data comparisons aims to show how small-scale findings in the New England corpus generalize at a large-scale to CHILDES.

[5] See in Methods the list of what we considered standard pairs of focused exchange, given the communicative intent categories available in the INCA-A coding scheme.

Figure 5.2.: Adjacent pairs for **children responding to caregivers**. Each plot should be read from left to right: the initiating intents on the left (the caregiver) and the responding intents on the right (the child). Communicative intents occurring less than 1% of the time were filtered out for a clear representation.

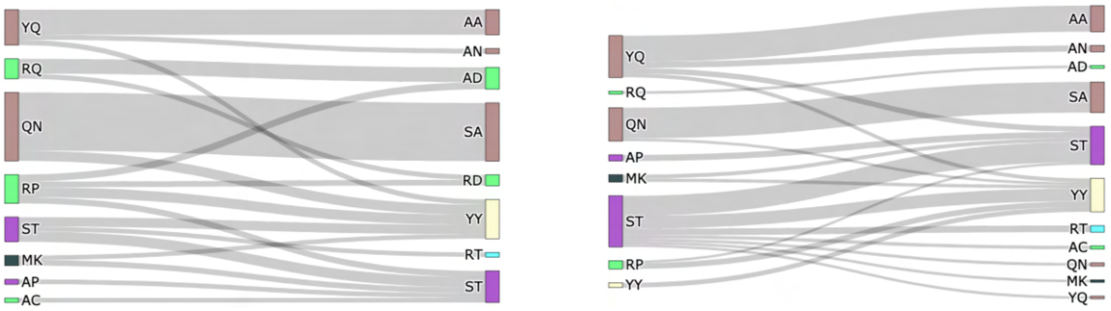

Figure 5.3.: Coherence of **children responding to caregivers** in a given adjacent pair. The rows represent the caregiver's communicative intent (initiations), and the columns represent the child's (responses). For readability, responses occurring less than 0.1% of the times in frequency are marked as -1.0 in the figure.

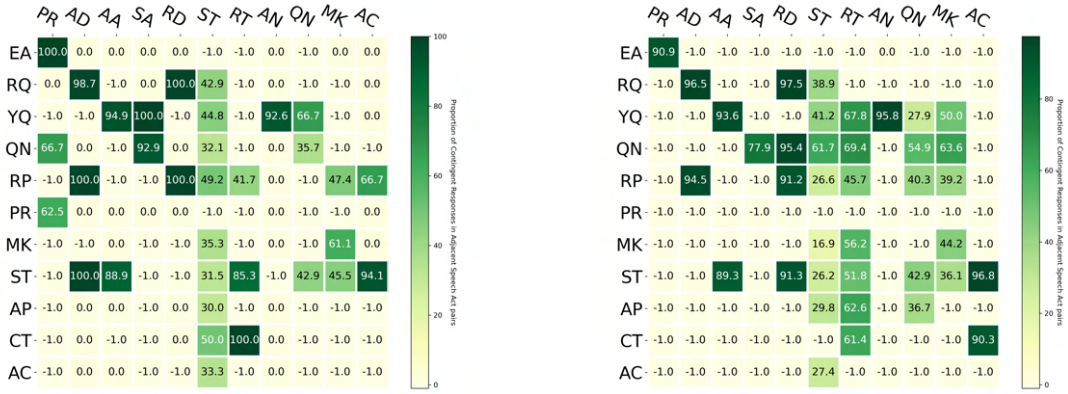

Table 5.2.: A list and short description of all the communicative intent labels displayed in the river plots and heat maps above.

Intent label	Short Description	Intent label	Short Description	Intent label	Short Description
EA	Elicit sound	RQ	Yes/No (Suggestion)	YQ	Yes/No (Question)
QN	Wh-question	RP	Request/Suggest	PR	Peform game move
MK	Social norm	ST	Statement	AP	Agree (proposition)
CT	Correct wrong form	AC	Show attentiveness	AD	Agree (act)
AA	Yes (Y/N question)	SA	Wh-answer	RD	Refuse (act)
RT	Imitate/Repeat	AN	No (Y/N question)	YY	Non-sensical utterance

Figures 5.2 and 5.4 show river plots indicating which category by the interlocutor in the immediately previous turn led to which category in the response in both the case of children's responses and caregivers' responses, respectively. The thickness of the bands provides a visual representation of the probability with which each response is given. For visual clarity, the figures include only the most frequent categories in the initiation and response. For just the river plots, we also consider the YY category which are word-like utterances without any clear function. This is to highlight the proportion of times that children respond to caregivers with non-sensical utterances.

Furthermore, Figures 5.3 and 5.5, show the coherence of the most frequently occurring pairs of communicative intents (based on the corresponding river plots): The numbers indicate the percentage of times a pair was coherent. Another thing to note here is that the data used for plotting the heatmaps for the New England corpus is a subset of the data that was used to plot the corresponding river plots for the New England corpus. This is because the amount of manually annotated data available for coherence is a smaller subset (N=19 children across 28 conversations) of the amount of manually annotated data available for communicative intents (N=45 children across 78 conversations).

Figure 5.4.: Adjacent pairs for **caregivers responding to children**. Each plot should be read from left to right: the initiating intents on the left (the child) and the responding intents on the right (the caregiver). Communicative intents occurring less than 1% of the time were filtered out for a clear representation.

Figure 5.5.: Coherence of **caregivers responding to children** in a given adjacent pair. The rows represent the child's communicative intent (initiations), and the columns represent the caregiver's (responses). Coherent responses occurring less than 0.1% of the times in frequency are marked as -1.0 in the figure.

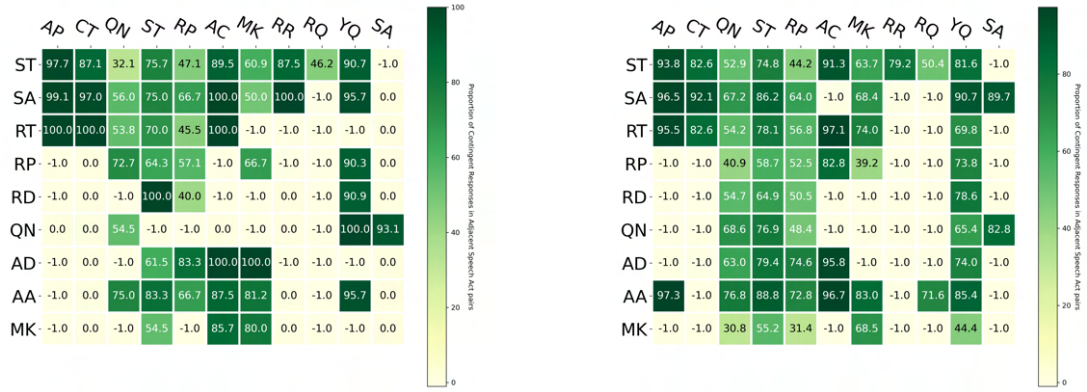

Table 5.3.: A list and short description of all the communicative intent labels displayed in the river plots and heat maps above.

Intent label	Short Description	Intent label	Short Description	Intent label	Short Description
RR	Request repetition	RQ	Yes/No (Suggestion)	YQ	Yes/No (Question)
QN	Wh-question	RP	Request/Suggest	YY	Non-sensical utterance
MK	Social norm	ST	Statement	AP	Agree (proposition)
CT	Correct wrong form	AC	Show attentiveness	AD	Agree (act)
AA	Yes (Y/N question)	SA	Wh-answer	RD	Refuse (act)
RT	Imitate/Repeat	PF	Prohibit act		

To understand the interactive dynamics that contribute to this structure (especially the asymmetry between children and adults), we examined finer-grained patterns at the level of pairs of communicative intents. Figures 5.2 and 5.4 show river plots indicating which category by the interlocutor in the immediately previous turn led to which category in the response.

For children's responses (Figure 5.2), we identify many common patterns in both focused and open pairs. For focused pairs, a highly common pattern is that of caregivers asking questions and children providing an answer: Question (YQ, QN) → Answer (AA, AN, SA). Another common pattern is when caregivers make requests and children respond by approving or refusing: Request (RQ, RP) → Approval/Refusal (AD/RD).

Regarding open pairs, the most typical is the one made of statements. Unlike questions, requests, or greetings, statements do not necessarily elicit a specific response. Here is an expert from the New England corpus of statement (ST) to statement (ST) exchange:

-two beds.
-I don't see see a other bed. (i)

Our definition of open pairs is not limited to sequences of statements, however. The river plot helps us uncover many other ones, such as when children respond to caregivers' Yes/No-questions (YQ) by a statement (ST) instead of the expected "yes" or "no":

- got it?
- too heavy. (ii)

For caregivers' responses (Figure 5.4), the major pairs do not look like what one would consider as focused exchanges, confirming the numbers shown in Figure 5.1. Indeed, only a minority of caregivers' responses are triggered by an elicitation, like questions or requests. Most sequences involve caregivers *following up* on the child's previous contributions, especially using follow-up questions:

-this one.
-who's that one? (iii)

Other common responses, besides follow-up questions, are about the caregivers agreeing with or acknowledging children's contributions:

-I have Cookie Monster.
-you sure do. (iv)

5.3.3. The coherence of child-caregiver interaction

The second aspect of our analysis concerns coherence. More specifically, we examine how much of the child-caregiver exchange is coordinated at the semantic level, comparing when this exchange falls within a focused vs. open pair. The results of this analysis are shown in Figure 5.6. We found that, in both children and caregivers, the overwhelming majority of focused pairs were coherent. The situation was different in the case of open pairs: For children, only about a third of their responses in this category were coherent, whereas, for the caregivers, the majority of responses remained coherent (though with a slightly lower proportion compared to focused pairs).

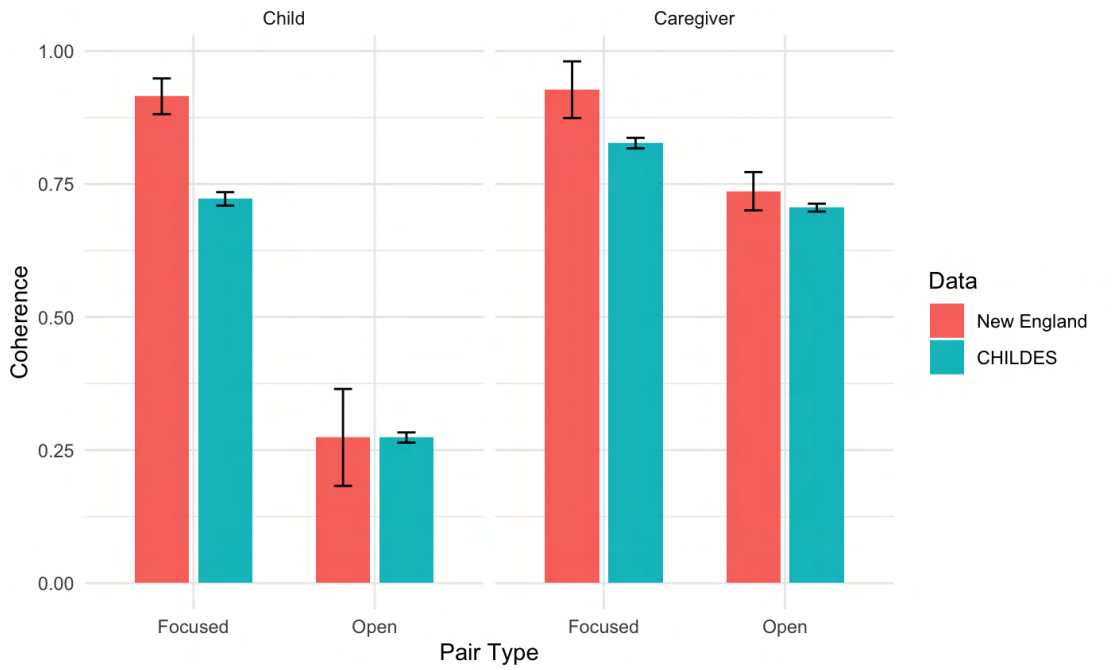

Figure 5.6.: The average semantic coherence of focused vs. open responses in adjacent pairs. Error bars show 95 % confidence intervals. Data comparisons aims to show how small-scale findings in the New England corpus generalize at a large-scale to CHILDES.

To further understand the interactive dynamics behind these patterns, and, similar to the previous section, we examine fine-grained patterns at the level of pairs of communicative intents. While in the previous sub-subsection, we focused on river plots, here we focus on the heat maps as shown in Figures 5.3 and 5.5, showing the coherence of the most frequently occurring pairs of communicative intents (based on the corresponding river plots): The numbers indicate the percentage of times a pair was coherent.

For children's responses, we see, indeed, different patterns in the heat maps depending on whether these responses are part of a focus vs. open exchange. For focused pairs like Questions (YQ, QN) → Answers (AA, AN, SA) and Requests (RQ, RP)

→ acceptance/refusal (AD, RD), or Social norms (MK) → Social norms (MK), the heat maps generally indicate very high coherence, confirming the numbers observed in Figure 5.6.

Concerning the case of open pairs, we found that the lower coherence scores in Figure 5.6 were not due to, say, some types of open exchange being coherent and other types being non-coherent. Rather, the heat maps show that almost all attested sequences can be both coherent and incoherent and, therefore, should be examined on a case-by-case basis. Starting by the most typical form of open exchange, i.e., Statement (ST) → Statement (ST), an excerpt of a coherent exchange was given in (1). Here is an excerpt of an incoherent, ill-coordinated exchange:

-I thought you just fell down.
-I want this Cookie Monster. (v)

The same can be said about other types of open pairs, such as when a question is followed by another question (instead of an answer, as in a typical focused exchange). We found that this sequence can be both coherent (here, as clarification request):

-what it is?
-what? (vi)

Or incoherent:

-what color is that?
-where the other crayon? (vii)

A final example is when a Request is followed by a Statement instead of the expected Acceptance or Refusal. This sequence can still be coherent:

-well let's see what's in the next box.
-toys in it. (viii)

As for caregivers, and despite the fact that almost all their responses do not fall within a typical focused exchange, the heat maps show that all the observed sequences are highly coherent.

5.3.4. Developmental patterns

In this section, we analyze developmental changes in meaning coordination. While Figures 5.1 and 5.6 compared average patterns between children and parents, here Figures 5.7 and 5.8 show how these same patterns develop within our age range in both children and caregivers.[6] Figure 5.7 shows there to be a moderate developmental increase in the rate of both children's and caregiver's focused interactions between 20 and 32 months of age. This rate develops from around a quarter of children's total replies to a third. For parents, it increases from around 5% to 10% of their total replies.

[6]For caregivers, "development" reflects how they change their behavior with children at different ages.

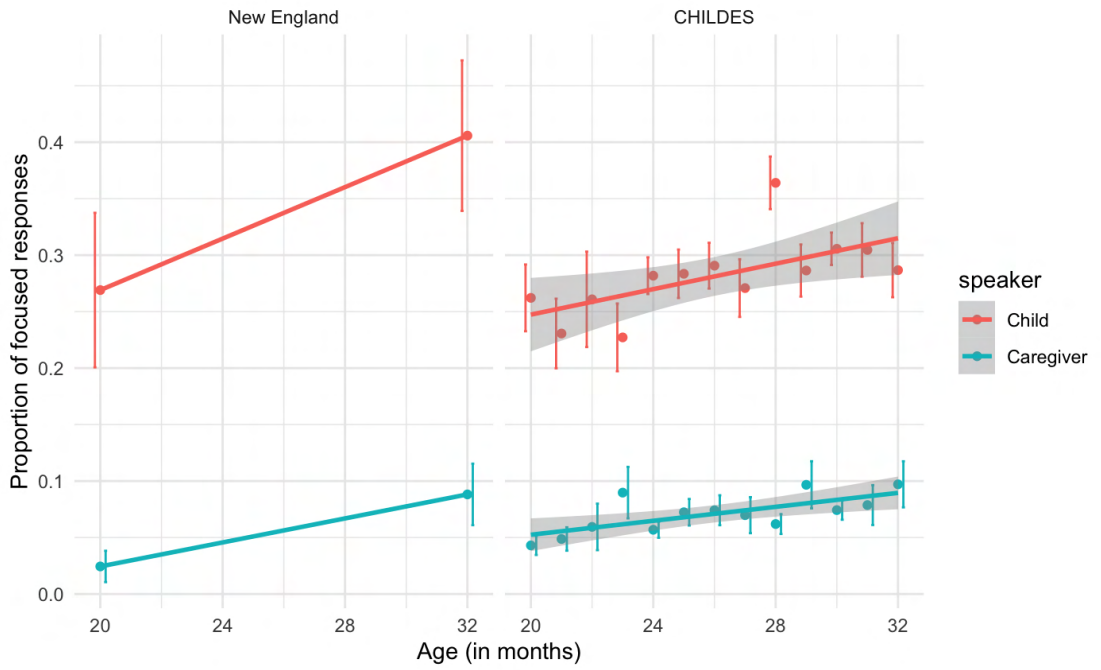

Figure 5.7.: The development of focused responses in adjacent pairs (out of the total made up of both focused and open responses). Error bars show 95% confidence intervals. Data comparisons aims to show how sparse developmental information in the New England corpus generalizes in developmentally dense data in CHILDES.

Figure 5.8 shows that, for children, there is a steady improvement in coherence. We observed this improvement in both focused and open pairs. For focused pairs, coherence increased from a little over half to more than 75% of total responses. In open pairs, it increases from around 15% to 37% on average. For caregivers, coherence is almost at ceiling, though we still observe a small increase in the coherence of open exchanges.

We confirmed this developmental observation using mixed-effects models where `coherence` is predicted by `age` and `pair_type` (fixed effects) and `conversation` (random effect). For children, we found a highly significant, positive effect of both `age` ($\beta = 0.02$, $p < 0.001$) and `pair_type` ($\beta = 0.4$, $p < 0.001$). A similar pattern was observed for parents, although—as apparent in the figure — both effects were smaller compared to children `age` ($\beta = 0.007$, $p < 0.001$) and `pair_type` ($\beta = 0.1$, $p < 0.001$).[7]

To further examine how changes occur at a finer-grained level, Figure 5.9 shows the development of children's response, focusing on how the top frequent categories of communicative intent change in terms of frequency and coherence. First, we observe that all categories gain in terms of coherence between 20 and 32 months of age, except

[7]More generally, all differences and trends reported with CHILDES in the results are statistically (highly) significant using mixed-effects testing.

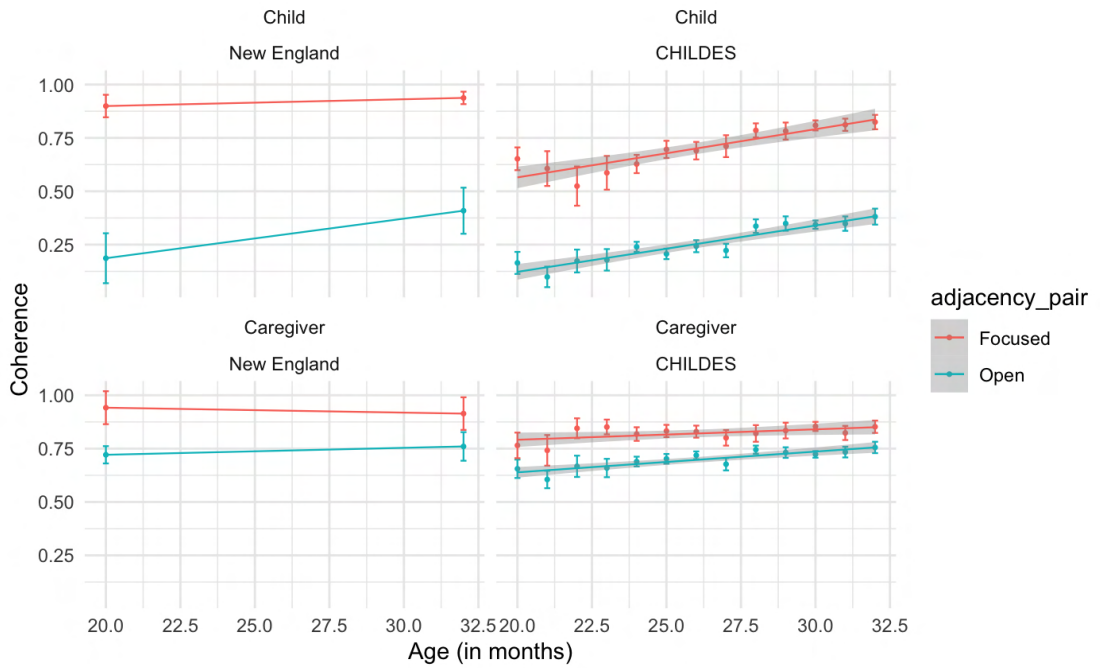

Figure 5.8.: The development of semantic coherence in focused vs. open responses within adjacent pairs. Error bars show 95% confidence intervals. Data comparisons aims to show how sparse developmental information in the New England corpus generalizes in the developmentally dense data of CHILDES.

responses to requests (Agree/Refuse), which are already at ceiling. This was not the case for frequency; some categories decreased, most notably imitations/repetitions, and others increased, such as answers to yes-no questions.

