



Gaze Dynamics in Middle Childhood

Dhia-Elhak Goumri

► To cite this version:

Dhia-Elhak Goumri. Gaze Dynamics in Middle Childhood. Computer Science [cs]. Aix-Marseille Université, 2026. English. ⟨NNT : 2026AIXM0063⟩. ⟨tel-05604310⟩

HAL Id: tel-05604310

<https://theses.hal.science/tel-05604310v1>

Submitted on 27 Apr 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

THÈSE DE DOCTORAT

Soutenue à AMU — Aix-Marseille Université
le 16 Mars 2026 par

Dhia-Elhak GOUMRI

Gaze Dynamics in Middle Childhood :
An Ecologically Grounded Account

Discipline
Informatique

École doctorale
ED 184 Mathématiques et Informatique

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

Composition du jury

Bahia GUELLAI
Professeur
Université Toulouse - Jean Jaurès
Rapporteuse

Mohamed CHETOUANI
Professeur
Sorbonne Université
Rapporteur

Mathilde FORT
Maîtresse de Conférences HDR
Université Grenoble Alpes
Examinatrice

Louise GOUPIL
Chargée de Recherche CNRS
Université Grenoble Alpes
Examinatrice

Philippe BLACHE
Directeur de Recherche CNRS
Aix-Marseille Université
Président du jury

Leonor BECERRA-BONACHE
Maîtresse de Conférences HDR
Aix-Marseille Université
Directrice de thèse

Invité

Abdellah FOURTASSI
Maître de Conférences HDR
Co-directeur de thèse

Affidavit

Je soussigné, Dhia Elhak GUMRI, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Leonor BECERRA, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte AMU relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 13 Janvier 2026



Cette œuvre est mise à disposition selon les termes de la [Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Liste de publications et participation aux conférences

Liste des publications et/ou brevets réalisées dans le cadre du projet de thèse :

1. **Automatic Detection of Gaze and Smile in Children's Video Calls.**
Goumri, D.-E., Jansson, T., Becerra-Bonache, L., & Fourtassi, A.
Companion Proceedings of the 25th International Conference on Multimodal Interaction (ICMI 2023), 2023.
2. **CHICA : A developmental corpus of child–caregiver face-to-face vs. video call conversations in middle childhood.**
Goumri, D., Agrawal, A., Nikolaus, M., Vu, H., Bodur, K., Emmar, E., Elias, A. C., Mazzocconi, C., Gupta, S., Prévot, L., Favre, B., Becerra, L., & Fourtassi, A.
Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC–COLING 2024), 2024.
3. **CHICA : un corpus de conversations face-à-face vs. visioconférence entre enfants et parents.**
Goumri, D., Agrawal, A., Nikolaus, M., Vu, H., Bodur, K., Emmar, E., Elias, A. C., Mazzocconi, C., Gupta, S., Prévot, L., Favre, B., Becerra, L., & Fourtassi, A.
Actes des 35^e Journées d'Études sur la Parole (JEP 2024), 2024.
4. **Child–Caregiver Gaze Dynamics in Naturalistic Face-to-Face Conversations.**
Goumri, D.-E., Becerra-Bonache, L., & Fourtassi, A.
Proceedings of the 46th Annual Meeting of the Cognitive Science Society (CogSci 2024), 2024.
5. **Children's Gaze in Face-to-Face Conversations with Caregivers : What Develops in Middle Childhood?**
Goumri, D.-E., Becerra-Bonache, L., & Fourtassi, A.
submitted to *Cognitive Science Journal*.

Participation aux conférences et écoles d'été au cours de la période de thèse :

1. Institute of Language, Communication and the Brain (ILCB) PhD program 2022-2025
2. Institute of Language, Communication and the Brain (ILCB) Summer School 2022
3. International Conference on Multimodal Interaction (ICMI 2023)
4. The 15th Lisbon Machine Learning School (LxMLS) 2023
5. Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC–COLING 2024)
6. Annual Meeting of the Cognitive Science Society (CogSci 2024)
7. Journées d'Études sur la Parole (JEP 2024)

Abstract and keywords

To be considered a competent conversationalist, one must acquire the ability to coordinate a repertoire of linguistic, vocal, and visual resources as part of a jointly managed activity. While early childhood and adult interaction have been extensively studied, the developmental period of middle childhood (6 to 12) remains comparatively under-characterized, particularly in ecologically valid settings. This thesis develops an ecologically grounded account of gaze behavior and dynamics in conversation during middle childhood.

We introduce the CHICA (Child Interpersonal Communication Analysis) corpus, a novel multimodal dataset of French-speaking child-caregiver dyads recorded in their homes, documenting gaze behavior in face-to-face interaction using mobile eye-tracking. The corpus was methodologically developed to directly challenge laboratory-based approaches in which experimental settings themselves reshape the organization of talk and gaze. In practice, the research overcomes the "annotation bottleneck" in ecological data by adopting a hybrid approach: combining mobile eye-tracking and machine learning, including automated face detection, speech segmentation, and gaze estimation, with targeted manual correction.

The thesis establishes baseline patterns of gaze behavior in middle childhood by first examining its overall prevalence. The analyses show that children aged 6-12 display a robust preference for directing their gaze toward the interlocutor's face. Beyond overall prevalence: (1) children modulate gaze as a function of their conversational role, exhibiting higher face-directed gaze while listening than while speaking, a pattern that mirrors that of adults. (2) gaze is organized with respect to turn-taking. Children are more likely to avert their gaze at speech onset when taking the turn, and to direct their gaze back toward the interlocutor at speech offset when yielding the turn. Taken together, these patterns indicate that gaze coordination in middle childhood reflects refinement of an already established competence, rather than the acquisition of a new skill.

Finally, we introduce a turn-taking prediction task to assess gaze contribution to floor management. Three approaches are introduced and evaluated: (1) a feature-based model operating on handcrafted acoustic and gaze features, (2) a sequence-based model that explicitly captures the temporal structure of speech and gaze, and (3) a representation-learning approach relying on learned acoustic embeddings. Across all three approaches, the inclusion of gaze information improves the prediction, showing that gaze contributes systematic information at turn transitions.

This thesis argues that the developmental work of middle childhood is not about learning where to look, but rather the qualitative fine-tuning of how to use gaze fluently in talk. These findings provide essential normative baselines for clinical applications

and the development of socially intelligent child-robot interfaces.

Keywords: middle childhood, gaze, conversational competence, machine learning, eye-tracking, ecological data, turn-taking, child development

Résumé et mots clés

Pour être considéré comme un interlocuteur compétent, il faut acquérir la capacité de coordonner, en temps réel, un ensemble de ressources linguistiques, vocales et visuelles au cours d'une activité conjointement organisée. Alors que l'interaction en petite enfance et à l'âge adulte a été largement étudiée, la période de l'enfance moyenne (6 à 12 ans) demeure relativement sous-caractérisée, en particulier dans des contextes à forte validité écologique. Cette thèse développe ainsi un compte rendu écologiquement fondé des comportements et des dynamiques de regard en conversation durant l'enfance moyenne.

Nous introduisons le corpus CHICA (Child Interpersonal Communication Analysis), un nouveau jeu de données multimodal de dyades enfant-parent francophones enregistrées à domicile, documentant le regard en interaction face à face à l'aide de lunettes avec un système d'eye-tracking. Le corpus a été conçu méthodologiquement pour proposer une alternative aux approches de laboratoire, dans lesquelles le cadre expérimental lui-même tend à reconfigurer l'organisation de la parole et du regard. Concrètement, cette thèse traite le "bottleneck" de l'annotation dans les données écologiques en adoptant une approche hybride : elle combine l'oculométrie mobile et des outils d'apprentissage automatique: incluant la détection automatique de visages, la segmentation de la parole et l'estimation du regard avec des corrections manuelles ciblées.

La thèse établit d'abord des régularités de base du regard en enfance moyenne en examinant sa prévalence globale. Les analyses montrent que les enfants de 6 à 12 ans présentent une préférence marquée pour fixer le regard sur le visage de l'interlocuteur. Au-delà de cette prévalence : (1) les enfants modulent leur regard en fonction de leur rôle conversationnel, avec davantage de regard dirigé vers le visage lorsqu'ils écoutent que lorsqu'ils parlent, selon un schéma comparable à celui des adultes. (2) Le regard est également organisé en fonction de la dynamique de prise de parole: les enfants ont davantage tendance à détourner le regard au début d'un segment de parole lorsqu'ils prennent le tour, et à le réorienter vers l'interlocuteur à la fin du segment lorsqu'ils cèdent la parole. Pris ensemble, ces résultats indiquent que la coordination du regard en enfance moyenne relève surtout du raffinement d'une compétence déjà en place, plutôt que de l'acquisition d'une nouvelle.

Enfin, nous introduisons une tâche prédictive de dynamique de prise de parole afin d'évaluer la contribution du regard à la gestion du tour de parole. Trois approches sont mises en œuvre et comparées: (1) un modèle fondé sur des variables explicites combinant des caractéristiques acoustiques et de regard, (2) un modèle séquentiel qui capture explicitement la structure temporelle de la parole et du regard, et (3) une approche par apprentissage de représentations s'appuyant sur des embeddings

acoustiques appris. Dans les trois cas, l'ajout de l'information de regard améliore la prédiction, ce qui montre que le regard apporte une information systématique au moment des transitions de tour.

Cette thèse soutient ainsi que le développement propre à l'enfance moyenne ne porte plus sur l'apprentissage de l'endroit où regarder, mais sur l'ajustement qualitatif de la manière d'utiliser le regard, avec fluidité, dans l'organisation de la conversation. Ces résultats fournissent des bases normatives essentielles pour des applications cliniques, ainsi que pour le développement d'interfaces socialement intelligentes, notamment dans les interactions enfant-robot.

Mots clés: enfance moyenne, regard, compétence conversationnelle, apprentissage automatique, oculométrie, données écologiques, prise de parole, développement de l'enfant

Acknowledgments

This thesis carries my name on its cover, but it was never mine alone. It grew out of the help and guidance of those who stood beside me. And now, as I ride off into the sunset, I want to honor those without whom these pages would never have existed.

First, I would like to express my deepest gratitude to the members of the jury: Bahia Guellai, Mohamed Chetouani, Mathilde Fort, Louise Goupil, and Philippe Blache. Having you read and discuss this work is something I value deeply. Thank you for dedicating your time and expertise to engaging with my research and for taking part in such an important moment of my life.

To my two supervisors, I owe more than words can capture.

Abdellah, you have been much more than a co-director. You have been an example, not only for me, but for many PhD students in TALEP and at the ILCB. Watching the way you think, the way you question, and the way you refuse intellectual shortcuts has shaped how I now approach research. Because of you, I leave this chapter not only with a PhD, but with a way of thinking and a standard I will carry with me wherever I go next.

Leonor, it is impossible to write these lines without emotion. You were first my teacher during my Master's, then my supervisor during my internship, and later became the director of this thesis. You have shaped this journey from its very beginning. You supervised this work while facing something far more serious than academic deadlines. Even during that time, you never stepped away. You remained present. You read drafts, you asked questions, you guided me forward. Seeing you back has meant the most to me in this entire journey.

I would also like to thank the members of my CSI, Magalie Ochs and Philippe Blache. You approached my work from distinct intellectual perspectives, often openly debating your positions. You consistently asked fundamental questions, sometimes as simple and as disarming as, "What is the title of your thesis?" Your intellectual clarity and rigor pushed me to sharpen my thinking and strengthen the foundations of this work at every stage.

To my lab mates — Abhishek, Ioana, Monica, Elie, Alice, Marion, Maiwenn, Eliot, Mitja, and Amaury — thank you for making these years feel shared. For the discussions, the debates, the moments of doubt we talked through, and the coffee breaks that saved the day. Research can be isolating, and as someone who tends to keep to

himself, I did not expect to find such ease and belonging in the lab. I never felt out of place, and it's because of you.

To my fiancée, you have been my anchor through every high and every low of this journey. You saw the doubts I did not always show to others, the stress, and the quiet frustrations that stayed behind closed doors. More than anyone, you know what this PhD truly cost... and what it truly meant. Thank you for carrying this with me.

And finally, to my parents. You sacrificed more than I will ever fully understand. You poured your energy, your comfort, and your dreams so that I could pursue mine. This thesis is mine in name, but it is yours in foundation. I hope it makes you proud.