Further, and regardless of development, this Figure shows an interesting dissociation between frequency and coherence at the level of many individual categories of communicative intents. For example, while statements are the most frequent, they are the least coherent. Conversely, while Agree/Refuse are almost always coherent, they are the least frequent (relative to the other selected examples).

5.4. Discussion

This study aimed to map the communicative landscape in child-caregiver early dialogues, combining the analysis of high-level communicative pairs with an examination of the verbal coherence within these pairs. Achieving this goal required developing domain-specific NLP tools that enable generalizable findings across large datasets.

This general research approach, which relies on NLP tools to investigate large-scale developmental data in language, has proven productive in many recent studies, advancing the ecological study of children's language learning and use (Bergelson et al.,

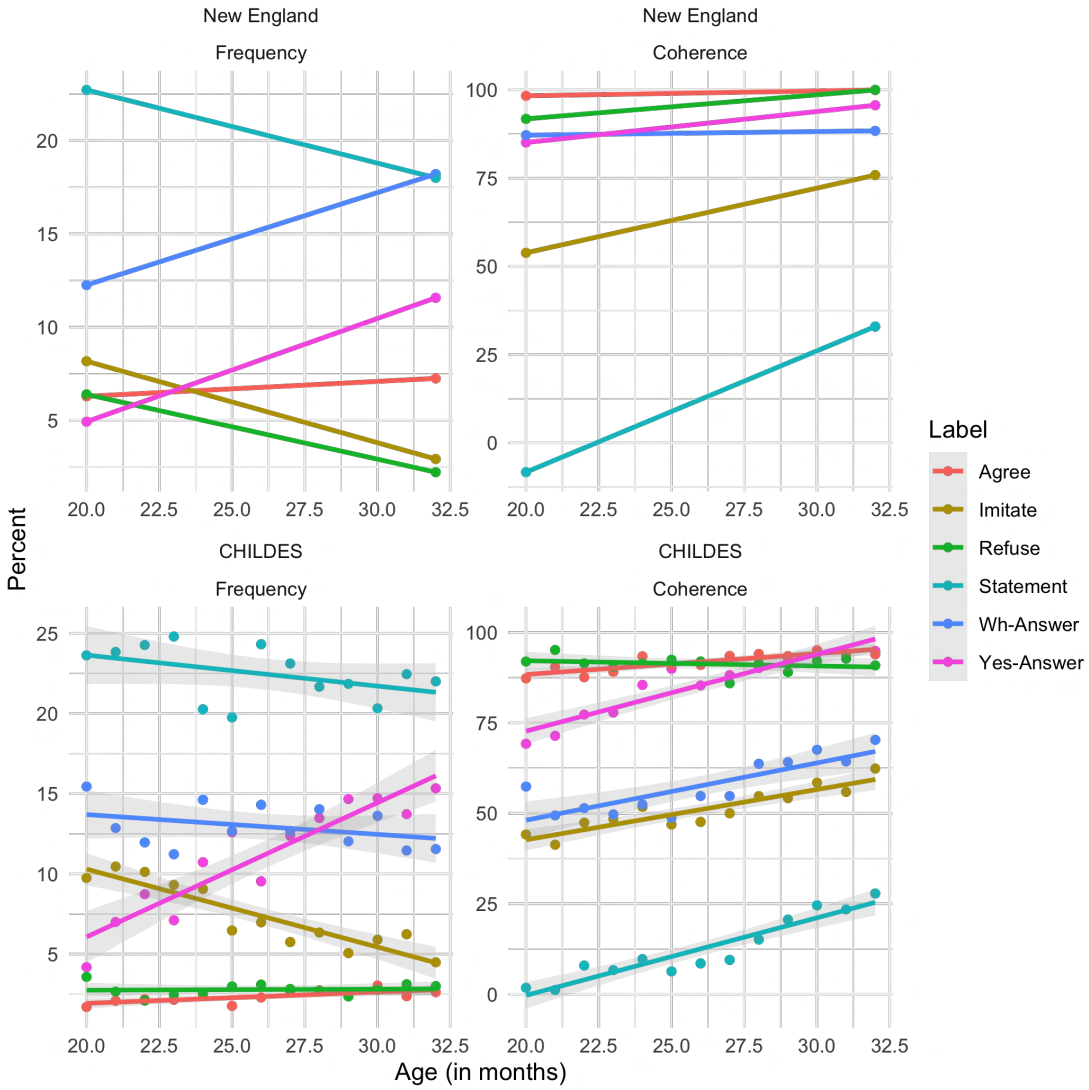

Figure 5.9.: The development of children's top frequent communicative intent categories between 20 and 32 months. We show – side by side – the development of their (relative) frequency and overall coherence given caregivers' initiations. The lines represent best linear fits, and the envelopes indicate 95% confidence intervals. Data comparisons aims to show how sparse developmental information in the New England corpus generalizes in the developmentally dense data of CHILDES.

2023; Cabiddu et al., 2025; Fusaroli et al., 2023b; Huebner et al., 2021; Lavechin et al., 2024; Misiek et al., 2020; Vong et al., 2024; Warlaumont et al., 2014; Warstadt et al., 2023). The main novelty of the current work is that we used tools that allow a bottom-up investigation of children's *meaning coordination* in early dialogues, filling a critical gap in this emergent literature, which has so far concentrated on the analysis of caregivers' input (outside the interactive context), on the analysis of pre-verbal temporal

interactive dynamics, or the examination of partial aspects of verbal coordination (e.g., mimicry).

This research enabled us to document several robust, generalizable patterns in children's early meaning coordination. For example, at 20 months old, about a quarter of children's responses fell within classic focused exchanges (or adjacency pairs) initiated by parents, such as answering questions and accepting or refusing requests. This proportion increased to about a third of total responses a year later. These findings generally agree with Dorval et al. (1984), who also reported a similar proportion and increase with age, though within an older age range (middle childhood) and different interlocutors (peers). We subsequently studied the coherence of these responses given the conversational context; we found they were generally coherent (unsurprisingly, given their highly structured nature), including at the earliest time in our age range. Coherence reached an adult level by 32 months.

That said, we observed a remarkable asymmetry between children and adults regarding their roles in focused pairs. The overwhelming majority of such sequences had parents as initiators (e.g., asking a question) and children as responders (giving an answer). Only a minority of these sequences had parents at the receiving end. The subsequent fine-grained examination showed that parent → child dynamics largely follow the standard pattern of focused exchange, i.e., **Solicitation → Response**. In contrast, the child → parent dynamics suggests a different overall organization: **Response → Follow-up**. This qualitative difference indicates that parents are the primary orchestrators of early dialogues by initiating and following up. While this fact is not entirely surprising, given that we know parents tend to scaffold children's early communication (e.g., E. V. Clark, 2018, 2020; Masek et al., 2021; Nikolaus & Fourtassi, 2023), such a strong imbalance appears to contrast with research suggesting that children are naturally inclined toward questioning and that this tendency plays a role in their social and cognitive development (Butler et al., 2020; M. Callanan et al., 2020; M. A. Callanan & Oakes, 1992; Chouinard et al., 2007; Kurkul & Corriveau, 2018) (But see Limitations below).

Focused exchanges have been widely used as a window into children's early communication (Bates et al., 1975; Bruner, 1985; Cameron-Faulkner, 2014; Dore, 1974; Dorval et al., 1984; Piaget, 1926; Snow et al., 1996; Stivers et al., 2018). However, they capture only part of the child-caregiver dialog dynamics. A sizable part of children's turns belongs to open exchange pairs (e.g., statement → statement). These pairs are much less structured, and therefore, analyzing children's coordination here depends on measuring coherence within the conversational context (e.g., Bloom et al., 1976; Keenan & Klein, 1975). Overall, we found children's responses in open pairs to be low on coherence: Less than one in four responses is coherent at 20 months. Though this number increased to a little less than 50% a year later, it remained much lower than in adults/caregivers, suggesting further development beyond the age range considered in this study. Indeed, research suggests a protracted development in topic coherence, especially as conversational topics increase in complexity and abstractness (Abbot-Smith et al., 2023; Baines & Howe, 2010; Dorval et al., 1984).

In addition to examining children's coordination in the context of focused vs. open

exchange, we also investigated the general properties of their communicative intent use, contrasting frequency of use and coherence. We replicated and generalized previous small-scale studies (e.g., Snow et al., 1996) showing that the frequency of communicative intent use undergoes major changes into the third year of life. A novel finding of the current study was the *dissociation* we observed between frequency and coherence: Some of the most frequent intent types (e.g., statements) were also the least coherent when considering the conversational context (and vice versa). These findings show that frequency of use, when considered alone, is not a good indicator of mastery and that other indicators, like coherence, should be included for a better account of communicative intent development in naturalistic studies.

Limitations

This study uses data from CHILDES, the largest repository of transcribed child-caregiver dialogues. This makes it ideal for large-scale studies focusing on verbal data, like ours. However, some findings may be influenced by the methods used to collect CHILDES data, which typically involve a short recording time (e.g., an hour) during which a researcher observes the child and caregiver participating in a joint activity. For instance, one of our finding was the asymmetry in the roles of children and caregivers during focused exchanges. We speculate this could be –in part– due to the time constraints and the presence of a researcher, prompting parents to sustain the conversation by asking more questions and follow-ups rather than waiting for the child to initiate interactions spontaneously. The extent to which these potential biases cause early child-caregiver dialogue dynamics to deviate from fully spontaneous interactions remains unclear (and a critical question to address in future research). Nevertheless, the fact that we reproduced this asymmetry at scale and across many corpora highlights its consistency and strength in shaping the communicative landscape, at least in the communicative context typical of short recordings.

A natural alternative to short recordings are long-form recordings (Cychosz et al., 2020), which aim to capture speech throughout the child's entire waking day. This approach increases the likelihood of capturing more spontaneous moments where children take a more active role in exchanges compared to typical CHILDES recordings. However, it presents challenges: Its share size makes detailed verbal transcriptions difficult. Even recent speech modeling tools struggle with this noisy data (Lavechin et al., 2024). This situation currently makes studies like ours, which focus on verbal interaction, unfeasible on long-form recordings.

The study adopted cutting-edge computational tools to classify communicative intent and evaluate semantic coherence. Yet, these tools are only effective on textual data; considering the visual context from unstructured and diverse sources/corpora remains an open technological question. Here, the omission of the shared visual context in the interaction may have underestimated children's coordination abilities; for instance, a seemingly incoherent response might still be relevant when viewed within the visual context (inaccessible to the models). This limitation was mitigated, at least partly, by the models having access to a conversational history, which serves

as a proxy for the general context, as this history tends to include references to aspects of the shared visual context.

A final limitation of this work lies in our exclusive focus on verbal interactions. We analyzed only turn pairs where both interlocutors produced intelligible verbal data, excluding interactions where one interlocutor relied on non-verbal signals instead of spoken language. While the authors of the New England corpus annotated non-verbal signals and their communicative intents—for instance, a child's head nod as a Yes-Answer (AA) to a Yes/No question (YQ) or a head shake as a No-Answer (AN)—scaling these annotations was not feasible. This limitation stems from the fact that non-verbal signals are not consistently textually described across corpora in CHILDES.

Excluding turn transitions involving non-verbal responses may have underestimated children's abilities in focused exchanges, particularly at younger ages. To gauge the magnitude of this effect, we counted the number of turns involving head nods or head shakes, as annotated in the New England corpus. Among 20-month-olds, less than 0.8% of their total turns consisted of head nods or head shakes, and among 32-month-olds, this number was less than 0.6%. These findings suggest that while our focus on verbal data may overlook some interactive opportunities, the impact of these omissions is minimal and unlikely to alter the main conclusions of the study.

5.5. Conclusion

To summarize, the current study aims to map the communicative landscape of early child-caregiver dialogues. This large-scale, bottom-up investigation represents a crucial step toward developing a comprehensive theory of communicative development. Such a theory should not only account for competencies identified in controlled lab settings but also explain how these competencies are effectively applied in everyday linguistic communication. We addressed a significant methodological challenge that has previously limited large-scale studies—experts' annotation—by leveraging advances in NLP for automation. While these tools are still limited in accurately capturing multimodal communication beyond text, the rapid progress in multimodal AI holds the potential to provide a more complete understanding of these interactions.

Part III.

Conversational grounding

Table of contents

6. Towards Understanding Conversational Grounding by Quantifying Caregiver's Repair

This chapter is based on the article "Identifying Repair Opportunities in Child-Caregiver Interactions" (Agrawal et al., 2025b), published in the *Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue*.

The final part of this thesis introduces an exploratory study on conversational grounding in child-caregiver interactions. From the literature (see Section 1.3) we know that repairs are one of the principle mechanisms for developing the common ground between interlocutors of a conversation. In fact, repairs are considered as important in both the coordination as convergence (Pickering & Garrod, 2004) and coordination as joint-action (H. H. Clark, 1996) theories for clearing up any misunderstandings that arise during the conversation. We also know how parents' repair helps scaffold children's early communication and also provides a learning signal to them (E. V. Clark, 2020). Often, developmental studies have only focused on linguistic markers of parents' repair and repair initiation (e.g., clarification requests); that too when the topic of the conversation at hand is concrete or physically present in the same space as the interlocutors.

In this study, we focused on analyzing the cases where parents initiated a repair compared to the overall repair opportunities that presented themselves as a means of quantifying the extent to which caregivers seize repair opportunities. One of the novelty of the study is that the weakly-structured setting of the child-caregiver interactions provide the ground truth intents which needs to be added to the shared knowledge and we can also study the repairs in cases where the intent is a complex abstract concept. Past studies have also shown that repairs are more frequently occurring in task-oriented conversations as a means to developing the common ground (Dideriksen et al., 2023) which is favorable for us since the corpus we choose to study involves children playing a word-guessing game with their caregiver.

Our analysis found that caregivers initiated repairs in only a small subset of the cases where a repair opportunity presented itself. We further wanted to test whether we could once again leverage ML as a tool for automatically annotating repair opportunities in children's utterances. However, due to the small amount of positive samples ($N = 154$) in our manually annotated data reflecting the actual repair opportunities, we didn't think fine-tuning a model would give us good results. Considering the mixed success of large language models (LLMs) on evaluating Theory of Mind (ToM) and

social reasoning (e.g., Gandhi et al., 2023; Shapira et al., 2024; Strachan et al., 2024) — which is useful in trying to understand the interlocutors beliefs and intents — and their capacity to be used in a zero-shot or few-shot setting with impressive results, we tasked several LLMs to test their capacity to recognize repair opportunities in children's utterances and found their performance to be lacking compared to human annotators.

This study provides an initial exploration that is valuable both for developmental studies and for researchers aiming to improve dialog agents for child–machine interaction.

6.1. Introduction

For an effective, intelligible, and fluent conversation, a key competency that the interlocutors must possess is the ability to successfully coordinate and negotiate their shared beliefs, knowledge, and assumptions (H. H. Clark, 1996; H. H. Clark & Schaefer, 1989; Stalnaker, 1978). This ability — also known as *conversational grounding* — allows the interlocutor to interpret an utterance accurately based on their shared knowledge with the speaker of the utterance and the dialog history, thereby letting the interlocutor respond in a coherent and effective manner. It helps the interlocutors resolve any ambiguity and clear up misunderstandings that occur during a conversation (Fried et al., 2023).

Interlocutors in a conversation start out with some shared belief space or *common ground* from shared culture, a social group, or previous interaction (Baker et al., 1999; H. H. Clark, 1996). The common ground is then further developed throughout the conversation by contributions from all the participants in the conversation (H. H. Clark & Brennan, 1991). For grounding any information, the interlocutors need to provide implicit or explicit evidence that information has been well communicated and understood. This evidence can take the form of acknowledgments (e.g., backchannels), initiation of the relevant next turn, by showing continued attention (e.g., through eye gaze), by issuing a clarification request, among other signals (H. H. Clark & Brennan, 1991; H. H. Clark & Krych, 2004; H. H. Clark & Schaefer, 1989).

Identifying and repairing breakdowns in early communication

Typically, an interlocutor has a communicative intent in mind that they need to get across to the other interlocutor. To illustrate, suppose Jane has an intent I in mind and tries to communicate it to Jack. If I has something to do with a situated object — as, for instance, is often the case when talking to a young child — then one of the actions Jane can take is to simply point to the object to indicate her intent, or by looking at the target, inviting gaze following (e.g., Frank et al., 2009). However, if I is not situated — e.g., an abstract idea or a displaced target — something that becomes more and more prevalent as children develop, then Jane and Jack need to ground I in their mutual understanding by more sophisticated means. To this end, Jane continually monitors Jack for signs of understanding of her intent. If Jack shows signs of misunderstanding, then she can step in and repair the misunderstanding. The objective of this exercise is to ensure that both Jack and Jane share their understanding of Jane's intent I.

Indeed, one of the primary mechanisms for maintaining common ground is identifying and repairing breakdowns in communication (Benotti & Blackburn, 2021; H. H. Clark & Krych, 2004; Dingemanse et al., 2015; Fusaroli et al., 2017; Purver et al., 2018; Schegloff, 1992). However, we know little about how this mechanism plays out in child development, especially in child-caregiver interactions. This is a significant gap given that many proposals suggest a role for caregivers' communicative feedback on children's production, especially when these productions are ill-coordinated or poorly constructed and potentially helping in furthering language use refinement (E. V. Clark,

2018, 2020; Nikolaus & Fourtassi, 2023). While there is a wealth of studies focusing on caregivers' role in guiding infants' understanding when in a situated context and the target is visually accessible (e.g., review in Çetinçelik et al., 2021), there is hardly any study quantifying this phenomenon when the target is abstract or not visually available to interlocutors.

A notable difficulty here, especially when analyzing spontaneous conversations (e.g., CHILDES, MacWhinney, 2000), is that the intent to be grounded (i.e., I) is not always apparent to a third party, namely the researcher, making the analysis fully dependent on the caregiver's reaction to what the child said (e.g., whether the caregiver asked for clarification). While a focus on the caregiver reactions allows for an estimate of actual repair initiation, this estimate can be misleading because it does not account for all *repair opportunities*, some of which may have been missed or ignored by caregivers.

To address the difficulty of identifying the caregiver's intent in fully unstructured settings, here we resort to using a weakly structured word-guessing game that allows us to maintain a (relatively) naturalistic conversational style while also providing access to the ground truth intent I (i.e., the word to be guessed). While this context — where the caregiver is making a child guess a word — is not fully naturalistic, it is meant to approximate the instance when the caregiver and child work collaboratively to ground a complex intent or idea (e.g., why limiting screen time is important) in shared understanding, only here this intent is operationalized, for simplicity, as a simple word that needs to be guessed.

The goals of the current study

Using this setup, a first goal of the current study is to quantify caregivers' actual repair relative to repair opportunities, as follows. First, we characterize all children's questions (e.g., "Does this object fly?") in terms of being well or ill-coordinated, thanks to our access to the caregiver's intent and the exchange history. In particular, the subset of children's questions that are ill-coordinated (e.g., asking "Does this object fly?" when it was already established that the object cannot take flight) provides the set of what we call **repair opportunities**. Second, we characterized instances of caregivers' actual repair (e.g., the caregiver reminding the child that their question is not valid given what has been discussed so far).