Contents

Affidavit	2
Liste de publications et participation aux conférences	3
Abstract and keywords	5
Résumé et mots clés	7
Acknowledgments	9
Contents	11
List of Figures	15
List of Tables	19
Introduction	20
0.1. From Linguistic Competence to Conversational Mastery	20
0.2. The Visual Architecture of Interaction	22
0.2.1. Defining the Behavior: Gaze and the Primacy of the Face	22
0.2.2. The Functional Duality: Acquisition and Signaling	22
0.3. The Gap: Middle Childhood	23
0.4. The Ecological Imperative: Moving Beyond the Lab	25
0.5. Motivation and Scientific Significance: Why This Research Matters	25
0.5.1. Theoretical Necessity: Closing the Developmental Loop	26
0.5.2. Applied Necessity: Establishing Valid Baselines	26
0.5.3. Societal Necessity: The Rise of Child-Robot Interaction	26
0.6. Breaking the Annotation Bottleneck	27
0.7. Central Research Questions	27
0.7.1. Methodological Inquiry	27
0.7.2. Developmental Inquiry	28
0.7.3. Predictive Inquiry	29
0.8. Roadmap of the manuscript	29
1. The CHICA Corpus: Ecological Validity in Middle Childhood	31
1.1. Introduction	32
1.1.1. The Multifaceted Nature of Conversational Competence	32
1.1.2. A Developmental Lacuna: Middle Childhood	32
1.1.3. The Imperative of Ecological Validity	33

1.2.	Design Principles and Rationale	34
1.2.1.	Ecological context	34
1.2.2.	Intuitive Elicitation Task	34
1.2.3.	Lightweight Measurement Device	35
1.2.4.	Face-to-face vs. Video Calls	35
1.3.	Contextualizing CHICA within the Landscape of Developmental Corpora	36
1.4.	Recording Methodology	37
1.4.1.	Participants	37
1.4.2.	Conversational Task	37
1.4.3.	Logistics and Equipment	38
1.4.4.	Procedure	39
1.5.	Data Processing and Annotation	41
1.5.1.	Synchronization	42
1.5.2.	Transcription, Diarization, and Forced Alignment	42
1.5.3.	Non-verbal behavior	43
1.6.	Analyses	44
1.6.1.	Methods	45
1.6.2.	Results and Discussion	46
1.7.	Conclusion	48
2.	Automatic Annotation of Smiling and Gaze in Unconstrained Video Call Data	50
2.1.	Introduction	51
2.2.	Methodological Positioning and Related Work	52
2.2.1.	Automatic Detection of Facial Expressions in Video	52
2.2.2.	Automatic Gaze Estimation from Video	53
2.2.3.	Challenges of Unconstrained and Developmental Video Data	54
2.2.4.	Summary and Positioning of this work	54
2.3.	Methods	54
2.3.1.	Data and Annotation	54
2.3.2.	Pre-processing and Face Extraction	55
2.3.3.	Modeling Approaches	56
2.3.4.	Experimental Design and Evaluation	59
2.4.	Results	60
2.4.1.	Automatic Detection Performance	60
2.4.2.	Smile Detection	60
2.4.3.	Gaze Detection	61
2.4.4.	Reproducing Aggregate Behavioral Patterns	61
2.5.	Discussion	62
2.5.1.	Smile Detection: Robustness Under Cross-Domain Transfer	63
2.5.2.	Gaze Direction: Structural Limits Imposed by Video-Call Geometry	63
2.5.3.	Implications for Automatic Annotation in Ecological Interaction Research	64
2.5.4.	Methodological Consequences for the Present Thesis	64

3. Gaze on Face in Middle Childhood: An Emergent Adult-like Preference	65
3.1. Introduction: The "Gaze-on-Face" Phenomenon	65
3.2. The Literature and the Middle Childhood Gap	66
3.3. Quantifying "Gaze-on-Face"	67
3.3.1. Estimating "Gaze-on-Face"	68
3.4. A Preference for "Gaze-on-Face"?	71
3.5. Results: A Robust and Early Adult-like Preference	72
3.5.1. Analysis 1: The Spatial Distribution of Gaze	72
3.5.2. Analysis 2: The Proportion of "Gaze on Face"	74
3.6. Discussion: Re-evaluating the Developmental Trajectory	75
4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling	77
4.1. From Prevalence to Function: Posing the Central Problem	77
4.2. Methodological Note: Segmentation of Speech and Gaze Data	79
4.2.1. Speech Segments	79
4.2.2. 'Speech Segment' vs. 'Conversational Turn': A Critical Distinction	81
4.2.3. Gaze Segments	82
4.3. Analysis 1: Role-Based Gaze Modulation (Listening vs. Speaking)	83
4.4. Analysis 2: Turn-Coordination Gaze (Turn Taking vs. Yielding)	86
4.5. Discussion	89
4.5.1. Role-Based Gaze Modulation as a Stable Component of Interactional Competence	89
4.5.2. Gaze and Turn Coordination: Graded Precision Rather Than Absent Competence	90
4.5.3. Developmental Continuity and Refinement of Gaze Use	90
4.5.4. Implications for Modeling Turn-Taking Behavior	90
4.5.5. Summary	91
5. Predicting Turn-Taking Dynamics in Middle Childhood	92
5.1. Introduction	93
5.2. Task Definition: Hold vs. Shift	95
5.2.1. Temporal Properties of Turn Transitions	95
5.3. Modeling Approaches	96
5.3.1. Method 1: Feature-Based Prediction	97
5.3.2. Method 2: LSTM-Based Prediction	98
5.3.3. Method 3: Wav2Vec2 Representation	98
5.4. Results	99
5.4.1. Overview of Model Performance	99
5.4.2. Baselines	99
5.4.3. Feature-Based Models	100
5.4.4. Sequence-Based Models	100
5.4.5. Representation-Learning Models	100

5.4.6. Summary of Results	101
5.5. Discussion	101
5.5.1. The task is structurally asymmetric: Shift is learnable; Hold is not as much	101
5.5.2. Acoustic-only and gaze-only are not competing explanations; they fail in different ways	102
5.5.3. Naive multimodal fusion does not solve the hard part — and that is informative	102
5.5.4. Why LSTMs did not help: the bottleneck is not temporal modeling but available information	102
5.5.5. Deep acoustic representations improve the right thing when com- bined with gaze	103
5.5.6. What the results say about turn-taking rather than about models	103
5.5.7. Contribution	103
6. Conclusions	105
6.1. From Feasibility to Function: Revisiting the Central Questions	105
6.2. Establishing an Ecological Baseline for Middle Childhood Conversation	106
6.3. Automation and the Annotation Bottleneck Revisited	106
6.4. Gaze on Face as an Early-Stabilized Conversational Anchor	107
6.5. Role-Based Modulation of Gaze: Functional Deployment in Conversation	107
6.6. Gaze Dynamics	107
6.7. Multimodal Prediction of Turn-Taking: Insights and Limits	108
6.8. Implications for Theories of Conversational Development	108
6.9. Limitations	108
6.10. Future Directions	109
6.11. Conclusion	110
Bibliography	111
ANNEXES	120
A. Complementary Results: Classification Reports for Chapter 5	121
A.1. Baselines	121
A.2. Feature-Based Models	122
A.3. Sequence-Based Models	123
A.4. Representation-Learning Models (Wav2Vec2)	124

List of Figures

1.1. The Pupil Labs Invisible eyewear used to capture eye movements and first-person scene video.	35
1.2. Annotated overview of the Pupil Labs Invisible eye-tracking glasses. The device integrates multiple sensing and recording modules, including an inertial measurement unit (IMU) embedded in the frame for head-motion tracking; an RGB indicator for system status feedback; infrared (IR) illuminators and inward-facing eye cameras for pupil detection; optical lenses; a USB-C port for data transfer and charging; a unique serial identifier; and a forward-facing scene camera mounted on the left temple to capture the scene.	39
1.3. The recording procedure involved (1) video-call recording where the child and caregiver communicated from different rooms at home, and (2) face-to-face recording at home using mobile eye-gaze and head movement detection, in addition to fixed frontal view using smartphone cameras on tripods.	40
1.4. Eye-tracking glasses were calibrated before each use by asking participants to fixate on an object in the room. The gaze marker (indicated by the red circle) was then adjusted to match the target object in real time via streaming from the camera and the gaze signal.	41
1.5. Mean normalized contributions of children and parents to Word Count, Turn Count, Word Duration, and Turn Duration for three age groups (7, 9, 11), shown separately for Face-to-Face and Video-Call modalities. Normalization was performed per dyad by dividing each speaker's measure by the total measure for that dyad. Error bars represent the standard deviation of normalized values across participants.	46
1.6. Average percentage of time that children and parents spent looking at the screen during video-call sessions, grouped by child age (7, 9, 11). For each session, looking-at-screen proportions were computed separately for the child and the parent, then averaged within age groups. Error bars represent standard deviations.	47
2.1. The (balanced) accuracy score for Smile and Gaze comparing Feature-based models to End-to-End models. Ranges indicate 95% confidence interval over 8 unique participants in the CHICO corpus, both parents and children. Scores for each individual were obtained using models that had not seen that individual during training. The dashed lines represent chance.	60

2.2. The proportion of frames in a video where children (top) or adults (bottom) were smiling (right) or/and averting gaze (left). We show the numbers obtained for each of the 8 dyads in CHICO and the average ('ALL'), using both hand-coded and automatically coded frames.	62
3.1. Illustration of the RetinaFace detection and interlocutor selection process in a single egocentric frame from the Pupil Invisible glasses. The algorithm detects two faces (green bounding boxes), but only the larger one (on the right) is selected as the interlocutor based on our heuristic. The red circle indicates the most recent fixation event, and the blue circles represent previous fixations with IDs: 1149 and 1150, respectively. (Note: Faces are redacted for privacy.)	68
3.2. Histogram of gaze distances to the face center for a single participant during a session. The striped bars represent the relative frequency (in %) of gaze distances within 0–1000 pixels, while the smooth curve indicates the Kernel Density Estimation (KDE) of the distribution.	70
3.3. Histogram of gaze distances to the face center for a participant during a session. The dashed vertical line marks the GOF threshold separating two gaze states: GOF (Gaze On Face) on the left, and AG (Averted Gaze) on the right.	71
3.4. Kernel density estimates (KDEs) of gaze distances to the face center across 15 sessions. Each subplot compares the gaze behavior of the child (solid line) and parent (dashed line) within the same session. The x-axis represents gaze distance in pixels (0–1000), and the y-axis shows the relative frequency (%).	72
3.5. Averaged Kernel Density Estimates (KDEs) of normalized gaze distances across 15 child–caregiver sessions. Distances were normalized by each session's gaze threshold (<i>distance/threshold</i>). The solid line represents the mean distribution for children, and the dashed line represents the mean distribution for parents. Shaded areas indicate the standard deviation across sessions, reflecting inter-session variability in gaze behavior.	73
3.6. Proportion of gaze samples where the gaze distance was below the session-specific threshold (i.e., gaze-on-face) for children and parents across 15 sessions. Each subplot shows the percentage of gaze-on-face instances for the corresponding session, with children represented by hatched white bars and parents by lightly hatched grey bars. This comparison illustrates inter-session variability and relative differences in gaze-on-face behavior between children and parents.	74
3.7. Mean proportion of gaze samples where the gaze distance was below the session-specific threshold (gaze-on-face) averaged across 15 sessions. Error bars represent standard deviation across sessions.	75

4.1. Alignment between the audio waveform and transcribed speech segments for the Parent and Child. Each colored bar represents a speech turn, positioned according to its start and end times. The waveform above corresponds to the acoustic signal. Dotted vertical lines mark segment boundaries, and light gray shaded areas indicate pauses in the conversation. Together, these elements illustrate the temporal structure and alternation of turns between speakers.	80
4.2. Example of overlapping parent–child speech segments. The figure illustrates a short exchange where both speakers talk simultaneously, with the overlapping region shaded in light coral. The parent’s utterance (“est-ce que. . .”) begins as the child continues speaking (“En fait, c’est quelque chose t’ai pas vraiment pas prêt”), before the child completes their turn.	81
4.3. Synchronized example of gaze and speech. Top: paired egocentric frames with overlaid gaze vectors from child (left) and parent (right). Bottom: The figure illustrates a brief segment of interaction, featuring alternating speech turns between the parent and the child. Green and beige horizontal bands represent moments when each participant’s gaze was directed toward the other’s face (GOF). Time in seconds.	83
4.4. Gaze-on-face proportions for children and parents while listening and speaking. Each boxplot reflects the distribution across 15 sessions. . .	84
4.5. Gaze-on-face percentages for children and parents while listening and speaking across 15 recorded sessions. Each subplot represents one session and includes two bars per interlocutor.	85
4.6. Mean cross-correlation functions (-20 to +20 s) between gaze-on-face and speech activity, averaged across all dyads and sessions. Each plot shows separate curves for child gaze and parent gaze. The left panel corresponds to moments when the speaker is producing speech (“Speaker’s Own Speech”), and the right plot corresponds to moments when the partner is speaking (“Partner’s Speech”). Vertical dashed lines indicate zero lag.	87
4.7. Mean cross-correlation functions (-20 to +20 s) between gaze-on-face and speech activity, computed across dyads. The speaker is producing speech (“Speaker’s Own Speech”.	87
4.8. Mean cross-correlation functions (-20 to +20 s) between gaze-on-face and speech activity, computed across dyads. The partner is speaking (“Partner’s Speech”).	88