A second goal was to study the extent to which Large Language Models (LLMs) can recognize repair opportunities in children's utterances; a fundamental task these models need to solve in order to be able to provide effective repair and help in children's learning (e.g., in an personalized educational setting), in a similar way that caregivers' repair help children learn (e.g., E. V. Clark, 2020). We examine the capabilities of current LLMs to identify whether the child's question is valid or not given the previous conversational context and the word picked by the caregiver. Figures 6.1 and 6.2 demonstrate this experimental setup.

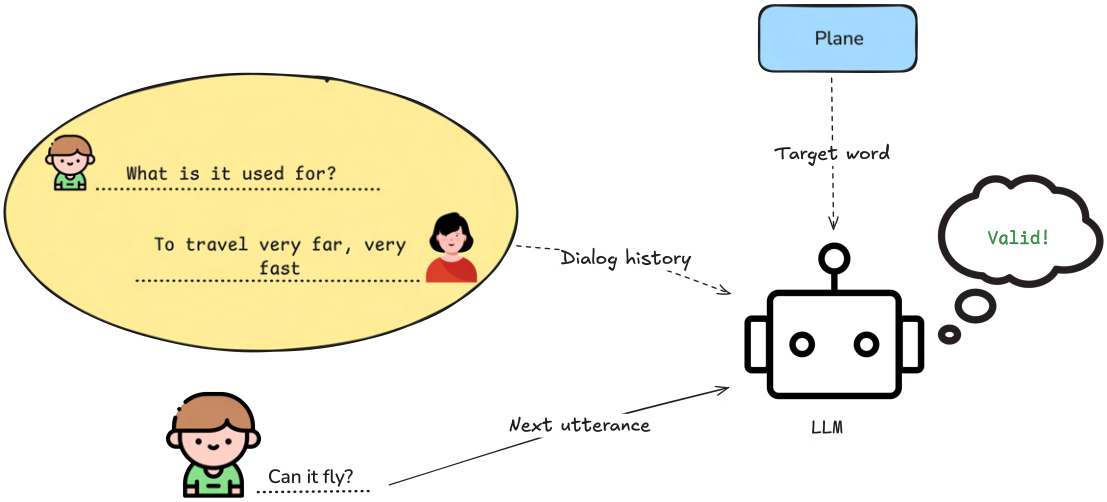

Figure 6.1.: Example of a valid question asked by the child.

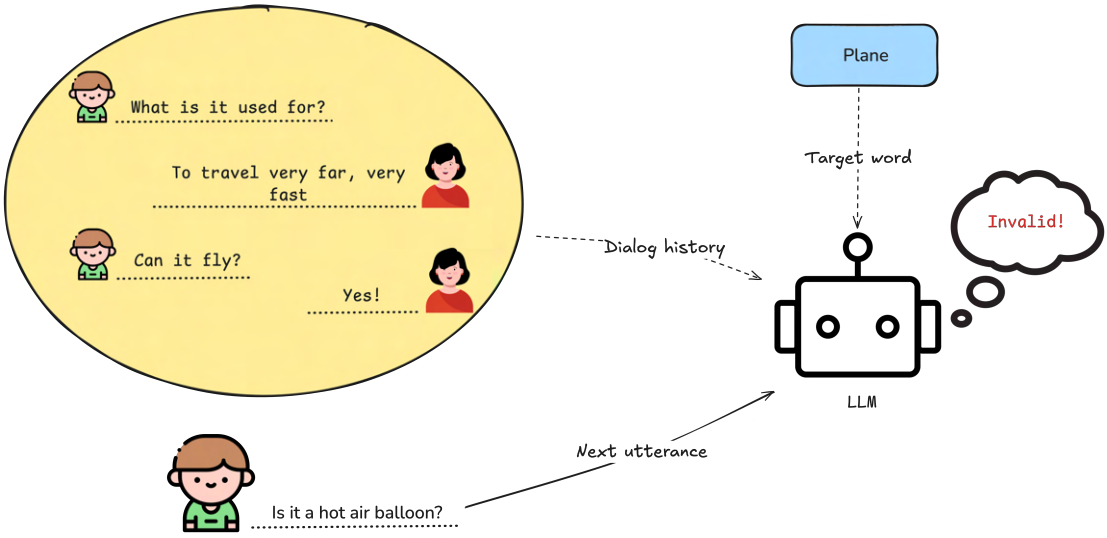

Figure 6.2.: Example of an invalid question asked by the child leading to a possible repair opportunity. Here, the question is invalid because a hot air balloon can neither travel very far nor is it very fast.

6.2. Methods

6.2.1. Data

We make use of the CHICA corpus (D. E. Goumri et al., 2024) which consists of recordings of child-caregiver face-to-face interactions in French. There are 15 dyads across three age groups in middle childhood (5 recordings per group) where the age of the child is around 7, 9 and 11 years old. The interlocutors take turns in picking a word and having the other interlocutor try to guess the word correctly by asking various questions about it.

6.2.2. Manual Annotation

After masking all the personal identifiers of the interlocutors in the data, we manually annotated all the questions asked by the child as either *"valid"* or *"invalid"* based on the previous dialog history and the word being guessed (which is known to the caregiver but not to the child). We considered only the transcript of the conversation while annotating the data. A child's question was marked as invalid if the question directly contradicted some information or a fact that was established by the parent and the child in the past dialog turns. Questions were also marked as invalid if they were repetitions of the same questions that were previously asked by the child. Two authors annotated approximately 25% of the data separately and obtained a Cohen's Kappa score of $\kappa = 0.75$. The first author annotated the rest of the data, leading to a total of $N = 739$ questions across the entirety of the 15 recordings. In addition to these repair opportunities, we also annotated whether the caregiver initiated a repair.[1]

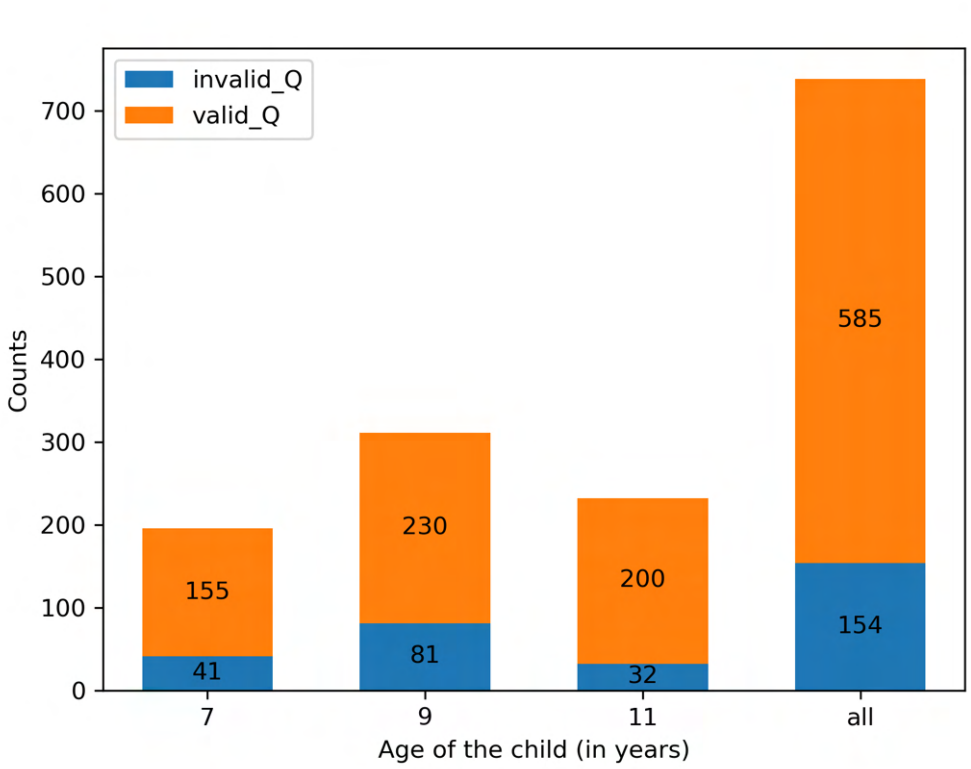

Figure 6.3.: Distribution of valid and invalid questions asked by the child across all age groups.

[1]These repairs are all other-repairs; we didn't annotate for self-repairs.

6.2.3. LLMs' testing

We tested a variety of models on our task of identifying breakdowns in child-caregiver interactions. Our selection of models are from amongst the most widely used set of models which have generally shown good performance across several tasks on various benchmarks and leader-boards (for e.g., the Chatbot Arena (Chiang et al., 2024)). All our models (except for GPT-4o) were downloaded from Ollama[2] and run locally on our system in inference mode (no fine-tuning). All the models downloaded from Ollama are 4-bit quantized versions by default. The models we tested in our study are as follows:

- Llama-3.1 8B[3]
- Llama-3.2 3B[4]
- Gemma-2 9B[5]
- Phi-3 14B[6]
- Mistral 7B[7]
- Mistral-nemo 12B[8]
- GPT-4o[9]

We used a few-shot prompting strategy to elicit from the LLMs whether a question posed by the child to the caregiver is valid or not based on all the previous relevant dialog history until that point. The prompt templates can be found in the Appendix D. We tested the LLMs both on the original French data as well as the English translation[10] to see if the language of communication affected the performance of the models.

6.3. Results and Analyses

6.3.1. Caregiver repairs vs. repair opportunities

First, we show the results of manual annotation. Figure 6.3 shows the distribution of children's breakdowns in our manually annotated data. 585 questions in total were "valid" and 154 questions were "invalid", showing insufficient grounding of prior

[2] https://ollama.com/
[3] https://huggingface.co/meta-llama/Llama-3.1-8B
[4] https://huggingface.co/meta-llama/Llama-3.2-3B
[5] https://huggingface.co/google/gemma-2-9b
[6] https://huggingface.co/microsoft/Phi-3-medium-128k-instruct
[7] https://mistral.ai/news/announcing-mistral-7b
[8] https://mistral.ai/news/mistral-nemo
[9] https://openai.com/index/hello-gpt-4o/
[10] Obtained through Google Translate and manual correction

Language Model	Balanced accuracy	
	English	French
Llama-3.1	0.60	0.59
Llama-3.2	0.57	0.57
Gemma-2	0.69	0.69
Mistral	0.61	0.62
Mistral-nemo	0.58	0.58
Phi-3	0.62	0.59
GPT-4o	**0.75**	**0.76**
Human score	0.84	

Table 6.1.: Balanced accuracy scores for few-shot prompting strategy.

information in around 26% of the time. This number varied across age groups (26% in the younger age group, 35% in the middle, 16% in the older group), but these numbers do not reflect a systematic developmental change.

The set of invalid questions represent what we call repair opportunities. We found that caregiver initiated $N = 59$ repairs, 95% of which followed invalid questions. Thus, out of a total of 154 repair opportunities, caregiver instantiated repair in about 36% of the time. Thus, while caregiver repair is not rare, it addresses only a minority of repair opportunities.

6.3.2. Can LLMs detect repair opportunities?

Table 6.1 shows the balanced accuracy scores for all the models when identifying whether a question by the child is valid or not. As seen in the table, the score of all models (except GPT-4o) are generally low and barely perform above chance, showcasing the difficulty of the task. This was the case both when using the original version in French and when using the English translation, showing that the reasons the models find the task difficult is not due to the use of French (as one may suspect, given that the models are trained primarily on English data scraped off the internet). While GPT-4o (and to some extent Gemma-2) shows a much better accuracy (around 0.75 in the case of GPT-4o), it is still lower than accuracy based on human inter-annotation agreement (Cohen's Kappa score of $\kappa = 0.75$ translates into an accuracy of 0.84).

Error analysis We analyzed the errors that these models make in their predictions. An interesting distinction to consider in the context of grounding is when the information to be grounded is a) common world knowledge that even strangers can have access to, and b) when this information is, instead, more dependent on the interlocutors sharing previous experiences that a third party may not have access to. We gave real examples of both cases from our data, shown in boxes 1 (Example 1) and 2 (Example 2). In example 1, which illustrates errors regarding common world

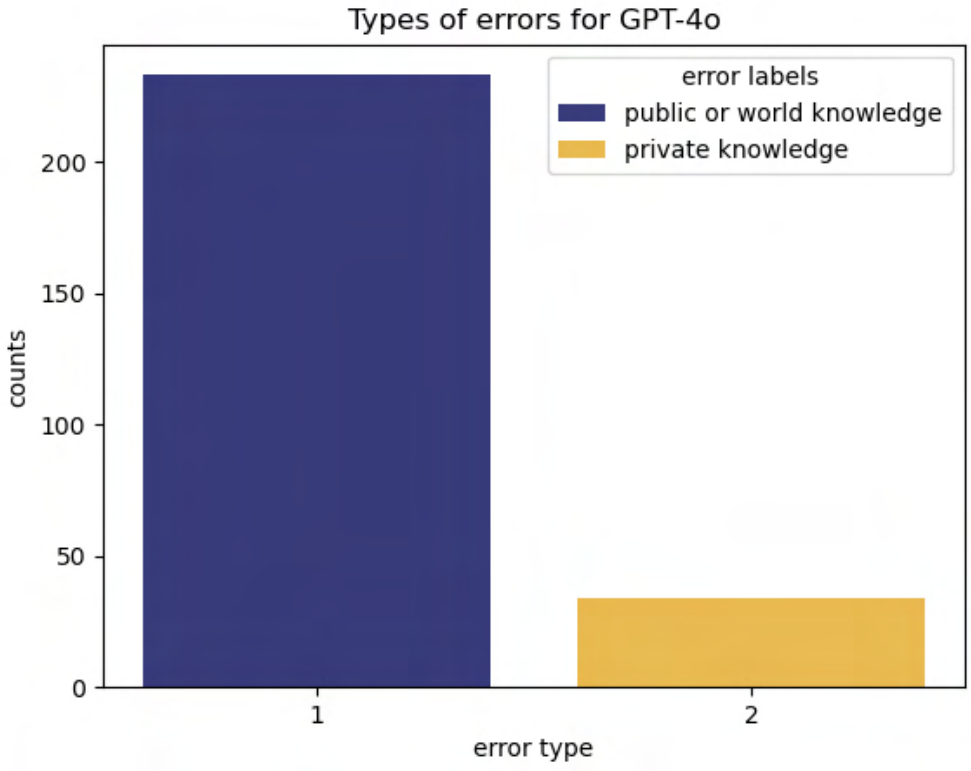

Figure 6.4.: Error analysis for GPT-4o model.

knowledge, the caregiver has a target word "book", they establish with the child that the object is rectangular, after which the child poses the question, "Does that mean a square?". The child's question can be easily classified as "invalid" by a third party. In example 2, which illustrates annotation difficulty due to private knowledge, the caregiver's target word is "cactus" and the child asks if they have the object available at home. The child's question relates to private knowledge (what the dyad has at home) and, therefore, makes the question less straightforward to categorize.

Given that the models do not have access to the interlocutors' private common ground, it could be understandable if most errors fall in this category. However, this was not the case. In fact, the overwhelming majority of the errors (Figure 6.4 for the best performing model) are related to common world knowledge (around 87%) and only a small subset concerns private knowledge shared by the interlocutor (around 13%), indicating ample room for improvement in these models regarding common world knowledge reasoning.

Example 1

Target word: **A book**

Caregiver: *Usually it is rectangular.*
Child: *What does a rectangle mean?*
Caregiver: *It's like a rectangle. And there are words inside.*
Child: *What is a rectangle again?*
Caregiver: *A rectangle is like that.*
Child: *Like that?*
Caregiver: *Like your presentation folder.*
Child: *Ok.*
Child: *Does that mean a square?* → [**Invalid**]
Caregiver: *No, a rectangle.*

— ChiCa corpus, ID_3.csv

Example 2

Target word: **A cactus**

Child: *Is it green?*
Caregiver: *Yeah, it's green.*
Child: *Is it a bit hard?*
Caregiver: *Yes, except when it rots, it becomes very soft.*
Caregiver: *It's hard and above all it has a special characteristic.*
Child: *Does it smell good?*
Caregiver: *No, not particularly.*
Child: *Is it hard?*
Caregiver: *Yes.*
Child: *Is that... I don't know.*
Caregiver: *I don't know, ask if we have any or does it grow, I don't know.*
Child: *Do we have any?* → [**Valid**]
Caregiver: *We have a whole one... In one of the planters, there are some very small ones.*

— ChiCa corpus, ID_2.csv

Does caregiver's repair indicate severity of communicative breakdown? So far, we tested the models on their ability to detect all repair opportunity and we found low to moderate performance. However, we know that caregivers initiate repair in only a minority of cases (as we saw above). One possibility is that caregivers initiate repair only for the subset of invalid questions that are more severe and risk to seriously disrupt the grounding process, in which case, we would expect the models to find it easier to classify those more apparent cases. As reported above, caregivers initiated

Language Model	Accuracy
Llama-3.1	0.60
Llama-3.2	0.55
Gemma-2	0.67
Mistral	0.63
Mistral-nemo	0.56
Phi-3	0.64
GPT-4o	0.75

Table 6.2.: Accuracy scores for repair initiating questions.

a total of $N = 59$ repairs, 56 of which followed invalid questions and 3 followed valid questions. To create a balanced testing data, we randomly sampled 56 valid questions and 3 invalid ones. Table 6.2 shows the results when we restricted our analysis to this subset. The accuracies are very similar to when the models were tested on the larger dataset; suggesting that the subset of repairs that caregivers initiate do not necessarily target more obvious cases of communicative breakdown, at least from the perspective of the LLMs we tested.

6.4. Conclusions

This study offers a first exploration into the identification of repair opportunities in child–caregiver interactions. The main finding is that caregivers address only a small portion—approximately one third—of the potential repair opportunities that arise during conversation.

We also evaluated the ability of several large language models (LLMs) to identify repair opportunities in children's utterances. Compared to human annotators, the models showed limited performance, underscoring the complexity of the task. Among the models tested, the larger, closed-weight model GPT-4o outperformed the smaller, open-weight models, consistent with prior findings on grounding-related tasks (Hakimov et al., 2025; Mohapatra, Kapadnis, et al., 2024). Error analysis revealed that GPT-4o's failures often stem from a lack of common world knowledge and/or limitations in reasoning over such knowledge.

As an initial exploration, this study comes with several limitations. Like any corpus-based analysis, as opposed to experimental approaches, it does not allow for the elicitation of specific phenomena and is constrained by what occurs naturally in the data. For example, although we annotated a relatively large number of data points (N=739), instances of our target phenomenon—invalid questions, or repair opportunities—were limited to just 154 cases. This smaller sample size limits the strength of our conclusions. Accordingly, our main finding, that parents respond to only a subset of available repair opportunities, should be interpreted with caution and awaits confirmation in future, larger-scale studies.

A key limitation in our evaluation of LLMs' ability to identify repair opportunities is that the models merely "overheard" the conversation, rather than actively participating in it. As Madureira and Schlangen (2024) rightly argue, developing common ground requires active engagement in the interaction. However, an 'overhearing' paradigm does not undermine the outcome of our approach, since the task focused solely on assessing the validity of questions based on prior dialogue context, rather than simulating the grounding process itself, a step that we did not address here.

Finally, our study relied solely on transcripts. However, the development of common ground is inherently a multimodal phenomenon. Evidence of grounding and repair initiation often appears in visual cues such as head nods, shakes, frowns, or pointing gestures. In naturalistic, free-flowing conversation, the signals for detecting and addressing communication breakdowns are typically more subtle and multimodal than what was captured here. That said, this limitation is partially mitigated by the design of the game, which required interlocutors to verbalize their repair initiations—making them almost always identifiable in the transcripts. While this reduces concerns about the internal validity of our operationalization (by focusing on the transcript), it leaves open the broader question of ecological validity.