5.1. Gap duration distributions for Hold and Shift transitions. Distribution of silent gap durations preceding Hold and Shift. Although Shift transitions tend to be associated with shorter gaps on average, the substantial overlap between the two distributions indicates that gap duration alone is insufficient to reliably distinguish between turn holding and turn yielding in naturalistic conversation.	96
--	----

List of Tables

1.1. The distribution of our 15 dyads across age groups, children's gender, and parents' gender. The children's average age is given in years and months ($\pm$ average deviation from the mean).	37
2.1. Distribution of video frames in the ChiCo corpus across behavioral categories (smile and gaze) and preprocessing stages. The table reports the number and proportion of frames in the original recordings, after face detection with RetinaFace, and after feature extraction with OpenFace. Differences across stages reflect detection and feature extraction failures.	57
5.1. Summary of model performance on the Hold/Shift prediction task. . .	100
A.1. Classification report for the majority baseline.	121
A.2. Classification report for the random baseline.	121
A.3. Classification report for the Feature-Based Logistic Regression model.	122
A.4. Classification report for the gaze MLP model.	122
A.5. Classification report for the combined features MLP model.	122
A.6. Classification reports for LSTM models	123
A.7. Classification report for Wav2Vec2 (LR).	124
A.8. Classification report for Wav2Vec2 (MLP).	124
A.9. Classification report for the ensemble of models.	124

Introduction

Table of contents

0.1. From Linguistic Competence to Conversational Mastery	20
0.2. The Visual Architecture of Interaction	22
0.2.1. Defining the Behavior: Gaze and the Primacy of the Face	22
0.2.2. The Functional Duality: Acquisition and Signaling	22
0.2.2.1. Information Acquisition: The Listener's Mandate	23
0.2.2.2. Information Signaling: The Speaker's Regulation	23
0.3. The Gap: Middle Childhood	23
0.4. The Ecological Imperative: Moving Beyond the Lab	25
0.5. Motivation and Scientific Significance: Why This Research Matters	25
0.5.1. Theoretical Necessity: Closing the Developmental Loop	26
0.5.2. Applied Necessity: Establishing Valid Baselines	26
0.5.3. Societal Necessity: The Rise of Child-Robot Interaction	26
0.6. Breaking the Annotation Bottleneck	27
0.7. Central Research Questions	27
0.7.1. Methodological Inquiry	27
0.7.1.1. RQ1: Ecological Feasibility	28
0.7.1.2. RQ2: Automation	28
0.7.2. Developmental Inquiry	28
0.7.2.1. RQ3: Prevalence of Gaze	28
0.7.2.2. RQ4: Role-Based Modulation	28
0.7.2.3. RQ5: Turn-Signaling Dynamics	29
0.7.3. Predictive Inquiry	29
0.7.3.1. RQ6: Predictability of Turn Transitions	29
0.8. Roadmap of the manuscript	29

0.1. From Linguistic Competence to Conversational Mastery

Considered by many scholars, language is the defining characteristic of homo sapiens. For much of the 20th century, the study of language was largely influenced by the structural perspective, which viewed it as an abstract system of code: signs and rules. Championed by Noam Chomsky, this tradition focused on “linguistic competence” and the innate knowledge of grammar and syntax that enabled an ideal

speaker-listener to generate an infinite number of sentences. However, as the fields of sociolinguistics and pragmatics matured, a limitation of the traditional view became apparent: possessing the code is insufficient for communication. To be considered a competent conversationalist, the child must acquire what Hymes 1972 termed "communicative competence": the ability to use language appropriately in specific social contexts.

Communicative competence is not a fixed repertoire of social rules. It is a dynamic, real-time accomplishment. This dissertation situates itself within this interactional paradigm. This thesis adopts the view, eloquently articulated by Clark 1996, that language use is fundamentally a form of joint action. We consider conversation as a synchronized endeavor that requires participants to align their behaviors with high precision to ensure mutual understanding. This coordination is not just a matter of taking turns to speak, it does involve a continuous, multimodal exchange of signals: verbal, vocal, and visual. This exchange does allow the interlocutors to establish and maintain a shared understanding of the situation.

Conversation is a highly complex task. To engage effectively in a face-to-face conversation, a speaker must handle and manage parallel streams of information. They must navigate the system of turn-taking: more often than not by predicting the end of their partner's utterance with millisecond precision (Levinson 2016). They also must provide continuous listener feedback ("backchannels") to signal attention without disrupting the flow (Yngve 1970; Bavelas et al. 2000). They must possess repair strategies to signal and fix misunderstandings as they arise (Schegloff et al. 1977; Dingemanse, Torreira, et al. 2015). Furthermore, they must achieve interactive alignment (Pickering et al. 2004), a process in which interlocutors prime each other at multiple levels of representation, ranging from phonetics to syntax, and subsequently reducing the cognitive cost of language processing. This orchestration is multimodal: it requires the seamless integration of visual signals (like gaze) with the temporal flow of speech (Holler et al. 2019).

This interactional perspective fundamentally changes how we view language acquisition. It proposes that learning to speak involves more than internalizing grammar; it requires mastering a complex social performance. The child must grasp not only the fixed rules of language but also the dynamic, real-time processes for using them effectively in social interactions with others. This prompts an important developmental question: how do children develop these sophisticated, multimodal coordination skills?

Although developmental psychology has extensively mapped the acquisition of phonology and syntax, still, the ontogeny of these skills remains comparatively under-characterized, especially beyond tightly controlled laboratory settings. How do children go through the transition from the heavily scaffolded interactions of infancy to the fluid and autonomous adult-like conversations? This thesis examines this question by focusing on one of the most vital, yet frequently disregarded, elements of the interactional system: **eye gaze**. We examine the dynamics of gaze behaviors during middle childhood (ages 6 to 12). This study seeks to connect early childhood learning and adult communication skills by examining the fundamental baseline of

gaze dynamics during this stage.

0.2. The Visual Architecture of Interaction

If speech is the vehicle of conversation, eye gaze is the traffic control system. Vision is often characterized as the primary sense for perceiving the world and as a mechanism for gathering information about our physical environment (Gibson 2014). However, in the context of human interaction, the eyes serve a second, equally vital function: they are a tool for transmission (Argyle and Cook 1976). When two people converse, they do not merely look to see; they look to signal.

However, before examining its functions, we must first define the behavior itself and its primary focus: the face.