To conclude, our corpus analysis reveals that caregivers draw on only a limited subset of the potential repair opportunities that arise during interactions with children. Moreover, we show that several large language models (LLMs) under-perform compared to humans in identifying these opportunities for repair in child–caregiver conversations. This highlights the need for further improvement, particularly in the context of applying LLMs to e-tutoring systems.

Part IV.

Discussion and conclusion

7. Discussion and Conclusion

In this chapter, we will discuss some of the challenges and results in terms of the broader context of this thesis. We will also suggest some potential avenues for research based on our studies and finally conclude by summarizing the major results from this thesis.

7.1. Turn-taking management in child-caregiver interactions

More attention needs to be given to data-driven quantitative models of turn-taking in child-caregiver interactions. These models enable us to study this phenomenon in its naturalistic environment where all the modalities play a role and where communication can be affected by the surrounding environment. Although our first study on turn-taking (see Chapter 2) was a starting step in finding evidence of children in their middle-childhood showing adult-like behavior in terms of taking the main channel or providing a backchannel in conversation, this evidence is merely correlational in nature and not causal. Further work needs to be done to confirm our findings. Since it has been argued that the protracted delay in learning turn-taking in children is due to a deficit in children being able to plan a response beforehand (Casillas et al., 2016), our evidence of adult-like maturity in turn-taking in children also implies that children in middle childhood are able to comprehend and process linguistic utterances and then formulate an appropriate response efficiently by that age.

Furthermore, considering the contribution of individual modalities in predicting turn-taking or backchanneling in our model and the importance awarded to cues from multiple modalities in the literature, we believe a better model could be developed that could better integrate these modalities like a multi-modal voice activity projection model (Onishi et al., 2023). We believe that developmental researchers could benefit by looking at the research done on turn-taking in the field of human-machine interaction (for e.g., Ekstedt & Skantze, 2022b; Inoue et al., 2024a; Roddy et al., 2018b) and leverage those models on child-caregiver interactions as predictive models of their behavior to support their findings and hypothesis from experimental and observational studies. In addition, these models allow us to easily conduct ablation studies which could help us to tease apart the contribution of the various cues in identifying turn-switches and backchannels.

7.2. Developmental dynamics of early child-caregiver interactions in terms of coherence

While our studies in Chapters 4 and 5 looked at the interaction dynamics of coherence in early child-caregiver conversations, we only analyzed data from children between the ages of 14-32 months. As we can see in Figure 4.4, there is still a gap in the coherence levels of the 32 month old child and the caregiver indicating that the child is still developing this skill. Thus, our study could be extended to older children to get the complete developmental trajectory of coherence in children.

Our studies restricted our analysis of coherence at turn-switches between the child and the caregiver; however, for a more comprehensive analysis of the interactional dynamics, we need to take into account the coherence for even consecutive utterances within the same turn from the same interlocutor and whether these utterances change the overall communicative intent of the interlocutor. For this purpose, we need to first devise a way to annotate communicative intents for the interlocutors at the turn level and not just the utterance level.

An interesting finding from our study was the dissociation between the frequency and coherence of several individual categories of communicative intents. This finding calls for delving deeper in the annotations to identify the reason for this dissociation.

7.3. Moving beyond repairs in conversational grounding

While repairs are an important mechanism in developing the common ground, there are other mechanisms such as linguistic entrainment and backchannels which are also relevant for improving our shared understanding (H. H. Clark & Brennan, 1991; Fusaroli et al., 2017; Pickering & Garrod, 2004; Yngve, 1970). Therefore, for a complete understanding of the process of grounding information, we need to take into account all of them and their interdependent nature. Although the conversational setting in Chapter 6 worked in our favor by ensuring that repairs were prominent in the data as a means of developing the common ground, it also limited our study since studies have shown that different conversational settings employ different degrees of the various mechanisms involved in developing the common ground (Dideriksen et al., 2023). Presumably, in a more free-style conversation, the number of both repair opportunities as well as the actual repairs by the caregiver would drop; however, this needs to be further investigated.

7.4. On using ML as a tool

Care needs to be taken while using ML models for automatically annotating data that the model's predictions are closely aligned with human judgments of the phenomena under consideration. For this reason, one can't be fully rid of the laborious task of manually annotating some small amount of data which still needs to be hand-coded by researchers[1] before training these models to automate the task. The importance of having a separate test dataset which the model hasn't come across during it's training cannot be overstated as this helps us test the generalization capabilities of the model to unseen data[2]. The quantity of data that needs to be collected and annotated depends on the type of ML model being used as well as the complexity of the phenomenon that we are trying to model. There is no fixed rule or universal guide as to how much data needs to be collected for a particular task at hand; the best approach here is the simplest approach of trial and error where you train the model with increasing amounts of data until you find that the performance of the model has plateaued.

Another important point to remember is that these models simply learn patterns and distribution of the data that they come across during their training so care must be taken to not use these models for annotating data that falls outside of the distribution that these models have been trained on as their predictions in this case cannot be trusted. For example, in Chapter 4, we show how the model fails to generalize on data from children above the age of 32 months since the model doesn't come across this kind of data during its training. A useful rule of thumb here is to randomly sample small amounts of data from the larger corpus that you are planning on automatically annotating and validating it with human judgments to ensure the validity of your model and automatic annotations.

7.5. Challenges and considerations while using ML as a predictive model

An important consideration to always keep in mind while using ML as a predictive model is that their findings are purely correlational in nature (mostly). Further evidence is required from either controlled experimental studies or other kinds of studies before one can claim a strong causal relationship between the variables they are modeling. For instance, in Chapter 2, since the model that we use is kind of a black-box (as are most modern neural-networks) although we find evidence in children for similar turn-taking behavior as adults, we can't be sure that they are indeed using the same coordination mechanisms. The cues and the modalities that the children are using might be different than the ones used by the adult. We can't know for sure until a more deeper analysis is conducted either by some experimental studies where the

[1] Also known as the training and the test set in ML parlance.

[2] More often than not, models can overfit on the training data giving excellent scores for all the evaluation metrics whereas in reality they just memorize the training data and perform poorly on unseen data

modalities and cues can be isolated or by using some simpler, more interpretable models. This problem — also known as the mimicry fallacy — alludes to the fact that a high performance (or performance similar to humans) by the model on a given task doesn't necessarily mean that it is using the same cognitive mechanisms or internal representations as humans do for the same task (see also Bowers et al., 2023; Lake et al., 2017).

One way of gaining causal insights from ML models can be with the use of probing (e.g., Linzen & Baroni, 2021; Manning et al., 2020; Pavlick, 2022) or by relating model representations with representations in the human brain obtained from neurophysiological recordings on the same task (e.g., Hosseini et al., 2024; McGrath et al., 2024; Schrimpf et al., 2021).

All ML models come with their own set of inductive biases[3] on account of their architecture or underlying algorithm which influences their predictions. For instance, convolutional neural networks have a locality bias whereas recurrent neural networks have a sequential ordering bias. Thus, one needs to be aware of these biases and how it might influence the bearing of the findings that we glean from the model's predictions.

7.6. Conclusion

In this thesis, we explored the development of communicative coordination in children on three different levels: i) turn-taking management, ii) coherence in conversation, and iii) conversational grounding. For our analyses, we utilised ML models in two ways: as a tool for automatically annotating data (see Chapters 3, 4 and 6) and as a predictive model of children's coordination (see Chapter 2).

We find evidence of adult-like behavior in children in mid-childhood in terms of turn-taking and backchanneling. We also find evidence that children in their early childhood wait for the caregiver to complete their turn before speaking themselves. We map out the communicate landscape of early child-caregiver interactions in terms of their communicative intents and coherency. We identify the increasing developmental trajectory of coherence in children between the ages of 20 and 32 months. Finally, we show that caregivers seize the opportunity to repair misunderstandings in the conversation in a very few cases as compared to the overall opportunities for repair that present themselves during the conversation. While these findings are important, our discussion in the previous sections show that this thesis barely scratches the surface of our understanding of the development of communicative coordination in children. A lot more work needs to be done to fully understand how these skills are used effectively by children so that educational programs and interventions can be designed to help children who are lagging in these skills (for e.g., Abbot-Smith et al., 2023).

As a part of this thesis, we showcase how ML models can be used to perform large-scale as well as more ecologically valid research on conversational skills in child-

[3]These biases are different from human inductive biases.

caregiver interactions. We demonstrate this by providing models for annotating TRPs in caregiver utterances (Chapter 3), models for annotating coherence in early child-caregiver interactions (Chapter 4), models for annotating repair opportunities in child-caregiver interactions (Chapter 6) and a predictive model for modeling turn-taking in child-caregiver interactions (Chapter 2).

Bibliography

Abbot-Smith, K., Dockrell, J., Sturrock, A., Matthews, D., & Wilson, C. (2023). Topic maintenance in social conversation: What children need to learn and evidence this can be taught. *First Language*, *43*(6), 614–642 (cit. on pp. 35, 64, 65, 78, 81, 83, 97, 118).

Abbot-Smith, K., Matthews, D., Bannard, C., Nice, J., Malkin, L., Williams, D. M., & William, H. (2024). Conversational topic maintenance and related cognitive abilities in autistic and neurotypical children. *Autism* (cit. on p. 33).

Adams, C., Lockton, E., Freed, J., Gaile, J., Earl, G., McBean, K., Nash, M., Green, J., Vail, A., & Law, J. (2012). The Social Communication Intervention Project: A randomized controlled trial of the effectiveness of speech and language therapy for school-age children who have pragmatic and social communication problems with or without autism spectrum disorder. *International Journal of Language & Communication Disorders*, *47*(3), 233–244 (cit. on p. 83).

Agrawal, A., Favre, B., & Fourtassi, A. (2025a). Mapping the communicative landscape of early child-caregiver dialogue (cit. on p. 80).

Agrawal, A., Favre, B., & Fourtassi, A. (2025b). Identifying repair opportunities in child-caregiver interactions. *Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue – Full Papers*, 48–59 (cit. on p. 102).

Agrawal, A., Liu, J., Bodur, K., Favre, B., & Fourtassi, A. (2023). Development of Multimodal Turn Coordination in Conversations: Evidence for Adult-like behavior in Middle Childhood. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45) (cit. on pp. 43, 56, 64).

Agrawal, A., Nikolaus, M., Favre, B., & Fourtassi, A. (2024). Automatic Coding of Contingency in Child-Caregiver Conversations. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 1856–1870). ELRA; ICCL. (Cit. on p. 63).

Alishahi, A. (2010). *Computational modeling of human language acquisition*. Morgan & Claypool Publishers. (Cit. on p. 23).

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, *34*(4), 351–366 (cit. on p. 28).

Apriliani, H., & Muslim, A. B. (2021). Grounding in Online Communication, 209–216 (cit. on p. 36).

Arkowitz, H., Lichtenstein, E., McGovern, K., & Hines, P. (1975). The behavioral assessment of social competence in males. *Behavior therapy*, *6*(1), 3–13 (cit. on p. 35).

Auer, P. (2021). Turn-allocation and gaze: A multimodal revision of the "current-speaker-selects-next" rule of the turn-taking system of conversation analysis. *Discourse Studies, 23*(2), 117–140 (cit. on pp. 29, 56).

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 (cit. on pp. 20, 45, 97).

Baker, M., Hansen, T. G. B., Joiner, R., & Traum, D. (1999). The role of grounding in collaborative learning tasks (cit. on pp. 36, 104)
[TLDR] A unifying perspective of mutual understanding mediated by material and semiotic tools that can be used for analysis as well as for design of collaborative learning tasks, especially those that are carried out via computer-mediated communication are built.

Baltrusaitis, T., Robinson, P., & Morency, L.-P. (2016). OpenFace: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10 (cit. on p. 22).

Barzilay, R., & Lapata, M. (2008). Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics, 34*(1), 1–34 (cit. on pp. 36, 65, 70).

Bates, E., Camaioni, L., & Volterra, V. (1975). The Acquisition of Performatives Prior to Speech. *Merrill-Palmer Quarterly of Behavior and Development, 21*(3), 205–226 (cit. on pp. 34, 81, 82, 97).

Bateson, M. C. (1975). Mother-infant exchanges: The epigenesis of conversational interaction. *Annals of the New York Academy of sciences, 263*(1), 101–113 (cit. on p. 31).

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology, 79*(6), 941–952 (cit. on pp. 30, 64).

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of communication, 52*(3), 566–580 (cit. on p. 30).

Beebe, B., Alson, D., Jaffe, J., Feldstein, S., & Crown, C. (1988). Vocal congruence in mother-infant play. *Journal of psycholinguistic research, 17*, 245–259 (cit. on p. 31).

Benoit, P. J. (1979). *A descriptive study of coherence in naturally-occurring and experimentally structured conversations of preschool children*. Wayne State University. (Cit. on p. 33).

Benotti, L., & Blackburn, P. (2021). Grounding as a Collaborative Process. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 515–531 (cit. on pp. 39, 104).

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramírez-Esparza, N., R. Hamrick, L., Marklund, E., Kalashnikova, M., Guez, A., Casillas, M., Benetti, L., Alphen, P. v., & Cristia, A. (2023). Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences, 120*(52) (cit. on pp. 79, 95).

Bergey, C., Marshall, Z., DeDeo, S., & Yurovsky, D. (2022). Learning Communicative Acts in Children's Conversations: A Hidden Topic Markov Model Analysis of the CHILDES Corpora. *Topics in Cognitive Science*, *14*(2), 388–399 (cit. on p. 82).

Bergey, C. A., O'Keeffe, M. E., & Hawkins, R. (2024). A longitudinal analysis of children's communicative acts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0) (cit. on p. 82).

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 (cit. on p. 47).

Birhane, A. (2022, June). Automating Ambiguity: Challenges and Pitfalls of Artificial Intelligence. (Cit. on p. 23).

Birhane, A., Dehdashtian, S., Prabhu, V., & Boddeti, V. (2024). The dark side of dataset scaling: Evaluating racial classification in multimodal models. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1229–1244 (cit. on p. 23).

Black, B., & Hazen, N. L. (1990). Social status and patterns of communication in acquainted and unacquainted preschool children. *Developmental Psychology*, *26*(3), 379–387 (cit. on pp. 19, 83).

Blain-Brière, B., Bouchard, C., & Bigras, N. (2014). The role of executive functions in the pragmatic skills of children age 4–5. *Frontiers in Psychology*, *5* (cit. on pp. 33, 83).

Bloom, L., Rocissano, L., & Hood, L. (1976). Adult-child discourse: Developmental interaction between information processing and linguistic knowledge. *Cognitive Psychology*, *8*(4), 521–552 (cit. on pp. 33, 35, 64, 65, 78, 81, 83, 97).

Bodur, K., Nikolaus, M., Kassim, F., Prévot, L., & Fourtassi, A. (2021). Chico: A multimodal corpus for the study of child conversation. *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 158–163 (cit. on pp. 15, 46, 146–148).

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* (cit. on pp. 45, 53, 64).

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, *52*, 46–57 (cit. on pp. 29, 30, 56).

Bögels, S., & Torreira, F. (2021). Turn-end estimation in conversational turn-taking: The roles of context and prosody. *Discourse processes*, *58*(10), 903–924 (cit. on p. 29).

Bohn, M., & Frank, M. C. (2019). The Pervasive Role of Pragmatics in Early Language. *Annual Review of Developmental Psychology*, *1*(Volume 1, 2019), 223–249 (cit. on p. 81).

Boland, J. E., Fonseca, P., Mermelstein, I., & Williamson, M. (2022). Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General*, *151*, 1272–1282 (cit. on p. 54).

Bornstein, M. H., Tamis-LeMonda, C. S., Hahn, C.-S., & Haynes, O. M. (2008). Maternal responsiveness to young children at three ages: Longitudinal analysis of a multidimensional, modular, and specific parenting construct. *Developmental psychology, 44*(3), 867 (cit. on p. 35).

Bornstein, M. H., Tamis-LeMonda, C. S., & Haynes, O. M. (1999). First words in the second year: Continuity, stability, and models of concurrent and predictive correspondence in vocabulary and verbal responsiveness across age and context. *Infant Behavior and Development, 22*(1), 65–85 (cit. on p. 35).

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences, 46*, e385 (cit. on p. 118).

Brennan, S. E. (1991). Conversation with and through computers. *User modeling and user-adapted interaction, 1*, 67–86 (cit. on p. 20).

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science, 1*(2), 274–291 (cit. on p. 20).

Brent, M. R. (1997). *Computational approaches to language acquisition*. MIT Press. (Cit. on p. 23).

Bruner, J. (1985). Child's Talk: Learning to Use Language. *Child Language Teaching and Therapy, 1*(1), 111–114 (cit. on pp. 38, 79, 81, 82, 97).

Brunner, L. J. (1979). Smiles can be back channels. *Journal of Personality and Social Psychology, 37*, 728–734 (cit. on p. 48).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, 77–91 (cit. on p. 23).

Butler, L. P., Ronfard, S., & Corriveau, K. H. (Eds.). (2020). *The Questioning Child: Insights from Psychology and Education*. Cambridge University Press. (Cit. on p. 97).

Cabiddu, F., Nikolaus, M., & Fourtassi, A. (2025). Comparing children and large language models in word sense disambiguation: Insights and challenges. *Language Development Research, 5*(1) (cit. on p. 96).

Callanan, M., Legare, C. H., Sobel, D. M., Jaeger, G. J., Letourneau, S., McHugh, S. R., Willard, A., Brinkman, A., Finiasz, Z., Rubio, E., Barnett, A., Gose, R., Martin, J. L., Meisner, R., & Watson, J. (2020). Exploration, Explanation, and Parent–Child Interaction in Museums. *Monographs of the Society for Research in Child Development, 85*(1), 7–137 (cit. on p. 97).

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development, 7*(2), 213–233 (cit. on p. 97).

Cameron-Faulkner, T. (2014, June). The development of speech acts. In *Pragmatic Development in First Language Acquisition* (pp. 37–52). John Benjamins. (Cit. on pp. 82, 97).

Capps, L., Kehres, J., & Sigman, M. (1998). Conversational Abilities Among Children with Autism and Children with Developmental Delays. *Autism*, *2*(4), 325–344 (cit. on pp. 33, 83).

Casillas, M., Bobb, S. C., & Clark, E. V. (2016). Turn-taking, timing, and planning in early language acquisition. *Journal of Child Language*, *43*(6), 1310–1337 (cit. on pp. 31, 57, 60, 64, 115).

Casillas, M., & Frank, M. (2013). The development of predictive processes in children's discourse understanding. *Proceedings of the annual meeting of the Cognitive Science Society*, *35*(35) (cit. on p. 31).

Casillas, M., & Hilbrink, E. (2020). 3. Communicative act development. In K. P. Schneider & E. Ifantidou (Eds.), *Developmental and Clinical Pragmatics* (pp. 61–88). De Gruyter Mouton. (Cit. on p. 82).

Castillo-López, G., de Chalendar, G., & Semmar, N. (2025). A survey of recent advances on turn-taking modeling in spoken dialogue systems. In M. I. Torres, Y. Matsuda, Z. Callejas, A. del Pozo, & L. F. D'Haro (Eds.), *Proceedings of the 15th international workshop on spoken dialogue systems technology* (pp. 254–271). Association for Computational Linguistics. (Cit. on p. 31).

Cathcart, N., Carletta, J., & Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, 51–58 (cit. on pp. 44, 48).

Cervone, A., & Riccardi, G. (2020). Is this Dialogue Coherent? Learning from Dialogue Acts and Entities. *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 162–174 (cit. on pp. 69, 70).

Cervone, A., Stepanov, E., & Riccardi, G. (2018). Coherence Models for Dialogue. *Interspeech 2018*, 1011–1015 (cit. on pp. 36, 65, 70).

Çetinçelik, M., Rowland, C. F., & Snijders, T. M. (2021). Do the eyes have it? a systematic review on the role of eye gaze in infant language development. *Frontiers in Psychology, Volume 11 - 2020* (cit. on pp. 38, 105).

Chandu, K. R., Bisk, Y., & Black, A. W. (2021). Grounding 'grounding' in NLP. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 4283–4305). Association for Computational Linguistics. (Cit. on pp. 39, 40).