0.2.1. Defining the Behavior: Gaze and the Primacy of the Face

In social interactions, "gaze" differs from the biological process of vision. Vision is a sensory system used to perceive the environment, whereas social gaze is a behavioral cue. It involves deliberately turning the eyes and head toward another person, acting as a visible indicator of attention.

Within the visual modalities of the conversation, we consider a very important stimulus: the face. The face is not just a physical object; it is considered a transmitter of social and linguistic information. It is a dual-channel display: the eye region conveys mental states, attention, and emotional nuance (Baron-Cohen 1995), while the mouth region provides visual articulatory cues essential for speech decoding (Lansing et al. 1999).

Therefore, throughout this thesis, we focus on "Gaze-On-Face" (GOF) as the fundamental unit of analysis. Gaze on Face is the "anchor" of dyadic interaction. It is the default state of social connection, the baseline from which all other gaze behaviors, such as gaze aversion or referencing objects, are measured. To "gaze-on-face" is to engage with the interlocutor's mind and words simultaneously. Understanding the development of this specific orientation: how much we do it, and when we break it, is the key to understanding the visual structure of conversation.

0.2.2. The Functional Duality: Acquisition and Signaling

Spanning social psychology, anthropology, and conversation analysis, the literature on social gaze has robustly established that it is a dynamic signal that serves a dual function. This function shifts from moment to moment based on the participant's role in the conversation. We conceptualize this duality through the framework of Information Acquisition and Information Signaling (Argyle and Cook 1976; Kendon 1967; Gobel et al. 2015).

0.2.2.1. Information Acquisition: The Listener's Mandate

When an interlocutor assumes the role of the listener, gaze primarily serves to acquire. The listener looks to see. Directing visual attention to the speaker's face provides a wealth of critical data.

First, it enables monitoring of social feedback. By tracking the speaker's facial expressions, the listener can gauge emotional nuance, irony, and intent, which are often absent from the acoustic signal alone.

Second, gaze facilitates speech processing. The "McGurk Effect" (McGurk et al. 1976) famously demonstrated that visual information can alter auditory perception. In natural conversation, seeing the speaker's articulatory movements greatly improves speech recognition, especially in noisy environments. For children, whose linguistic processing systems are still refining, this visual anchor may be even more crucial than it is for adults.

0.2.2.2. Information Signaling: The Speaker's Regulation

When the interlocutor transitions to the role of the speaker, the function of gaze changes. It becomes a signaling and regulatory mechanism. Here, gaze aversion becomes paramount. Research has consistently shown that adults tend to avert their gaze at the onset of a speaking turn (Kendon 1967).

This aversion fulfills a cognitive function: by disengaging from the complex visual stimulus of the partner's face, the speaker reduces cognitive load, freeing up resources for lexical retrieval and syntactic planning (Glenberg et al. 1998; Doherty-Sneddon et al. 2012).

Simultaneously, it serves a social function: gaze aversion signals to the partner that the speaker is "holding the floor" and that they are not yet ready to yield. Conversely, as the speaker approaches the end of a turn, they typically shift their gaze to the listener. This behavior serves as a handover signal to indicate that the floor is open (Kendon 1967; Ho et al. 2015).

This intricate exchange of looking and looking away is the hallmark of adult conversational competence. It allows for seamless turn transitions with minimal gaps and overlaps. While these patterns are well-documented in adult interactions, their developmental status remains unclear. Whether such gaze-based regulatory mechanisms are already in place, still emerging, or functionally reorganized in middle childhood is an open empirical question that this thesis directly addresses. Do children intuitively understand and deploy this signal system? Do they use gaze to manage their cognitive load in the same way adults do?

0.3. The Gap: Middle Childhood

Work on the development of social gaze does not form a smooth developmental account. Instead, it clusters around two periods that have received sustained and care-

ful attention, leaving a fragmented temporal gap in the middle that is comparatively thinly documented.

The Infancy Period: On one end, a vast body of literature exists on early childhood. We know that infants exhibit an early sensitivity to eyes (Batki et al. 2000), that gaze-following is a precursor to Joint Attention (Carpenter et al. 1998), and that caregivers' gaze is a critical scaffolding tool for early word learning (Yu et al. 2013). In this stage, gaze is primarily studied as a mechanism for learning and bonding.

The Adult Standard: On the other hand, we have extensive models of adult interaction. In this context, the research community is no longer treating gaze as a channel for learning, but rather, as a mechanism for regulation and coordination (Duncan 1972; Goodwin 1981).

The Gap: middle childhood lies between the two poles; it spans roughly ages 6 to 12. This period remains comparatively under-characterized in the literature on non-verbal communication, particularly in ecologically valid settings. This neglect is paradoxical, as middle childhood is a period of profound socio-cognitive expansion. It marks the transition from the family-centric social world of the toddler to the peer-centric world of the school-age child. As a consequence, our understanding of gaze behaviour during middle childhood is often extrapolated from findings obtained either in infancy or in adulthood, rather than being grounded in direct observation of children's interactions at this age. Cognitively, this period is characterized by the refinement of Theory of Mind (the ability to attribute mental states to others), the maturation of executive functions (inhibitory control, working memory), and the development of sophisticated pragmatic skills (Wang et al. 2016; Matthews et al. 2018).

Given these cognitive leaps, one might expect children in this age group to exhibit fully mature gaze patterns. However, the limited research to date suggests a more complex picture. Some laboratory-based eye-tracking studies have reported a non-linear trajectory, showing that adolescents look less at their interlocutors than young adults (De Lillo et al. 2021). Importantly, these findings are primarily based on laboratory-based tasks and older age groups, leaving open the question of whether similar patterns are present earlier in development or whether they reflect context-specific artifacts rather than a genuine reduction in gaze.

This finding is theoretically challenging. Does it imply that the development of social cognition paradoxically suppresses face-directed gaze? Or does it reflect a "dip" in social confidence associated with adolescence? Or (and this is a central hypothesis of this thesis) are these findings artifacts of the specific contexts in which these subjects were observed? This ambiguity highlights the central problem this thesis seeks to resolve: to understand the true nature of gaze in middle childhood, we must confront the limitations of how we have historically studied it.

0.4. The Ecological Imperative: Moving Beyond the Lab

The data on children's gaze is conflicting. The trade-off between experimental control and ecological validity highlights a fundamental tension in behavioral science.