Cheng, Q., İnan, M., Mbarki, R., Grmek, G., Choi, T., Sun, Y., Persaud, K., Wang, J., & Alikhani, M. (2024, October). Learning Multimodal Cues of Children's Uncertainty. (Cit. on p. 39).

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., & Stoica, I. (2024). Chatbot arena: An open platform for evaluating llms by human preference. *Proceedings of the 41st International Conference on Machine Learning* (cit. on p. 108).

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 6 (cit. on p. 72).

Chieng, A. C. J., Wynn, C. J., Wong, T. P., Barrett, T. S., & Borrie, S. A. (2024). Lexical alignment is pervasive across contexts in non-weird adult–child interactions. *Cognitive Science*, *48*(3), e13417 (cit. on p. 64).

Chouinard, M. M., Harris, P. L., & Maratsos, M. P. (2007). Children's Questions: A Mechanism for Cognitive Development. *Monographs of the Society for Research in Child Development*, *72*(1), i–129 (cit. on pp. 35, 97).

Clark, E. V. (2015). Common ground. *The handbook of language emergence*, 328–353 (cit. on p. 38).

Clark, E. V. (2018). Conversation and Language Acquisition: A Pragmatic Approach. *Language Learning and Development*, *14*(3), 170–185 (cit. on pp. 38, 79, 81, 97, 104).

Clark, E. V. (2020). Conversational repair and the acquisition of language. *Discourse Processes*, *57*(5-6), 441–459 (cit. on pp. 20, 38, 64, 97, 102, 105).

Clark, H. H. (1992). *Arenas of language use*. University of Chicago Press. (Cit. on p. 19).

Clark, H. H. (1996). *Using Language*. Cambridge University Press. (Cit. on pp. 20, 36, 81, 102, 104).

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition.* (pp. 127–149). American Psychological Association. (Cit. on pp. 37, 104, 116).

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81 (cit. on pp. 39, 104).

Clark, H. H., & Schaefer, E. F. (1989). Contributing to Discourse. *Cognitive Science*, *13*(2), 259–294 (cit. on pp. 36–38, 104).

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39 (cit. on pp. 19, 20).

Cruz Blandón, M. A., Cristia, A., & Räsänen, O. (2023). Introducing meta-analysis in the evaluation of computational models of infant language development. *Cognitive Science*, *47*(7), e13307 (cit. on p. 23).

Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., Casillas, M., de Barbaro, K., Bang, J. Y., & Weisleder, A. (2020). Longform recordings of everyday life: Ethics for best practices. *Behavior Research Methods*, *52*(5), 1951–1969 (cit. on p. 98).

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., & Batra, D. (2017). Visual dialog. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 326–335 (cit. on p. 40).

De Felice, S., Hamilton, A. F. d. C., Ponari, M., & Vigliocco, G. (2023). Learning from others is good, with others is better: The role of social interaction in human acquisition of new knowledge. *Philosophical Transactions of the Royal Society B*, *378*(1870), 20210357 (cit. on p. 19).

De Ruiter, J.-P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*(3), 515–535 (cit. on pp. 29, 30).

De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. (2017). GuessWhat?! Visual Object Discovery through Multi-modal Dialogue. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4466–4475 (cit. on p. 40).

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). Llm.int8(): 8-bit matrix multiplication for transformers at scale. *Proceedings of the 36th International Conference on Neural Information Processing Systems* (cit. on p. 79).

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 (cit. on p. 45).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (cit. on pp. 65, 71).

Dideriksen, C., Christiansen, M. H., Tylén, K., Dingemanse, M., & Fusaroli, R. (2023). Quantifying the interplay of conversational devices in building mutual understanding. *Journal of Experimental Psychology: General*, *152*(3), 864–889 (cit. on pp. 102, 116).

Dingemanse, M. (2024). Interjections at the heart of language. *Annual Review of Linguistics*, *10*(1), 257–277 (cit. on p. 30).

Dingemanse, M., & Enfield, N. J. (2024). Interactive repair and the foundations of language. *Trends Cogn. Sci.*, *28*(1), 30–42 (cit. on p. 64).

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., & Enfield, N. J. (2015). Universal Principles in the Repair of Communication Problems. *PLOS ONE*, *10*(9), e0136100 (cit. on pp. 39, 104).

Donnellan, E., Bannard, C., McGillion, M. L., Slocombe, K. E., & Matthews, D. (2020). Infants' intentionally communicative vocalizations elicit responses from caregivers and are the best predictors of the transition to language: A longitudinal investigation of infants' vocalizations, gestures and word production. *Developmental science*, *23*(1), e12843 (cit. on p. 35).

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*, *92*(2), 609–625 (cit. on p. 79).

Dore, J. (1973). A Developmental Theory of Speech Act Production*. *Transactions of the New York Academy of Sciences*, *35*(8 Series II), 623–630 (cit. on p. 34).

Dore, J. (1974). A pragmatic description of early language development. *Journal of Psycholinguistic Research*, *3*(4), 343–350 (cit. on pp. 82, 97).

Dorval, B., Eckerman, C. O., & Ervin-Tripp, S. (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), 1–91 (cit. on pp. 33, 81–83, 85, 97).

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*, 283–292 (cit. on pp. 29, 44, 48).

Dunn, J., & Shatz, M. (1989). Becoming a conversationalist despite (or because of) having an older sibling. *Child development*, 399–410 (cit. on p. 31).

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition, 173*, 43–59 (cit. on p. 23).

Ekstedt, E., & Skantze, G. (2020). TurnGPT: A Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2981–2990 (cit. on pp. 32, 55–58).

Ekstedt, E., & Skantze, G. (2022a). How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In O. Lemon, D. Hakkani-Tur, J. J. Li, A. Ashrafzadeh, D. H. Garcia, M. Alikhani, D. Vandyke, & O. Dušek (Eds.), *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 541–551). Association for Computational Linguistics. (Cit. on p. 32).

Ekstedt, E., & Skantze, G. (2022b). Voice Activity Projection: Self-supervised Learning of Turn-taking Events. *Interspeech 2022*, 5190–5194 (cit. on pp. 32, 115).

Elmlinger, S. L., Goldstein, M. H., & Casillas, M. (2023). Immature vocalizations simplify the speech of tseltal mayan and u.s. caregivers. *Topics in Cognitive Science, 15*(2), 315–328 (cit. on p. 79).

Elmlinger, S. L., Schwade, J. A., Vollmer, L., & Goldstein, M. H. (2023). Learning how to learn from social feedback: The origins of early vocal development. *Developmental Science, 26*(2), e13296 (cit. on p. 81).

Erel, Y., Potter, C. E., Jaffe-Dax, S., Lew-Williams, C., & Bermano, A. H. (2022). Icatcher: A neural network approach for automated coding of young children's eye movements. *Infancy, 27*(4), 765–779 (cit. on p. 22).

Erel, Y., Shannon, K. A., Chu, J., Scott, K., Kline Struhl, M., Cao, P., Tan, X., Hart, P., Raz, G., Piccolo, S., 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 (cit. on p. 22).

Ervin-Tripp, S. (1979). Children's verbal turn-taking. *Developmental pragmatics*, 391–414 (cit. on p. 31).

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & 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 (cit. on p. 48).

Feldman, R. (2007). Parent–infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions. *Journal of Child Psychology and Psychiatry, 48*(3-4), 329–354 (cit. on p. 81).

Feldman, R. (2015). Sensitive periods in human social development: New insights from research on oxytocin, synchrony, and high-risk parenting. *Development and psychopathology*, *27*(2), 369–395 (cit. on p. 35).

Ferguson, B., & Lew-Williams, C. (2016). Communicative signals support abstract rule learning by 7-month-old infants. *Scientific reports*, *6*(1), 25434 (cit. on p. 35).

Fernandez, R., & Grimm, R. (2014). Quantifying Categorical and Conceptual Convergence in Child-Adult Dialogue. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36) (cit. on p. 83).

Ferreira, V. S., & Bock, K. (2006). The functions of structural priming. *Language and cognitive processes*, *21*(7-8), 1011–1029 (cit. on p. 20).

Ford, C., & Thompson, S. (1996). Interactional units in conversation: Syntactic, intonational and pragmatic resources. *Interaction and grammar*, (13), 134 (cit. on pp. 28, 29, 44, 48, 55, 56).

Fourtassi, A. (2023). Understanding children's multimodal conversational development: Challenges and opportunities (cit. on pp. 12, 20–22).

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning. *Psychological Science*, *20*(5), 578–585 (cit. on p. 104).

Fried, D., Tomlin, N., Hu, J., Patel, R., & Nematzadeh, A. (2023). Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12619–12640 (cit. on p. 104).

Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. *Annual Meeting of the Cognitive Science Society* (cit. on pp. 39, 104, 116).

Fusaroli, R., Weed, E., Rocca, R., Fein, D., & Naigles, L. (2023a). Caregiver linguistic alignment to autistic and typically developing children: A natural language processing approach illuminates the interactive components of language development. *Cognition*, *236*, 105422 (cit. on pp. 46, 64).

Fusaroli, R., Weed, E., Rocca, R., Fein, D., & Naigles, L. (2023b). Repeat After Me? Both Children With and Without Autism Commonly Align Their Language With That of Their Caregivers. *Cognitive Science*, *47*(11), e13369 (cit. on pp. 83, 96).

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, *36*, 13518–13529 (cit. on p. 103).

Garvey, C. (1975). Requests and responses in children's speech. *Journal of Child Language*, *2*(1), 41–63 (cit. on p. 82).

Garvey, C., & Berninger, G. (1981). Timing and turn taking in children's conversations. *Discourse processes*, *4*(1), 27–57 (cit. on p. 31).

Garvey, C., & Hogan, R. (1973). Social Speech and Social Interaction: Egocentrism Revisited. *Child Development*, *44*(3), 562–568 (cit. on p. 81).

Garzaniti, I., Pearce, G., & Stanton, J. (2011). Building friendships and relationships: The role of conversation in hairdressing service encounters (J. Finsterwalder & T. Garry, Eds.). *Managing Service Quality: An International Journal, 21*(6), 667–687 (cit. on pp. 19, 35, 64).

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics, 1*, 1–68 (cit. on p. 20).

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology, 26*(2), 248–265 (cit. on p. 35).

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. *Acoustics, speech, and signal processing, ieee international conference on, 1*, 517–520 (cit. on p. 66).

Gokaslan, A., & Cohen, V. (2019). Openwebtext corpus. (Cit. on p. 71).

Golland, D., Liang, P., & Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In H. Li & L. Màrquez (Eds.), *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 410–419). Association for Computational Linguistics. (Cit. on p. 40).

Goswami, M., Manuja, M., & Leekha, M. (2020). Towards social & engaging peer learning: Predicting backchanneling and disengagement in children. (Cit. on p. 48).

Goumri, D. E., Agrawal, A., Nikolaus, M., Vu, H. D. T., Bodur, K., Emmar, E., Armand, C., Mazzocconi, C., Gupta, S., Prévot, L., Favre, B., Becerra-Bonache, L., & Fourtassi, A. (2024). CHICA: A Developmental Corpus of Child-Caregiver's Face-to-face vs. Video Call Conversations in Middle Childhood. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3153–3164). ELRA; ICCL. (Cit. on pp. 25, 106).

Goumri, D.-E., Janssoone, T., Becerra-Bonache, L., & Fourtassi, A. (2023). Automatic detection of gaze and smile in children's video calls. *Companion Publication of the 25th International Conference on Multimodal Interaction*, 383–388 (cit. on p. 22).

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language, 25*(3), 601–634 (cit. on pp. 29, 30, 44, 56).

Grice, H. P. (1975). Logic and Conversation. In D. Davidson (Ed.), *The logic of grammar* (pp. 64–75). Dickenson Pub. Co. (Cit. on pp. 33, 64, 81).

Grice, H. P. (1957). Meaning. *Philosophical Review, 66*(3), 377–388 (cit. on p. 34).

Grice, P. (1991). *Studies in the way of words*. Harvard University Press. (Cit. on p. 33).

Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., & Fernández, R. (2019). The PhotoBook dataset: Building common ground through visually-grounded dialogue. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the*

57th annual meeting of the association for computational linguistics (pp. 1895–1910). Association for Computational Linguistics. (Cit. on p. 40).

Hakimov, S., Abdullayeva, Y., Koshti, K., Schmidt, A., Weiser, Y., Beyer, A., & Schlangen, D. (2025). Using Game Play to Investigate Multimodal and Conversational Grounding in Large Multimodal Models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5686–5718). Association for Computational Linguistics. (Cit. on pp. 40, 112).

Hale, C. M., & Tager-Flusberg, H. (2005). Social communication in children with autism: The relationship between theory of mind and discourse development. *Autism*, *9*(2), 157–178 (cit. on pp. 33, 64, 83).

Harris, R. (1996). *Signs, language, and communication: Integrational and segregational approaches*. Psychology Press. (Cit. on p. 20).

Hazen, N. L., & Black, B. (1989). Preschool peer communication skills: The role of social status and intervention context. *Child Development*, *60*(4), 867–876 (cit. on pp. 19, 35, 64).

He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *The Eleventh International Conference on Learning Representations* (cit. on pp. 70, 71, 83, 86).

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, *abs/2006.03654* (cit. on p. 71).

Helland, W. A., Lundervold, A. J., Heimann, M., & Posserud, M.-B. (2014). Stable associations between behavioral problems and language impairments across childhood–the importance of pragmatic language problems. *Research in developmental disabilities*, *35*(5), 943–951 (cit. on p. 35).

Hess, L. J., & Johnston, J. R. (1988). Acquisition of back channel listener responses to adequate messages. *Discourse Processes*, *11*(3), 319–335 (cit. on pp. 20, 44, 45, 53).

Higashinaka, R., Meguro, T., Imamura, K., Sugiyama, H., Makino, T., & Matsuo, Y. (2014). Evaluating coherence in open domain conversational systems. *Interspeech 2014*, 130–134 (cit. on p. 69).

Hilbrink, E. E., Gattis, M., & Levinson, S. C. (2015). Early developmental changes in the timing of turn-taking: A longitudinal study of mother–infant interaction. *Frontiers in psychology*, *6*, 1492 (cit. on p. 31).

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., Yust, P. K. S., & Suma, K. (2015). The Contribution of Early Communication Quality to Low-Income Children's Language Success. *Psychological Science*, *26*(7), 1071–1083 (cit. on p. 35).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, *9*(8), 1735–1780 (cit. on p. 48).

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639–652 (cit. on pp. 29, 47).

Hömke, P., Holler, J., & Levinson, S. C. (2018). Eye blinks are perceived as communicative signals in human face-to-face interaction. *PloS one*, *13*(12), e0208030 (cit. on p. 30).

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024). Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, *5*(1), 43–63 (cit. on p. 118).

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (pp. 2790–2799, Vol. 97). PMLR. (Cit. on p. 79).

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In A. Bisazza & O. Abend (Eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 624–646). Association for Computational Linguistics. (Cit. on p. 96).

Ilinykh, N., Zarrieß, S., & Schlangen, D. (2019). Meet up! a corpus of joint activity dialogues in a visual environment. *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers* (cit. on p. 40).

Inoue, K., Jiang, B., Ekstedt, E., Kawahara, T., & Skantze, G. (2024a, January). Real-time and Continuous Turn-taking Prediction Using Voice Activity Projection. (Cit. on pp. 32, 115).

Inoue, K., Jiang, B., Ekstedt, E., Kawahara, T., & Skantze, G. (2024b). Multilingual turn-taking prediction using voice activity projection. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 11873–11883). ELRA; ICCL. (Cit. on p. 32).

Inoue, K., Lala, D., Skantze, G., & Kawahara, T. (2025). Yeah, un, oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 conference of the nations of the americas chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 7171–7181). Association for Computational Linguistics. (Cit. on p. 32).

Isaacs, S. (2013). *Social development in young children*. Routledge. (Cit. on p. 31).

Jain, V., & Leekha, M. (2021). Exploring semi-supervised learning for predicting listener backchannels. *Conference on Human Factors in Computing Systems - Proceedings* (cit. on p. 48).

Jasnow, M., & Feldstein, S. (1986). Adult-like temporal characteristics of mother-infant vocal interactions. *Child development*, 754–761 (cit. on p. 31).

Jeknic, I., Schlangen, D., & Koller, A. (2024). A dialogue game for eliciting balanced collaboration. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, & K. Komatani (Eds.), *Proceedings of the 25th annual meeting of the special interest group on discourse and dialogue* (pp. 477–489). Association for Computational Linguistics. (Cit. on p. 40).

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7b. (Cit. on p. 79).

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 (cit. on p. 46).

Johansson, M., & Skantze, G. (2015). Opportunities and obligations to take turns in collaborative multi-party human-robot interaction. *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 305–314 (cit. on p. 56).

Jokinen, K., Schneider, P., & Mori, T. (2024). Towards harnessing large language models for comprehension of conversational grounding. *CoRR* (cit. on p. 39).

Kaukomaa, T., Peräkylä, A., & Ruusuvuori, J. (2013). Turn-opening smiles: Facial expression constructing emotional transition in conversation. *Journal of Pragmatics*, *55*, 21–42 (cit. on p. 29).

Kaukomaa, T., Peräkylä, A., & Ruusuvuori, J. (2014). Foreshadowing a problem: Turn-opening frowns in conversation. *Journal of Pragmatics*, *71*, 132–147 (cit. on p. 29).

Keenan, E. O., & Klein, E. (1975). Coherency in children's discourse. *Journal of Psycholinguistic Research*, *4*(4), 365–380 (cit. on pp. 33, 35, 64, 65, 78, 81, 83, 97).

Keitel, A., Prinz, W., Friederici, A. D., Von Hofsten, C., & Daum, M. M. (2013). Perception of conversations: The importance of semantics and intonation in children's development. *Journal of Experimental Child Psychology*, *116*(2), 264–277 (cit. on p. 31).

Kelley, K. (1967). Early syntactic acquisition (tech. rep. no. p-3719). *Santa Monica, California: Rand Corp* (cit. on p. 23).

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*, 22–63 (cit. on pp. 29, 48, 56).

Kendrick, K. H., Holler, J., & Levinson, S. C. (2023). Turn-taking in human face-to-face interaction is multimodal: Gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1875), 20210473 (cit. on p. 29).

Kennington, C., & Schlangen, D. (2015). Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In C. Zong & M. Strube (Eds.), *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 292–301). Association for Computational Linguistics. (Cit. on p. 40).

Ketelaars, M. P., Cuperus, J., Jansonius, K., & Verhoeven, L. (2010). Pragmatic language impairment and associated behavioural problems. *International Journal of Language & Communication Disorders*, *45*(2), 204–214 (cit. on p. 35).

Kim, J.-H., Kitaev, N., Chen, X., Rohrbach, M., Zhang, B.-T., Tian, Y., Batra, D., & Parikh, D. (2019). CoDraw: Collaborative drawing as a testbed for grounded

goal-driven communication. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6495–6513). Association for Computational Linguistics. (Cit. on p. 40).

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and speech*, *41*(3-4), 295–321 (cit. on p. 29).

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*(1), 110–120 (cit. on p. 81).

Kumar, H., Agarwal, A., Dasgupta, R., & Joshi, S. (2018). Dialogue Act Sequence Labeling Using Hierarchical Encoder With CRF. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1) (cit. on p. 36).

Kurkul, K. E., & Corriveau, K. H. (2018). Question, Explanation, Follow-Up: A Mechanism for Learning From Others? *Child Development*, *89*(1), 280–294 (cit. on pp. 81, 97).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253 (cit. on p. 118).

Langley, P., & Carbonell, J. G. (1987). Language acquisition and machine learning. In *Mechanisms of language acquisition*. (Cit. on p. 23).

Lavechin, M., de Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E., & Cristia, A. (2024). Modeling early phonetic acquisition from child-centered audio data. *Cognition*, *245*, 105734 (cit. on pp. 96, 98).