We argue that conversation is inherently context-sensitive. As Argyle and Dean 1965 and subsequent researchers have shown, gaze behavior is modulated by intimacy, hierarchy, and comfort. When we study children in a university laboratory, we introduce a host of confounding variables. The child is in an unfamiliar environment (the "stranger effect") (Ainsworth et al. 2015); they are often interacting with an adult researcher they do not know; and they may be encumbered by heavy stationary eye-tracking equipment or required to sit unnaturally still, also known as the "chin-rest constraint".¹

In such a context, a child's gaze aversion may not reflect their conversational competence. Instead, it may reflect social anxiety, deference to an authority figure, or simply the cognitive demand of performing a strange task. As Schegloff 1992 warned in his critique of experimental methods in pragmatics, strictly controlled designs risk "compromising the naturally occurring constitution of talk-in-interaction." By stripping away the natural context, we may be stripping away the very behavior we seek to measure.

To accurately characterize the conversational skills of school-age children, we face an ecological imperative. We must move from the laboratory into the real world. To understand how children truly interact, we must observe them in a context where they feel competent and comfortable: in their own homes, interacting with their natural interlocutors (their parents or caregivers).

This thesis is founded on this methodological shift. It introduces the CHICA (Child Interpersonal Communication Analysis) corpus, a novel dataset created specifically to address these limitations. The corpus features French-speaking dyads recorded in their homes. Furthermore, to avoid the asymmetry inherent in unstructured child-adult conversation, in which adults often dominate the floor, we used a prompt-free elicitation task—a word-guessing game—that enforces a balanced exchange of roles (Speaker/Listener) while maintaining the spontaneity of natural speech.

0.5. Motivation and Scientific Significance: Why This Research Matters

Why is it necessary to undertake such a complex, ecologically demanding study now? The motivation for this thesis extends beyond the simple desire to fill a gap

1. The "chin-rest constraint" refers to experimental setups in which participants' heads are immobilized using a chin rest, typically in eye-tracking or vision research. While this increases measurement precision, it restricts natural head movements, thereby limiting the ecological validity of observed gaze behaviour.

in the literature. It is driven by three urgent imperatives: theoretical completeness, clinical application, and the changing landscape of human interaction.

0.5.1. Theoretical Necessity: Closing the Developmental Loop

The current state of knowledge presents a fragmented view of human social development. We possess sophisticated models of how infants learn to connect (joint attention) and how adults manage the interaction, but we lack a coherent model of the transition. Without understanding middle childhood, we cannot answer fundamental questions about the ontogeny of social cognition. Do we rely less on visual cues as our linguistic processing improves? Or does visual social signaling become more refined as we age? By mapping this terra incognita, this thesis provides the missing link, the missing piece of the puzzle, that is required to construct a continuous, unified theory of human interactional development.

0.5.2. Applied Necessity: Establishing Valid Baselines

The stakes of this research are also practical. In clinical psychology and special education, diagnosing neurodevelopmental conditions such as Autism Spectrum Disorder (ASD) often relies on detecting deviations from "typical" gaze behavior. However, suppose our baselines for what is "typical" behavior are derived from artificial laboratory settings where even neurotypical children avert their gaze due to being shy or in discomfort. In that case, we risk constructing a flawed standard of comparison. By establishing what children actually do in comfortable, naturalistic settings, this thesis contributes to more accurate, ecologically valid normative baselines.

0.5.3. Societal Necessity: The Rise of Child-Robot Interaction

Finally, this research is driven by a forward-looking imperative to integrate technology in developmental contexts. Social robots and embodied Artificial Intelligence are emerging rapidly in sensitive domains such as education and therapy. However, a critical bottleneck in Human-Robot Interaction (HRI) is the "interactional mismatch": artificial agents often fail to sustain engagement with children because their underlying models are trained almost exclusively on adult data (Kennedy et al. 2015; Belpaeme et al. 2018). They lack the specific multimodal timing and gaze contingencies that characterize interactions in middle childhood.

This thesis addresses this adult-centric bias by providing the necessary "ground truth." By quantifying precisely how children regulate conversation in natural settings, we offer the behavioral parameters needed to model child-specific interaction dynamics. This is essential for developing socially intelligent interfaces that can adapt to the child's developmental reality, rather than forcing the child to adjust to the machine's rigid, adult-normative protocols.

0.6. Breaking the Annotation Bottleneck

Historically, the demand for ecological validity presented an insurmountable analytical barrier. Naturalistic conversation is "messy." It involves variable lighting, unconstrained head movements, overlapping speech, and dynamic environments. In the past, analyzing such data required manual frame-by-frame annotation, a process so labor-intensive that it limited sample sizes to a handful of minutes, rendering statistical generalization almost impossible.

This thesis represents a departure from this constraint, enabled by the convergence of two technological advancements: Mobile Eye-Tracking and Machine Learning.

Mobile Eye-Tracking: Lightweight, wearable glasses (such as Pupil Invisible) allow us to record gaze from the "first-person" perspective without restraining the child's head or movement, preserving the natural flow of interaction.

Machine Learning: Recent advances in deep learning (e.g., RetinaFace (Deng et al. 2020) for face detection, WhisperX (Bain et al. 2023) for speech segmentation) allow us to automate the processing of this "wild" data. We can now detect faces, map gaze vectors, and segment speech across hours of video with speed and consistency that manual coding could never achieve.

In this thesis, Machine Learning allows us to quantify social behavior at scales and resolutions previously reserved for laboratory studies, while retaining the ecological validity of the experiment.

0.7. Central Research Questions

This thesis investigates the development of face-directed gaze in middle childhood by combining an ecological study design with advanced computational analysis. By observing children aged 6 to 12 interacting with their parents in their home environment, and by using machine learning techniques to process the resulting multimodal data, this work aims to characterize gaze behavior as it unfolds in natural conversation. Crucially, we do not infer a predefined developmental trajectory; rather, we aim to decouple the influence of social context from the reality of cognitive development.

The dissertation is organized around a set of central research questions that mirror a progression from methodological feasibility to descriptive analysis, to interactional functionality of gaze dynamics, and finally to predictive modeling. This organization reflects a deliberate progression from establishing how interaction can be observed in everyday settings to examining how children coordinate their behavior within the flow of conversation.

0.7.1. Methodological Inquiry

The first phase of this thesis is dedicated to validating the instruments and protocols required for state-of-the-art developmental science. We will be addressing the methodological challenges that are inherent in studying conversation outside

the laboratory. We pose two fundamental questions regarding the feasibility of our approach.

0.7.1.1. RQ1: Ecological Feasibility

Our first research question asks: Is it possible to capture high-fidelity, multimodal behavioral data in an uncontrolled home environment that is robust enough for quantitative analysis? This inquiry challenges the field's traditional reliance on laboratory settings. We seek to determine whether we can successfully "move beyond the lab" without sacrificing the data quality necessary for rigorous scientific inference.