Lee, S., Schulz, H., Atkinson, A., Gao, J., Suleman, K., El Asri, L., Adada, M., Huang, M., Sharma, S., Tay, W., & Li, X. (2019). Multi-domain task-completion dialog challenge. *Dialog System Technology Challenges 8* (cit. on p. 57).

LERNER, G. H. (2003). Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society*, *32*(2), 177–201 (cit. on p. 28).

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press. (Cit. on p. 20).

Levinson, S. C. (2016). Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, *20*(1), 6–14 (cit. on p. 64).

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6* (cit. on pp. 29, 30).

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Interspeech*, *2011*, 3081–3084 (cit. on p. 20).

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In G. Kondrak & T. Watanabe (Eds.), *Proceedings of the eighth international joint conference on natural language processing (volume 1: Long papers)* (pp. 986–995). Asian Federation of Natural Language Processing. (Cit. on p. 57).

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., & Stoica, I. (2018). Tune: A research platform for distributed model selection and training. *CoRR* (cit. on p. 49).

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212 (cit. on p. 118).

Liu, J., Nikolaus, M., Bodur, K., & Fourtassi, A. (2022). Predicting backchannel signaling in child-caregiver multimodal conversations. *Companion Publication of the 2022 International Conference on Multimodal Interaction*, 196–200 (cit. on pp. 23, 45, 48, 50, 51, 53, 64).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, July). RoBERTa: A Robustly Optimized BERT Pretraining Approach. (Cit. on p. 71).

Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), 1191 (cit. on p. 23).

Long, B. L., Kachergis, G., Agrawal, K., & Frank, M. C. (2022). A longitudinal analysis of the social information in infants' naturalistic visual experience using automated detections. *Developmental Psychology*, *58*(12), 2211 (cit. on p. 22).

López Pérez, D., Leonardi, G., Niedźwiecka, A., Radkowska, A., Rączaszek-Leonardi, J., & Tomalski, P. (2017). Combining recurrence analysis and automatic movement extraction from video recordings to study behavioral coupling in face-to-face parent-child interactions. *Frontiers in psychology*, *8*, 2228 (cit. on p. 22).

Luchkina, E., Simon, L. R., & Waxman, S. R. (2025). Catching up with icatcher: Comparing analyses of infant eye tracking based on trained human coders and icatcher+ automated gaze coding software. *Behavior Research Methods*, *57*(6), 1–9 (cit. on p. 22).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. (Cit. on p. 54).

Lutzenberger, H., De Wael, L., Omardeen, R., & Dingemanse, M. (2024). Interactional infrastructure across modalities: A comparison of repair initiators and continuers in british sign language and british english. *Sign Language Studies*, *24*(3), 548–581 (cit. on p. 30).

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed.* Lawrence Erlbaum Associates Publishers. (Cit. on pp. 26, 65, 66, 105).

MacWhinney, B. (2014, January). *The Childes Project: Tools for Analyzing Talk, Volume II: The Database* (3rd ed.). Psychology Press. (Cit. on pp. 80, 81, 84).

Madureira, B., & Schlangen, D. (2024). It couldn't help but overhear: On the limits of modelling meta-communicative grounding acts with supervised learning. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, & K. Komatani (Eds.), *Proceedings of the 25th annual meeting of the special interest*

group on discourse and dialogue (pp. 149–158). Association for Computational Linguistics. (Cit. on p. 113).

Magyari, L., & De Ruiter, J. P. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in psychology*, *3*, 376 (cit. on p. 29).

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054 (cit. on p. 118).

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*(1), 59–76 (cit. on pp. 20, 45).

Masek, L. R., McMillan, B. T. M., Paterson, S. J., Tamis-LeMonda, C. S., Golinkoff, R. M., & Hirsh-Pasek, K. (2021). Where language meets attention: How contingent interactions promote learning. *Developmental Review*, *60*, 100961 (cit. on pp. 79, 97).

Masumura, R., Asami, T., Masataki, H., Ishii, R., & Higashinaka, R. (2017). Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. *Interspeech*, *2017*, 1661–1665 (cit. on p. 32).

Masumura, R., Tanaka, T., Ando, A., Ishii, R., Higashinaka, R., & Aono, Y. (2018). Neural dialogue context online end-of-turn detection. In K. Komatani, D. Litman, K. Yu, A. Papangelis, L. Cavedon, & M. Nakano (Eds.), *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue* (pp. 224–228). Association for Computational Linguistics. (Cit. on pp. 32, 56).

Matthews, D. (2014). Pragmatic Development in First Language Acquisition, 1–400 (cit. on p. 82).

Matthews, D., Biney, H., & Abbot-Smith, K. (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 (cit. on pp. 20, 33, 64, 83).

McGillion, M., Pine, J. M., Herbert, J. S., & Matthews, D. (2017). A randomised controlled trial to test the effect of promoting caregiver contingent talk on language development in infants from diverse socioeconomic status backgrounds. *Journal of Child Psychology and Psychiatry*, *58*(10), 1122–1131 (cit. on p. 35).

McGrath, S. W., Russin, J., Pavlick, E., & Feiman, R. (2024). How can deep neural networks inform theory in psychological science? *Current directions in psychological science*, *33*(5), 325–333 (cit. on p. 118).

McGuinness, L., Abbot-Smith, K., & Gambi, C. (2023). Autistic and neurotypical children's social impressions of off-topic and delayed responding (cit. on p. 35).

Meena, R., Skantze, G., & Gustafson, J. (2014). Data-driven models for timing feedback responses in a map task dialogue system. *Computer Speech & Language*, *28*(4), 903–922 (cit. on pp. 32, 56).

Mehri, S., Choi, J., D'Haro, L. F., Deriu, J., Eskenazi, M., Gasic, M., Georgila, K., Hakkani-Tur, D., Li, Z., Rieser, V., et al. (2022). Report from the nsf future directions

workshop on automatic evaluation of dialog: Research directions and challenges. *arXiv preprint arXiv:2203.10012* (cit. on pp. 36, 65, 69).

Mehri, S., & Eskenazi, M. (2020). Unsupervised evaluation of interactive dialog with DialoGPT. *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 225–235 (cit. on pp. 36, 64, 65, 70).

Melander, H., & Sahlström, F. (2009). In tow of the blue whale: Learning as interactional changes in topical orientation. *Journal of Pragmatics*, *41*(8), 1519–1537 (cit. on p. 64).

Mezza, S., Cervone, A., Stepanov, E., Tortoreto, G., & Riccardi, G. (2018). ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3539–3551). Association for Computational Linguistics. (Cit. on p. 36).

Miczo, N., Segrin, C., & Allspach, L. E. (2001). Relationship between nonverbal sensitivity, encoding, and relational satisfaction. *Communication Reports*, *14*(1), 39–48 (cit. on pp. 19, 35, 64).

Misiek, T., Favre, B., & Fourtassi, A. (2020). Development of Multi-level Linguistic Alignment in Child-adult Conversations. In E. Chersoni, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 54–58). Association for Computational Linguistics. (Cit. on pp. 64, 96).

Misiek, T., & Fourtassi, A. (2022). Caregivers exaggerate their lexical alignment to young children across several cultures. *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers* (cit. on pp. 46, 64, 83).

Mohapatra, B., Hassan, S., Romary, L., & Cassell, J. (2024). Conversational Grounding: Annotation and Analysis of Grounding Acts and Grounding Units. *LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (cit. on p. 40).

Mohapatra, B., Kapadnis, M. N., Romary, L., & Cassell, J. (2024). Evaluating the effectiveness of large language models in establishing conversational grounding. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 9767–9781). Association for Computational Linguistics. (Cit. on pp. 40, 112).

Mondada, L. (2007). Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse studies*, *9*(2), 194–225 (cit. on pp. 29, 56).

Monroe, W., Hawkins, R. X., Goodman, N. D., & Potts, C. (2017). Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, *5*, 325–338 (cit. on p. 40).

Morency, L. P., de Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, *20*, 70–84 (cit. on p. 48).

Murphy, S. M., Faulkner, D. M., & Farley, L. R. (2014). The behaviour of young children with social communication disorders during dyadic interaction with peers. *Journal of abnormal child psychology, 42*, 277–289 (cit. on p. 19).

Murray, M., Walker, N., Nanavati, A., Alves-Oliveira, P., Filippov, N., Sauppe, A., Mutlu, B., & Cakmak, M. (2022). Learning backchanneling behaviors for a social robot via data augmentation from human-human conversations. *Conference on Robot Learning*, 513–525 (cit. on p. 48).

Musso, M. F., Moyano, S., Rico-Picó, J., Conejero, Á., Ballesteros-Duperón, M. Á., Cascallar, E. C., & Rueda, M. R. (2023). Predicting effortful control at 3 years of age from measures of attention and home environment in infancy: A machine learning approach. *Children, 10* (cit. on p. 23).

Nadig, A., Lee, I., Singh, L., Bosshart, K., & Ozonoff, S. (2010). How does the topic of conversation affect verbal exchange and eye gaze? A comparison between typical development and high-functioning autism. *Neuropsychologia, 48*(9), 2730–2739 (cit. on p. 64).

Narayan-Chen, A., Jayannavar, P., & Hockenmaier, J. (2019). Collaborative dialogue in Minecraft. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5405–5415). Association for Computational Linguistics. (Cit. on p. 40).

Nelson, K. (2007). *Young minds in social worlds: Experience, meaning, and memory*. Harvard University Press. (Cit. on p. 79).

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, Id Rather Do It Myself: Some Effects and Non-Effects of Maternal Speech Style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to Children: Language Input and Acquisition* (pp. 109–149). Cambridge University Press. (Cit. on p. 30).

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 (cit. on pp. 20, 44, 53, 56, 81).

Nikolaus, M., & Fourtassi, A. (2021). Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. *Proceedings of the 25th Conference on Computational Natural Language Learning* (cit. on p. 79).

Nikolaus, M., & Fourtassi, A. (2023). Communicative Feedback in language acquisition. *New Ideas in Psychology, 68*, 100985 (cit. on pp. 38, 79, 97, 105).

Nikolaus, M., Maes, E., Auguste, J., Prévot, L., & Fourtassi, A. (2022). Large-scale study of speech acts' development in early childhood. *Language Development Research, 2*(1) (cit. on pp. 20, 22, 36, 64, 80, 83, 85, 86).

Ninio, A., & Snow, C. (1996). *Pragmatic Development*. Routledge. (Cit. on p. 81).

Ninio, A., Snow, C. E., Pan, B. A., & Rollins, P. R. (1994). Classifying communicative acts in children's interactions. *Journal of Communication Disorders, 27*(2), 157–187 (cit. on pp. 34, 67, 69, 82, 83, 85).

Nota, N., Trujillo, J. P., & Holler, J. (2023). Conversational eyebrow frowns facilitate question identification: An online study using virtual avatars. *Cognitive Science, 47*(12), e13392 (cit. on p. 29).

Onishi, K., Tanaka, H., & Nakamura, S. (2023). Multimodal Voice Activity Prediction: Turn-taking Events Detection in Expert-Novice Conversation. *International Conference on Human-Agent Interaction*, 13–21 (cit. on p. 115).

OpenAI. (2023). Gpt-4 technical report. (Cit. on p. 79).

Paggio, P., & Navarretta, C. (2013). Head movements, facial expressions and feedback in conversations: Empirical evidence from danish multimodal data. *Journal on Multimodal User Interfaces, 7*(1), 29–37 (cit. on p. 48).

Pagmar, D., Abbot-Smith, K., & Matthews, D. (2022). Predictors of children's conversational contingency. *Language Development Research, 2*(1) (cit. on p. 64).

Pang, B., Nijkamp, E., Han, W., Zhou, L., Liu, Y., & Tu, K. (2020). Towards Holistic and Automatic Evaluation of Open-Domain Dialogue Generation. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3619–3629). Association for Computational Linguistics. (Cit. on pp. 36, 65, 70).

Park, H. W., Gelsomini, M., Lee, J. J., & Breazeal, C. (2017). Telling stories to robots: The effect of backchanneling on a child's storytelling. *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, 100–108 (cit. on pp. 23, 44, 48, 53).

Parsons, L., Cordier, R., Munro, N., & Joosten, A. (2019). A Randomized Controlled Trial of a Play-Based, Peer-Mediated Pragmatic Language Intervention for Children With Autism. *Frontiers in Psychology, 10* (cit. on p. 83).

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics, 8*(1), 447–471 (cit. on p. 118).

Paxton, A., & Dale, R. (2013a). Argument disrupts interpersonal synchrony. *Quarterly Journal of Experimental Psychology, 66*(11), 2092–2102 (cit. on p. 19).

Paxton, A., & Dale, R. (2013b). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior research methods, 45*, 329–343 (cit. on p. 19).

Peirolo, M., Xu, Z., & Fourtassi, A. (2024). Development of flexible role-taking in conversations across preschool. *Proceedings of the annual meeting of the cognitive science society, 46* (cit. on p. 35).

Pellegrini, A. D., Symons, F., & Hoch, J. (2012). *Observing children in their natural worlds: A methodological primer*. Psychology Press. (Cit. on p. 65).

Peterson, C. (1990). The who, when and where of early narratives. *Journal of child language, 17*(2), 433–455 (cit. on pp. 44, 53).

Peterson, C., Jesso, B., & McCabe, A. (1999). Encouraging narratives in preschoolers: An intervention study. *Journal of child language, 26*(1), 49–67 (cit. on p. 30).

Piaget, J. (1926). *The language and thought of the child*. Harcourt, Brace. (Cit. on pp. 81, 97).

Piaget, J. (1989). *The Child's Conception of the World*. Rowman & Littlefield. (Cit. on p. 81).

Piaget, J. (2005). *Language and thought of the child: Selected works vol 5*. Routledge. (Cit. on pp. 64, 65, 78).

Pickering, M. J., & Branigan, H. P. (1999). Syntactic priming in language production. *Trends in cognitive sciences*, *3*(4), 136–141 (cit. on p. 20).

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–226 (cit. on pp. 19, 20, 33, 64, 102, 116).

Pickering, M. J., & Garrod, S. (2021). *Understanding Dialogue: Language Use and Social Interaction*. Cambridge University Press. (Cit. on pp. 20, 81).

Pika, S., Wilkinson, R., Kendrick, K. H., & Vernes, S. C. (2018). Taking turns: Bridging the gap between human and animal communication. *Proceedings of the Royal Society B*, *285*(1880), 20180598 (cit. on p. 28).

Place, K. S., & Becker, J. A. (1991). The influence of pragmatic competence on the likeability of grade-school children. *Discourse Processes*, *14*(2), 227–241 (cit. on pp. 19, 35, 64, 83).

Purver, M. (2004). Clarie: The clarification engine. *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, 77–84 (cit. on p. 40).

Purver, M., Hough, J., & Howes, C. (2018). Computational Models of Miscommunication Phenomena. *Topics in Cognitive Science*, *10*(2), 425–451 (cit. on pp. 39, 104).

Putallaz, M., & Gottman, J. M. (1981). An interactional model of children's entry into peer groups. *Child development*, 986–994 (cit. on p. 35).

Qiu, S., Liu, M., Li, H., Zhu, S.-C., & Zheng, Z. (2024). MindDial: Enhancing Conversational Agents with Theory-of-Mind for Common Ground Alignment and Negotiation. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, & K. Komatani (Eds.), *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 746–759). Association for Computational Linguistics. (Cit. on p. 39).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners (cit. on pp. 32, 58, 65, 70).

Rane, S., Nencheva, M. L., Wang, Z., Lew-Williams, C., Russakovsky, O., & Griffiths, T. (2023). Predicting word learning in children from the performance of computer vision systems. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45) (cit. on p. 23).

Reed, J., Hirsh-Pasek, K., & Golinkoff, R. M. (2016). 24 meeting children where they are: Adaptive contingency builds early communication skills. *Communication and learning*, 601 (cit. on p. 35).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992 (cit. on p. 70).

Rieser, V., & Moore, J. D. (2005). Implications for generating clarification requests in task-oriented dialogues. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 239–246 (cit. on p. 40).

Riest, C., Jorschick, A. B., & de Ruiter, J. P. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in psychology*, *6*, 89 (cit. on p. 29).

Robledo, J. P., Hawkins, S., Cornejo, C., Cross, I., Party, D., & Hurtado, E. (2021). Musical improvisation enhances interpersonal coordination in subsequent conversation: Motor and speech evidence. *PloS one*, *16*(4), e0250166 (cit. on p. 20).

Roddy, M., Skantze, G., & Harte, N. (2018a). Investigating speech features for continuous turn-taking prediction using lstms. *Interspeech* (cit. on pp. 32, 56).

Roddy, M., Skantze, G., & Harte, N. (2018b). Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 186–190 (cit. on pp. 32, 115).

Rodríguez, K. J., & Schlangen, D. (2004). Form, intonation and function of clarification requests in german task-oriented spoken dialogues. *Proceedings of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)* (cit. on p. 40).

Rosnay, M. D., Fink, E., Begeer, S., Slaughter, V., & Peterson, C. (2014). Talking theory of mind talk: Young school-aged children's everyday conversation and understanding of mind and emotion. *Journal of Child Language*, *41*(5), 1179–1193 (cit. on pp. 33, 83).

Ruede, R., Müller, M., Stüker, S., & Waibel, A. (2017). Enhancing backchannel prediction using word embeddings. *Interspeech*, 879–883 (cit. on p. 48).

Sacks, H. (1967). Transcribed lectures. *March 9th, University of California, Irvine* (cit. on p. 69).

Sacks, H., Jefferson, G., et al. (1995). *Lectures on conversation* (Vol. 1). Wiley Online Library. (Cit. on p. 33).

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735 (cit. on pp. 20, 28–30, 44, 55, 56, 81).

Sai, A. B., Mohankumar, A. K., Arora, S., & Khapra, M. M. (2020). Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining. *Transactions of the Association for Computational Linguistics*, *8*, 810–827 (cit. on pp. 36, 65).

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, *51*, 1928–1941 (cit. on p. 66).

Sato, R., Higashinaka, R., Tamoto, M., Nakano, M., & Aikawa, K. (2002). Learning decision trees to determine turn-taking by spoken dialogue systems. *INTERSPEECH*, 861–864 (cit. on p. 32).

Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature*, *253*(5489), 265–266 (cit. on p. 35).

Schegloff, E. A. (1986). The routine as achievement. *Human Studies*, *9*(2), 111–151 (cit. on p. 35).

Schegloff, E. A. (1987). Some sources of misunderstanding in talk-in-interaction. *25*(1), 201–218 (cit. on p. 39).

Schegloff, E. A. (1992). Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *American Journal of Sociology*, *97*(5), 1295–1345 (cit. on pp. 39, 104).

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, *29*(1), 1–63 (cit. on pp. 28, 30).

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, *53*(2), 361–382 (cit. on p. 39).

Schegloff, E. A., & Sacks, H. (1973). Opening up Closings. *8*(4), 289–327 (cit. on pp. 33, 64, 69, 82).

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118 (cit. on p. 118).

Selting, M. (1996). Prosody as an activity-type distinctive cue in conversation: The case of so-called'astonished'questions in repair initiation. *STUDIES IN INTERACTIONAL SOCIOLINGUISTICS*, *12*, 231–270 (cit. on pp. 29, 56).

Selting, M. (2000). The construction of units in conversational talk. *Language in society*, *29*(4), 477–517 (cit. on p. 29).

Shaikh, O., Gligoric, K., Khetan, A., Gerstgrasser, M., Yang, D., & Jurafsky, D. (2024). Grounding Gaps in Language Model Generations. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 6279–6296). Association for Computational Linguistics. (Cit. on pp. 39, 40).

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2024). Clever hans or neural theory of mind? stress testing social reasoning in large language models. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2257–2273 (cit. on p. 103).

Skantze, G. (2017). Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks. In K. Jokinen, M. Stede, D. DeVault, & A. Louis (Eds.), *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 220–230). Association for Computational Linguistics. (Cit. on pp. 32, 56).

Skantze, G. (2021). Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, *67*, 101178 (cit. on pp. 28, 31, 32, 44, 47, 53).

Slomkowski, C., & Dunn, J. (1996). Young children's understanding of other people's beliefs and feelings and their connected communication with friends. *Developmental Psychology*, *32*(3), 442–447 (cit. on pp. 33, 64, 83).

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, *4*(1), 1–22 (cit. on pp. 35, 46).

Snow, C. E., Pan, B. A., Imbens-Bailey, A., & Herman, J. (1996). Learning How to Say What One Means: A Longitudinal Study of Children's Speech Act Use*. *Social Development*, *5*(1), 56–84 (cit. on pp. 25, 34, 55, 57, 66, 69, 82–86, 97, 98).

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Citeseer. (Cit. on p. 64).

Sperber, D., & Wilson, D. (2002). Pragmatics, Modularity and Mind-reading. *Mind & Language*, *17*(1-2), 3–23 (cit. on p. 81).

Stalnaker, R. C. (1978, December). Assertion. Brill. (Cit. on pp. 36, 104).

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592 (cit. on p. 28).

Stivers, T., Sidnell, J., & Bergen, C. (2018). Children's responses to questions in peer interaction: A window into the ontogenesis of interactional competence. *Journal of Pragmatics*, *124*, 14–30 (cit. on pp. 31, 35, 81, 97).

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., et al. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285–1295 (cit. on p. 103).

Streeck, J., & Hartge, U. (1992). Previews: Gestures at the transition place. *The contextualization of language*, 135–157 (cit. on pp. 29, 56).

Tice, M., & Henetz, T. (2011). Turn-boundary projection: Looking ahead. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33) (cit. on p. 31).

Tolins, J., Namiranian, N., Akhtar, N., & Fox Tree, J. E. (2017). The role of addressee backchannels and conversational grounding in vicarious word learning in four-year-olds. *First Language*, *37*(6), 648–671 (cit. on p. 30).

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press. (Cit. on p. 79).

Tomasello, M. (2009, June). *Constructing a Language*. Harvard University Press. (Cit. on p. 81).

Tomasello, M. (2010). *Origins of human communication*. MIT press. (Cit. on p. 38).

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child development*, *78*(3), 705–722 (cit. on p. 38).

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. (Cit. on p. 79).

Traum, D. R., & Allen, J. F. (1992). A "speech acts" approach to grounding in conversation. *ICSLP* (cit. on p. 40).

Trinh, T. H., & Le, Q. V. (2019, September). A Simple Method for Commonsense Reasoning. (Cit. on p. 71).

Udagawa, T., & Aizawa, A. (2019). A natural language corpus of common grounding under continuous and partially-observable context. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 7120–7127 (cit. on p. 40).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. u., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30* (cit. on p. 36).

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511 (cit. on p. 96).

Vygotsky, L. S. (2012, July). *Thought and Language, revised and expanded edition*. MIT Press. (Cit. on p. 81).

Wang, Z., Devine, R. T., Wong, K. K., & Hughes, C. (2016). Theory of mind and executive function during middle childhood across cultures. *Journal of experimental child psychology*, *149*, 6–22 (cit. on p. 20).

Ward, N. (2004). Pragmatic functions of prosodic features in non-lexical utterances. *Speech prosody*, *4*, 325–328 (cit. on p. 30).

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, *32*(8), 1177–1207 (cit. on pp. 30, 44).

Ward, N. G. (2019). *Prosodic patterns in english conversation*. Cambridge University Press. (Cit. on pp. 29, 56).

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A Social Feedback Loop for Speech Development and Its Reduction in Autism. *Psychological Science*, *25*(7), 1314–1324 (cit. on p. 96).

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, & R. Cotterell (Eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 1–34). Association for Computational Linguistics. (Cit. on p. 96).

Wolters, N., Knoors, H., Cillessen, A. H., & Verhoeven, L. (2014). Behavioral, personality, and communicative predictors of acceptance and popularity in early adolescence. *The Journal of Early Adolescence*, *34*(5), 585–605 (cit. on p. 35).

Yeh, Y.-T., Eskenazi, M., & Mehri, S. (2021). A Comprehensive Assessment of Dialog Evaluation Metrics. *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, 15–33 (cit. on p. 65).

Yi, S., Goel, R., Khatri, C., Cervone, A., Chung, T., Hedayatnia, B., Venkatesh, A., Gabriel, R., & Hakkani-Tur, D. (2019). Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators. *Proceedings of the 12th International Conference on Natural Language Generation*, 65–75 (cit. on pp. 36, 65, 70).

Yngve, V. H. (1970). On getting a word in edgewise. *Chicago Linguistics Society, 6th Meeting, 1970*, 567–578 (cit. on pp. 30, 44, 116).

Yu, Y., Bonawitz, E., & Shafto, P. (2019). Pedagogical Questions in Parent-Child Conversations. *Child Development*, *90*(1), 147–161 (cit. on p. 81).

Zellers, M., Gorisch, J., House, D., & Peters, B. (2019). Hand gestures and pitch contours and their distribution at possible speaker change locations: A first investigation. *Gesture and Speech in Interaction-6th edition (GESPIN 2019), 11-13 September, University of Paderborn* (cit. on pp. 29, 56).

Zhao, C., Serratrice, L., Lieven, E., Steele, C., Malik, N., An, Y., Hayden, E., Neumegen, J., & Cameron-Faulkner, T. (2024). Communicative function in child directed speech: A cross-cultural analysis. *First Language*, 01427237241259065 (cit. on p. 82).

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. *2015 IEEE International Conference on Computer Vision (ICCV)*, 19–27 (cit. on p. 71).

ANNEXES

A. Appendix A

A.1. Turn-shift Plots

Figures A.1, A.2, and A.3 display the turn-taking latencies between the two sets of dyads in the ChiCo corpus (Bodur et al., 2021).

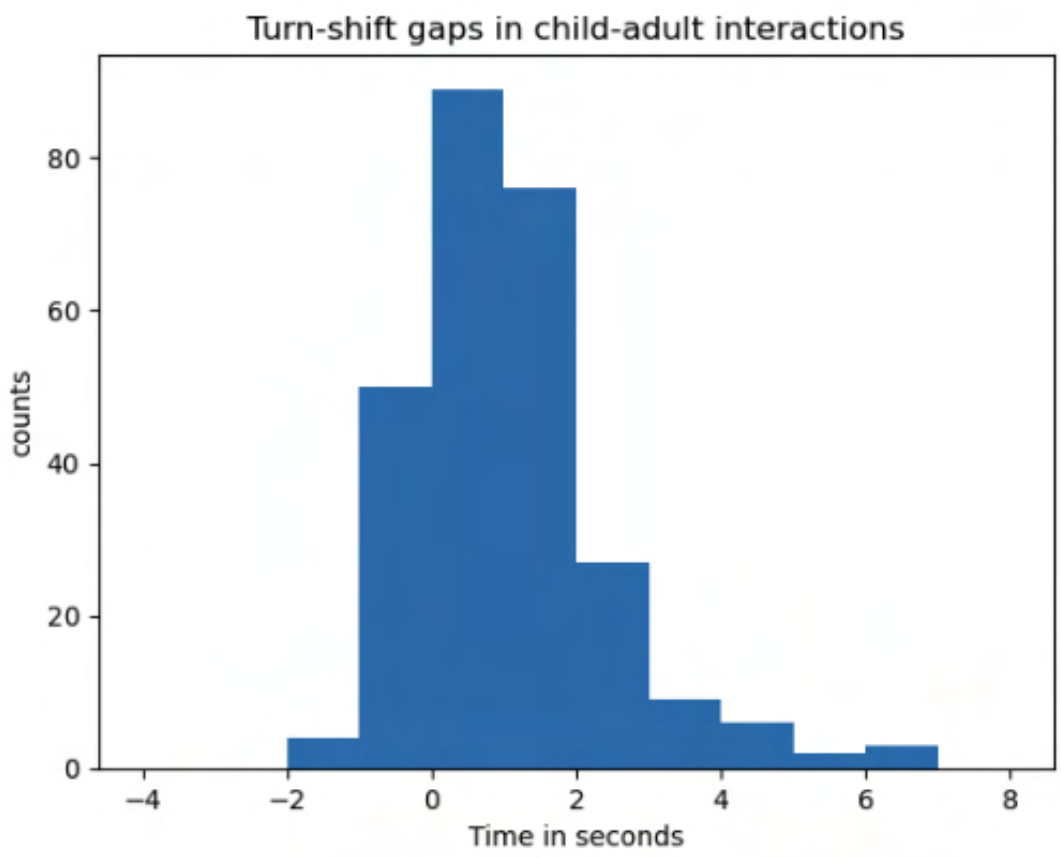

Figure A.1.: Turn-taking latency in child-caregiver interactions for the ChiCo corpus (Bodur et al., 2021). Negative latencies represent overlaps and positive latencies gaps.

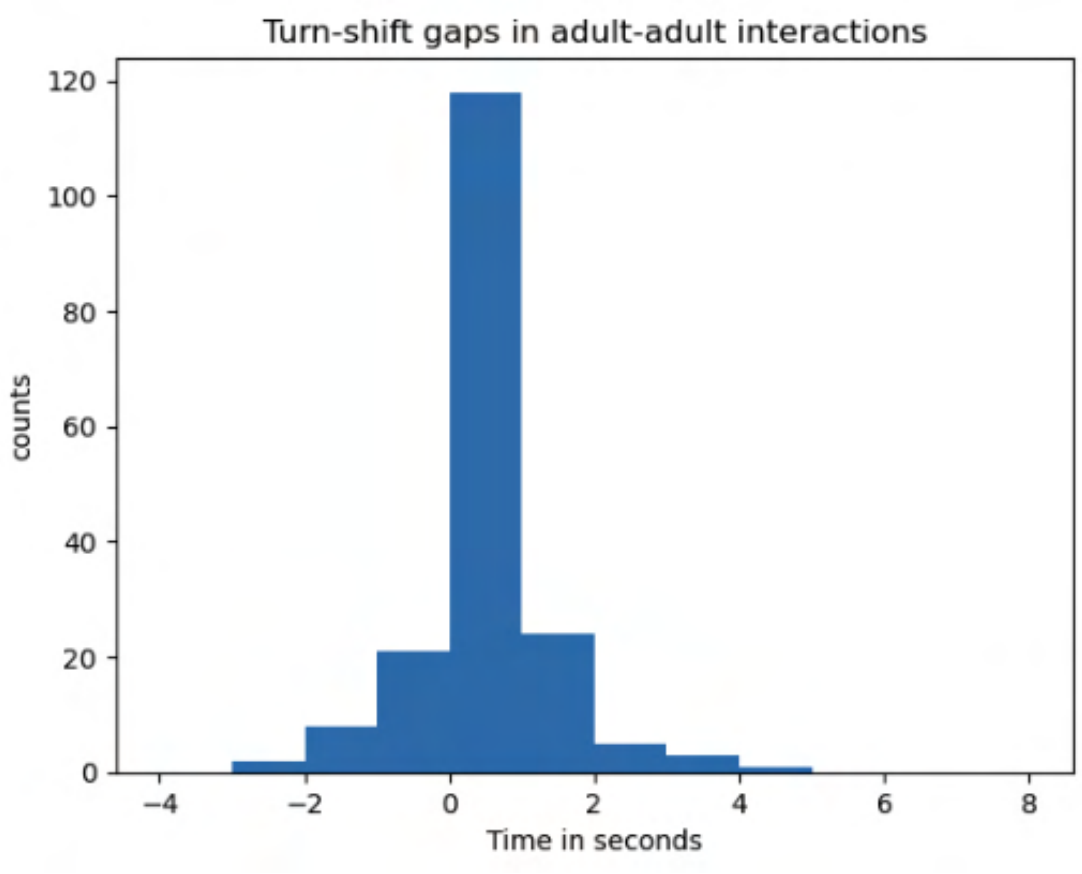

Figure A.2.: Turn-taking latency in caregiver-adult interactions for the ChiCo corpus (Bodur et al., 2021). Negative latencies represent overlaps and positive latencies gaps.

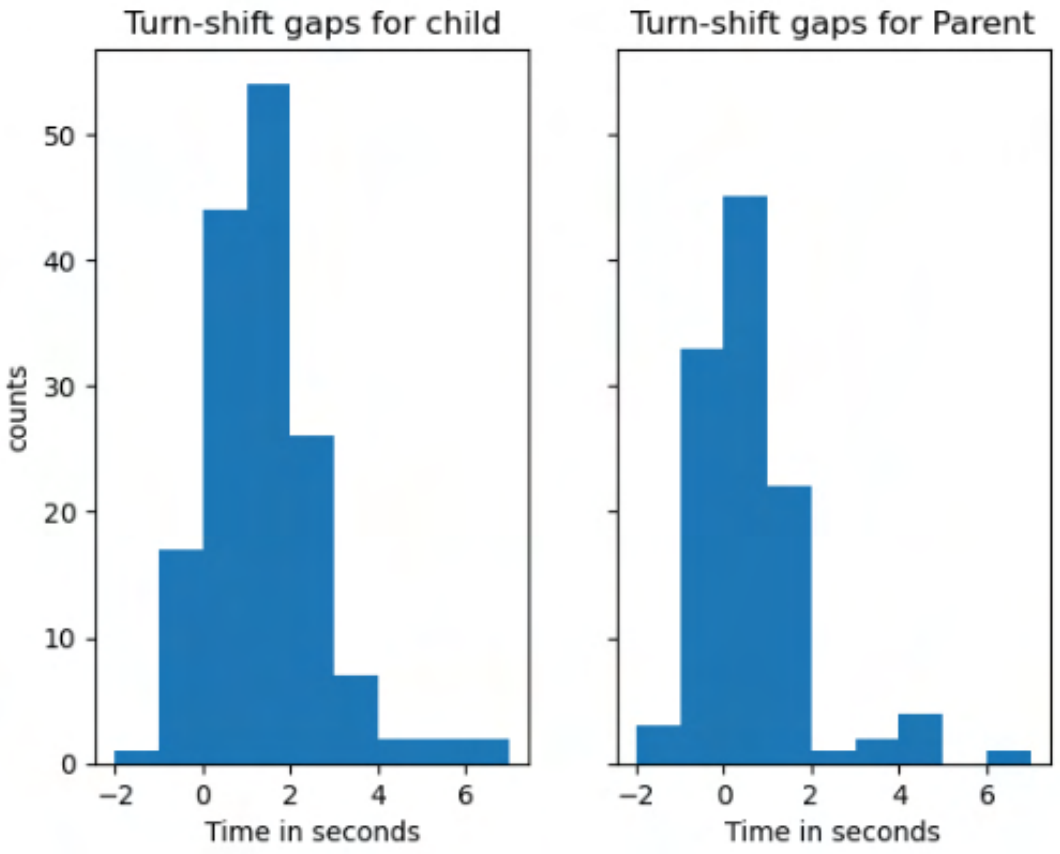

Figure A.3.: Turn-taking latency in child-caregiver interactions split by each interlocutor for the ChiCo corpus (Bodur et al., 2021). The plot on the left shows the latencies for the child taking the turn after the caregiver and the plot on the right shows the latencies for caregiver taking the turn after the child. Negative latencies represent overlaps and positive latencies gaps.

A.2. Additional Experiments

A.2.1. Experiment 4a: MC vs. BC vs. no signal

In this set of experiments, we predict either the listener's turn, BC or when they are not taking any action based on the speaker's cues. Here the model was trained and tested on an equal number of samples for each outcome. As depicted in the graph below, the results show an above chance performance (chance here is 0.33) for all groups across all modalities.

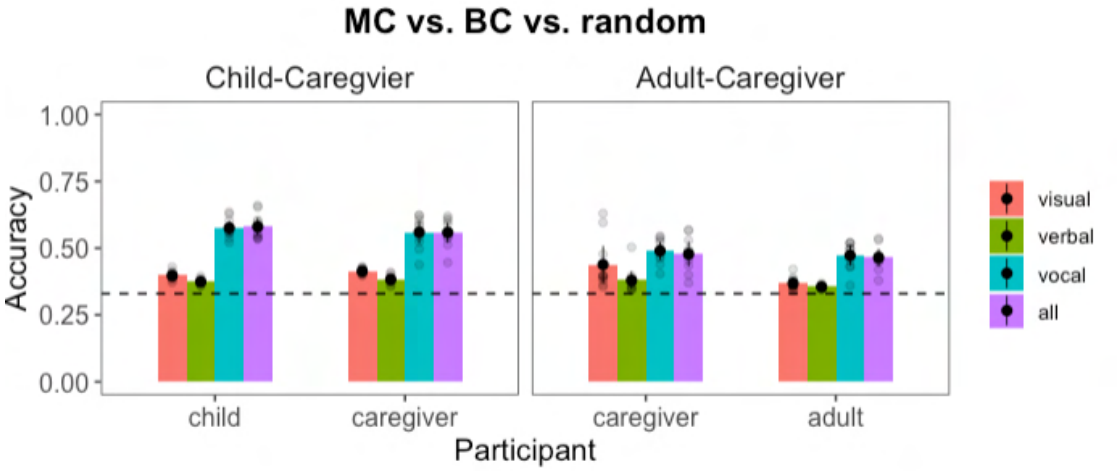

Figure A.4.: Accuracy scores of the MC vs. BC vs. no signal predicting models (Experiment 4a) both in the child-caregiver condition (left) and the adult-caregiver condition (right). In each case, we show the results of the modality-specific models (i.e., the models using predictors from a single modality) 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.

A.2.2. Experiment 4b: Cross-dyad setting

We also performed a set of experiments using the same three label prediction task as described in Experiment 4a, where we took models trained on adult-caregiver conversations and used them to predict the child and caregiver's use of the appropriate channel in a child-caregiver conversation and vice versa. For this purpose, we saved the models that had the highest accuracy in the LOOCV setup for a particular dyad (either adult-caregiver or child-caregiver) and then used these models to predict MC, BC or no signal for all four groups of participants by feeding the model the cues from the corresponding other interlocutor. For example, if we consider the Adult model, it has been trained to predict if the Adult is going to take the MC, do a BC or nothing based on the caregiver's cues and the model with the best accuracy is saved. This saved model is then used to predict the Child's coordination by feeding it cues from the Caregiver, the Caregiver's coordination by feeding it cues from the Child and so on. The table below contains the results for this experiment. Note that the accuracy values here are for the entire data (across all 10 participants) for each group and not the accuracy values in a LOOCV setting. All the models for this experiment used the hyperparameters which were tuned with respect to the Adult.

Interlocutor	Adult model	Child model	Caregiver/C model
Child	0.549	0.647	0.638
Caregiver/C	0.494	0.589	0.609
Adult	0.599	0.589	0.603
Caregiver/A	0.583	0.555	0.592

Table A.1.: The accuracy scores for Experiment 4b. For the Adult-Caregiver/A dyad we train only one model as the both models would behave the same in theory.

The results are all above chance level (which is 0.333 in this case) indicating that there might be some overlap in the inviting cues used by children and adults.

A.3. Ablation study

We also trained models on just the speaker's gaze and speaker's F0 semitone for the same experimental setups as experiments 1, 2, and 3. We chose the gaze and F0 semitone since most studies in the literature (see Chapter 1) consider them as important in the visual and vocal modalities respectively. The results for experiments 1, 2, and 3 can be seen in Figures A.5, A.6, and A.7 respectively.

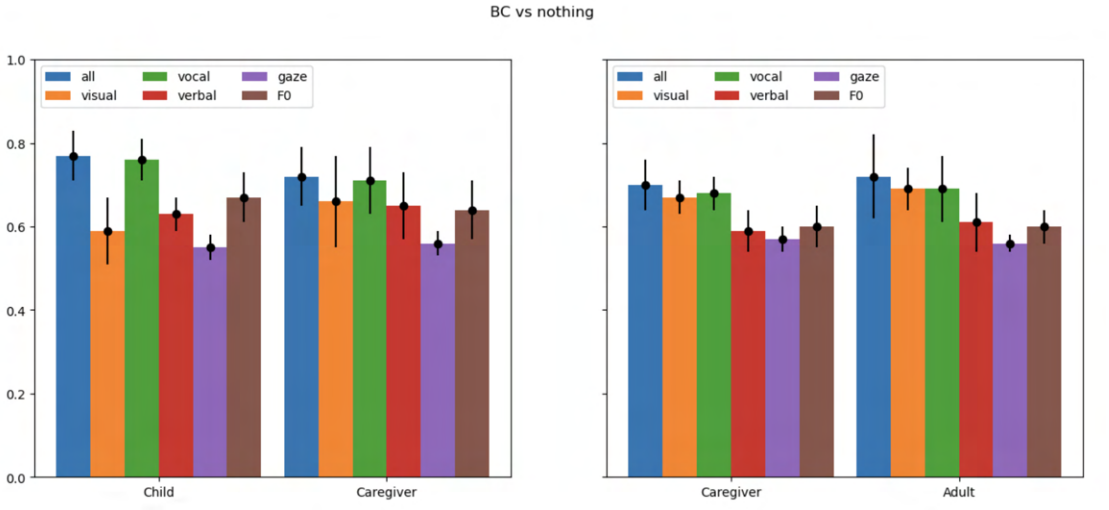

Figure A.5.: Accuracy scores of the BC vs. no signal predicting models both in the child-caregiver condition (left) and the adult-caregiver condition (right). In each case, we show the results of the modality-specific models (i.e., the models using predictors from a single modality) and the full model (where cues from all modalities are combined) in addition to just using the gaze of the speaker and the F0 semitone. We show the mean accuracy scores obtained with different training/testing configurations according to the LOOCV cross-validation and the error bars show the standard deviation over these scores.

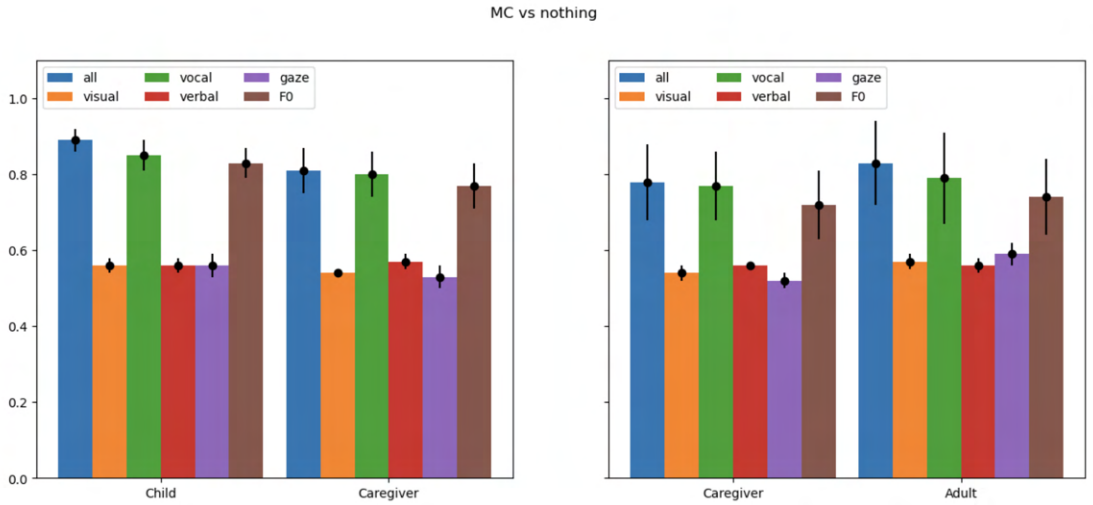

Figure A.6.: Accuracy scores of the MC vs. no signal predicting models both in the child-caregiver condition (left) and the adult-caregiver condition (right). In each case, we show the results of the modality-specific models (i.e., the models using predictors from a single modality) and the full model (where cues from all modalities are combined) in addition to just using the gaze of the speaker and the F0 semitone. We show the mean accuracy scores obtained with different training/testing configurations according to the LOOCV cross-validation and the error bars show the standard deviation over these scores.

Hyperparameter	Tuned w.r.t. child	Tuned w.r.t. adult
train batch size	16	128
test batch size	2	16
number of epochs	23	45
learning rate	2.1e-05	1e-05
weight decay	4.3e-05	1.6e-05
LSTM hidden dims	90	100
LSTM layers	1	2
input dropout	0.53	0.26
output dropout	0.1	0.36

Table A.2.: Hyperparameters for experiment 1)

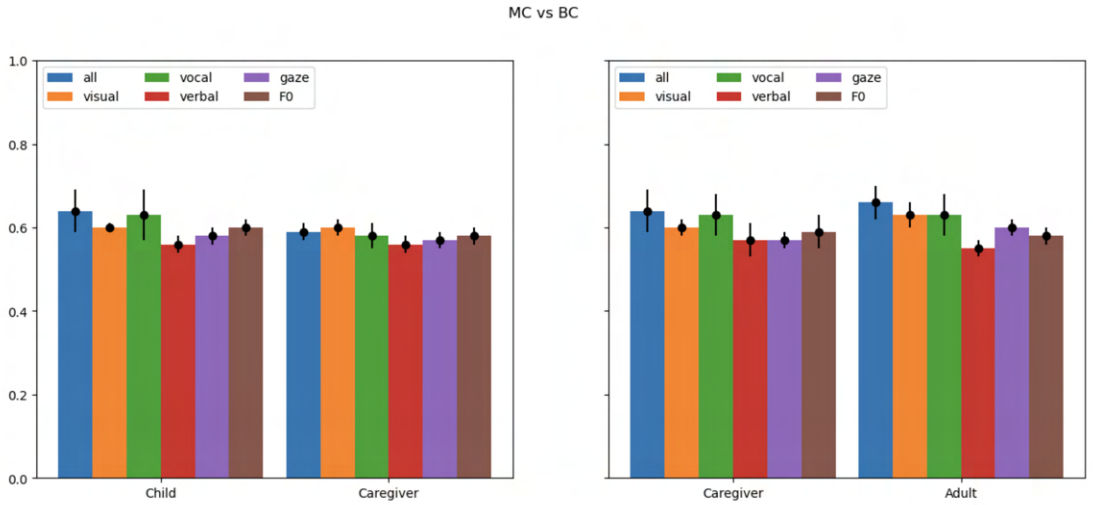

Figure A.7.: Accuracy scores of the MC vs. BC predicting models both in the child-caregiver condition (left) and the adult-caregiver condition (right). In each case, we show the results of the modality-specific models (i.e., the models using predictors from a single modality) and the full model (where cues from all modalities are combined) in addition to just using the gaze of the speaker and the F0 semitone. We show the mean accuracy scores obtained with different training/testing configurations according to the LOOCV cross-validation and the error bars show the standard deviation over these scores.

A.4. Model Hyperparameters

For each experiment, we tuned a set of hyperparameters for the model using Ray Tune. We tuned the set of hyperparameters once using the children's data and once using the adult's data. The following tables of hyperparameters gave us the best results.

Hyperparameter	Tuned w.r.t. child	Tuned w.r.t. adult
train batch size	128	64
test batch size	2	8
number of epochs	51	54
learning rate	5.3e-05	5.1e-05
weight decay	1.6e-05	2.9e-05
LSTM hidden dims	80	110
LSTM layers	3	1
input dropout	0.21	0.73
output dropout	0.22	0.73

Table A.3.: Hyperparameters for experiment 2)

Hyperparameter	Tuned w.r.t. child	Tuned w.r.t. adult
train batch size	32	128
test batch size	8	8
number of epochs	53	49
learning rate	6.8e-05	0.0009
weight decay	0.0002	0.005
LSTM hidden dims	190	60
LSTM layers	1	2
input dropout	0.15	0.7
output dropout	0.49	0.31

Table A.4.: Hyperparameters for experiment 3)

Hyperparameter	Tuned w.r.t. child	Tuned w.r.t. adult
train batch size	64	128
test batch size	4	16
number of epochs	59	33
learning rate	0.0018	0.0013
weight decay	0.0009	0.029
LSTM hidden dims	100	80
LSTM layers	3	1
input dropout	0.11	0.27
output dropout	0.81	0.65

Table A.5.: Hyperparameters for experiment 4)

B. Appendix B

B.1. Annotation Scheme

We develop an annotation scheme to annotate contingency for the New England corpus which is one of several corpora of child-caregiver interactions available in the CHILDES databank. We annotate utterances only at the turn switch level between the child and the caregiver i.e., when the role of the speaker in the conversation changes from child to caregiver or vice versa. We don't consider a fixed past context length in terms of the number of utterances that we consider while annotating a target utterance.

We consider topic shifts on a case by case basis. Generally, we consider minor topic shifts to be ambiguous in nature. For instance, while reading the animal picture book, if the caregiver keeps asking "what is this?" we annotate it as ambiguous. Any smooth topic shifts which fall in line with something from the recent past context is annotated as contingent. Consider the below example:

Caregiver: *what is it?*
Caregiver: *a book!*
Child: *yeah.*
Caregiver: *oh you want me to read it?*

In the above example, we consider the turn switch from child to caregiver as contingent since the topic shifts smoothly from the book to reading the book. An example of a non-contingent topic shift is shown below:

Caregiver: *what are you going to do now?*
Child: *going to do.*
Caregiver: *what's that?*
Caregiver: *is that a block?*

In the above example, the turn switch from the child to the caregiver is non-contingent since it is an abrupt change of topic.

If a turn switch can be considered contingent on the assumption that the person is pointing/gesturing to something then we mark it as ambiguous (since we rely only on the transcripts and not visual data for our annotations). If an utterance is a repetition of the previous utterance then we consider it as contingent as it can be a confirmation or acknowledgment of the previous utterance. However, if the interlocutor/s keep

repeating an utterance redundantly then we annotate it as ambiguous since we cannot be sure of the intention behind this repetition. Consider the below example:

Caregiver: *this is a no-no.*
Child: *no-no.*
Caregiver: *no-no.*
Child: *no-no.*
Caregiver: *no-no.*

In the above example, we consider the first repetition done by the child to be contingent but all the other repetitions we mark as ambiguous since they are redundant.

We consider all clarification requests to be contingent. If there are two back to back utterances from the same interlocutor where the second utterance can be considered as a continuation of the first utterance and the turn switch is contingent with the second utterance then we mark the turn switch as contingent. Consider the below example:

Caregiver: *what's this?*
Caregiver: *what's in this box?*
Child: *oh.*
Child: *oh this.*

In the above example, we annotate the turn switch as contingent since the second utterance by the child indicates that the child has an idea of what could be in the box.

We annotate any random or off topic responses to questions as non-contingent. If the response to a question is another question then we mark it as non-contingent unless the question in the response is a clarification request. If we are unsure whether the response question is a clarification request then we mark it as ambiguous.

We consider backchannels (short verbal utterances like “mhm", “mm", “uh-huh", “oh", etc.) on a case by case basis. We never treat a backchannel as non-contingent. If there is any doubt concerning the contingent nature of a backchannel response, then we annotate it as ambiguous. Consider the example below:

Caregiver: *is that a cow?*
Child: *mhm.*
Caregiver: *mm.*
Caregiver: *and a baby donkey on the farm.*

In the above example, we consider the child turn switch as contingent since the child is responding to the question. However, consider the following example:

Caregiver: *what's that?*
Caregiver: *wanna sit down and read the book?*

Child: *oh.*
Caregiver: *come here.*

This was marked as ambiguous because one cannot be sure – based on the transcript alone – what the child is trying to express.

B.2. Classifier Results Segregated by Age of Child

We also trained separate models for data segregated by the age of the child. Table B.1 displays the results for the models trained on data from the 20 months old children. The classifier used in the feature-based methods and for the baselines was the logistic regression classifier. In instances where we compute the perplexity with the GPT-2 model, we then further fit a logistic regression classifier to predict the contingency label from the perplexity values. As you can see in the table, the feature-based models perform quite poorly while the best model is the supervised DeBERTaV3 model for both children and adults.

Table B.2 displays the results for the models trained on data from the 32 months old children. Once again, the feature-based models perform quite poorly while the best model is the supervised DeBERTaV3 model for both children and adults.

Classifier	Child		Adult	
	F1 score	MCC score	F1 score	MCC score
Majority classifier	0.35 ± 0.07	0.00 ± 0.00	0.66 ± 0.06	0.00 ± 0.00
Chance classifier	0.34 ± 0.05	-0.01 ± 0.07	0.40 ± 0.04	-0.01 ± 0.03
Speech acts (SA)	0.35 ± 0.07	0.03 ± 0.04	0.66 ± 0.06	0.00 ± 0.00
Noun phrase reps. (NP)	0.44 ± 0.05	0.12 ± 0.09	0.25 ± 0.25	0.00 ± 0.03
Cosine similarity (CS)	0.28 ± 0.04	0.12 ± 0.05	0.55 ± 0.05	0.17 ± 0.03
NP + CS	0.37 ± 0.07	0.07 ± 0.03	0.43 ± 0.09	0.10 ± 0.05
GPT-2 (no fine-tuning)	0.06 ± 0.04	0.00 ± 0.00	0.31 ± 0.26	-0.01 ± 0.01
GPT-2 (self-supervised, PPL)	0.46 ± 0.08	0.14 ± 0.09	0.55 ± 0.15	-0.03 ± 0.04
PPL + NP	0.49 ± 0.05	0.18 ± 0.09	0.20 ± 0.20	0.00 ± 0.02
PPL + CS	0.51 ± 0.03	0.20 ± 0.05	0.53 ± 0.05	0.14 ± 0.05
PPL + NP + CS	0.52 ± 0.05	0.21 ± 0.07	0.44 ± 0.10	0.09 ± 0.06
DeBERTaV3 (default)	0.19 ± 0.14	0.01 ± 0.01	0.26 ± 0.29	0.03 ± 0.04
DeBERTaV3 (supervised)	**0.64 ± 0.07**	**0.41 ± 0.08**	**0.33 ± 0.05**	**0.73 ± 0.06**

Table B.1.: The mean weighted F1 scores and MCC scores along with the standard deviation across a 5 fold cross-validation for children aged 20 months and for adults conversing with 20 months old children. The results for the feature based models are with a logistic regression classifier.

Classifier	Child		Adult	
	F1 score	MCC score	F1 score	MCC score
Majority classifier	0.58 ± 0.06	0.00 ± 0.00	0.71 ± 0.02	0.00 ± 0.00
Chance classifier	0.38 ± 0.03	-0.02 ± 0.04	0.41 ± 0.02	0.01 ± 0.02
Speech acts (SA)	0.58 ± 0.06	0.00 ± 0.00	0.71 ± 0.02	-0.01 ± 0.01
Noun phrase reps. (NP)	0.57 ± 0.05	0.05 ± 0.03	0.41 ± 0.05	0.05 ± 0.03
Cosine similarity (CS)	0.51 ± 0.08	0.05 ± 0.09	0.55 ± 0.02	0.14 ± 0.04
NP + CS	0.55 ± 0.08	0.04 ± 0.08	0.50 ± 0.05	0.10 ± 0.03
GPT-2 (no fine-tuning)	0.02 ± 0.01	0.00 ± 0.00	0.28 ± 0.32	0.01 ± 0.01
GPT-2 (self-supervised, PPL)	0.35 ± 0.21	0.08 ± 0.08	0.45 ± 0.19	-0.03 ± 0.04
PPL + NP	0.58 ± 0.05	0.08 ± 0.04	0.42 ± 0.05	0.04 ± 0.04
PPL + CS	0.43 ± 0.15	0.01 ± 0.08	0.55 ± 0.03	0.14 ± 0.05
PPL + NP + CS	0.56 ± 0.07	0.07 ± 0.06	0.51 ± 0.05	0.10 ± 0.04
DeBERTaV3 (default)	0.26 ± 0.25	0.02 ± 0.02	0.31 ± 0.34	0.01 ± 0.02
DeBERTaV3 (supervised)	**0.59 ± 0.14**	**0.24 ± 0.14**	**0.73 ± 0.04**	**0.24 ± 0.11**

Table B.2.: The mean weighted F1 scores and MCC scores along with the standard deviation across a 5 fold cross-validation for children aged 32 months and for adults conversing with 32 months old children. The results for the feature based models are with a logistic regression classifier.

C. Appendix C

C.1. Corpora in English-language CHILDES

Table C.1.: A list of all the corpora in the English CHILDES dataset.

Corpus	No. of transcripts	Corpus	No. of transcripts
Bates	100	McCune	29
Belfast	15	McMillan	3
Bernstein	13	Morisset	138
Bliss	1	Nadig	15
Bloom	24	Nelson	40
Braunwald	438	NewEngland	78
Brown	51	NewmanRatner	121
Clark	21	Peters	45
Conti1	18	Post	29
Demetras1	12	Providence	196
Demetras2	31	Sachs	67
EllisWeismer	116	Snow	14
Feldman	6	Suppes	27
Forrester	15	TD	31
Gelman	26	Tardif	19
Gleason	14	Thomas	168
Higginson	6	Tommerdahl	10
Howe	16	Valian	42
Kuczaj	30	VanHouten	55
Lara	58	Warren	7
MPI-EVA-Manchester	231	Weist	56
MacWhinney	19	Wells	126

D. Appendix D

D.1. Prompt Template

The prompt templates for the English and French transcripts are shown in Fig D.1 and D.2 respectively.

```
Two people are playing a word guessing game where player 1 picks a word and
 player 2 doesn't know this word. Player 2 needs to ask questions to player 1
 to guess the word correctly. Given the dialog history in terms of the turns taken
 by player 1 and player 2 and the word picked by player 1, you need to decide whether
 the next question asked or statement made by player 2 or the object mentioned by player 2 is valid or
 not based on the dialog history until that point. You need to give a boolean binary response (True or False) whether
 the question is valid or not in JSON format. Use the following template: {valid: ""}.

 Here are some examples to help you out.

 Example 1:      Word picked by player 1: A balloon.
 Dialog history:  player 2 turn: Is it a living being? player 1 turn: No. player 2 turn: Is it an object? player 1 turn: Yes.
 Next question:  Can you play with it?
 {valid: True}

 Example 2:      Word picked by player 1: A cat.
 Dialog history:  player 2 turn: Is it a living being? player 1 turn: Yes. player 2 turn: Can it be a pet? player 1 turn: Yes.
 Next question:  a cat?
 {valid: True}

 Example 3:      Word picked by player 1: A car.
 Dialog history:  player 2 turn: Is it a living being? player 1 turn: No.
 Next question:  is it an insect?
 {valid: False}

 End of examples.

 Word picked by player 1: <TARGET_WORD>
 Dialog history: <DIALOG_HISTORY>
 Next question: <QUESTION>
```

Figure D.1.: Prompt template with English examples and transcript.

```
Two people are playing a word guessing game in the French language where player 1 picks a word and
 player 2 doesn't know this word. Player 2 needs to ask questions to player 1
 to guess the word correctly. Given the dialog history in terms of the turns taken
 by player 1 and player 2 and the word picked by player 1, you need to decide whether
 the next question asked or statement made by player 2 or the object mentioned by player 2 is valid or
 not based on the dialog history until that point. You need to give a boolean binary response (True or False) whether
 the question is valid or not in JSON format. Use the following template: {valid: ""}.

 Here are some examples to help you out.

 Example 1:      Word picked by player 1: Un ballon.
 Dialog history: player 2 turn: Est-ce que ça un être vivant? player 1 turn: Non.
                 player 2 turn: Est-ce que ça un objet? player 1 turn: Oui.
 Next question:  Peux-tu jouer avec ça?
 {valid: True}

 Example 2:      Word picked by player 1: Un chat.
 Dialog history: player 2 turn: Est-ce que ça un être vivant? player 1 turn: Oui.
                 player 2 turn: Est-ce que ça peut être un animal de compagnie? player 1 turn: Oui.
 Next question:  un chat?
 {valid: True}

 Example 3:      Word picked by player 1: Une voiture.
 Dialog history: player 2 turn: Est-ce que ça un être vivant? player 1 turn: Non.
 Next question:  Est-ce que ça un insecte?
 {valid: False}

 End of examples.

 Word picked by player 1: <TARGET_WORD>
 Dialog history: <DIALOG_HISTORY>
 Next question: <QUESTION>
```

Figure D.2.: Prompt template with French examples and transcript.