0.7.1.2. RQ2: Automation

Closely linked to acquisition is the challenge of annotation. Our second question asks: Can we reliably automate the detection of social signals (specifically gaze direction and smiles) in naturalistic, video-mediated data using modern computer vision techniques? This question addresses the "annotation bottleneck" that has historically hindered interaction research. This question evaluates the extent to which automated methods can support large-scale developmental research.

0.7.2. Developmental Inquiry

Once the methodological framework is established, the focus shifts to the empirical reality of middle childhood. To address this, I adopt a deliberately multi-perspective approach, examining gaze dynamics from several complementary levels. This choice reflects my view that gaze behavior in conversation unfolds at different levels of organization and cannot be reduced to a single dimension: overall prevalence, role-dependent organization, and temporal coordination each capture distinct aspects of interaction, and considering only one would provide an incomplete account of children's conversational behavior.

0.7.2.1. RQ3: Prevalence of Gaze

Our third question investigates the baseline behavior: When observed in a familiar, ecological context, do children demonstrate an adult-like preference for gazing at the face? This inquiry directly addresses the controversial hypothesis of a "developmental dip" in social attention during these years, positing that context may play a decisive role.

0.7.2.2. RQ4: Role-Based Modulation

Moving into the functional mechanics of interaction, our fourth question examines cognitive load management: Do children modulate their gaze based on their conversational role (Listening vs. Speaking)? Specifically, we investigate whether children, like

adults, avert their gaze when speaking to manage language planning and production demands, or if their gaze patterns are indifferent to their production state.

0.7.2.3. RQ5: Turn-Signaling Dynamics

Finally, we examine the fine-grained temporal coordination of gaze. Our fifth question asks: Have children mastered the dynamic use of gaze as a signal for turn-taking and turn-yielding? We analyze the temporal-level alignment between gaze and speech at turn boundaries to determine whether children actively use visual cues to facilitate smooth floor transitions.

0.7.3. Predictive Inquiry

The final phase of the thesis adopts a predictive perspective, building on the descriptive analyses of gaze behavior and conversational roles. This phase examines whether multimodal cues carry sufficient regularity to inform computational models of turn-taking.

0.7.3.1. RQ6: Predictability of Turn Transitions

The sixth research question asks whether turn transitions in child–caregiver conversation can be predicted from multimodal features, including gaze behavior and acoustic cues. Specifically, it assesses how well different modeling approaches can discriminate between turn-holding and turn-shifting events. Importantly, predictive performance is not treated as evidence of communicative intent or mature signaling strategies. Instead, it serves as an analytical tool: limitations in predictability help clarify the extent to which turn-taking behavior in middle childhood is systematically organized and accessible to computational modeling.

0.8. Roadmap of the manuscript

In this thesis, I set out to develop a clear and coherent account of gaze behavior in middle childhood. The structure of the manuscript reflects this stance: the chapters move from establishing the empirical basis of the study, structured around two guiding commitments: to study interaction as it occurs in everyday settings, and to approach gaze as a behavior that operates at several interacting levels. For that reason, we move progressively to more focused analyses of gaze coordination, to functional and predictive perspectives on turn-taking. Finally, we bring these perspectives together to clarify what the findings contribute to broader accounts of social interaction and development.

Chapter 1 – The CHICA Corpus: Ecological Validity in Middle Childhood. This chapter presents the design and implementation of the CHICA corpus. It discusses the recruitment of French-speaking child–caregiver dyads, the recording procedures in both face-to-face and video-mediated settings, and the technical choices underlying

the use of mobile eye trackers and multimodal recording. It details the data processing pipeline: synchronisation, transcription, diarisation, forced alignment, and preliminary non-verbal annotations. This chapter ends with initial descriptive analyses of conversational dynamics across ages and modalities.

Chapter 2 – Automatic Annotation of Smiling and Gaze in Unconstrained Video Call Data. This chapter addresses a central methodological challenge of the thesis: how to analyze non-verbal behavior in large-scale, naturalistic video data without relying exclusively on manual annotation. Using children’s video-call data as a test case, I examine the extent to which current computer vision methods can detect socially relevant signals such as gaze direction and smiling in unconstrained recording conditions. This chapter does not present automation as a definitive solution, but rather, it evaluates what these models reliably capture, where they break down, and how these limitations informed the design and interpretations of subsequent analyses.

Chapter 3 – Gaze on Face in Middle Childhood: An Emergent Adult-like Preference. When children interact face-to-face with a caregiver in everyday settings, where do they look? Building on the face-to-face portion of CHICA, this chapter focuses on the “gaze-on-face” phenomenon. It introduces a measure of gaze-on-face by quantifying the spatial distribution of gaze in conversation. The analyses reveal a robust preference for gaze on face already in middle childhood, prompting a reconsideration of assumptions about the protracted development of conversational gaze.

Chapter 4 – Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signalling. Where Chapter 3 establishes overall prevalence, this chapter moves beyond overall prevalence to examine how gaze unfolds over the conversation. It focuses on the relation between gaze behavior, conversational role, and turn transitions. The chapter analyzes gaze patterns during speaking and listening, and around moments of turn change.

Chapter 5 — Predicting Turn-Taking Dynamics in Middle Childhood. This chapter adopts a predictive perspective to examine the extent to which conversational structure in middle childhood can be captured from observable multimodal cues. Building on the descriptive analyses of gaze dynamics developed in earlier chapters, it introduces a turn-taking task to classify turn-shifting and turn-holding events. The chapter explores different modeling approaches and feature representations, using prediction not as an end in itself, but as a way to assess whether gaze behavior provides reliable cues to turn transitions.

The thesis concludes with a final chapter that steps back from the individual analyses to reflect on what this work has made possible. It brings together the methodological choices, empirical observations, and modeling results developed throughout the thesis to construct an account of gaze behavior in middle childhood. I will be revisiting the research questions stated above to clarify how they are addressed across the different parts of the manuscript and to construct a final argument for my stances. Finally, this chapter identifies the methodological and empirical boundaries of the present work and uses them to motivate concrete directions for future research.

1. The CHICA Corpus: Ecological Validity in Middle Childhood

Table of contents

1.1. Introduction	32
1.1.1. The Multifaceted Nature of Conversational Competence	32
1.1.2. A Developmental Lacuna: Middle Childhood	32
1.1.3. The Imperative of Ecological Validity	33
1.2. Design Principles and Rationale	34
1.2.1. Ecological context	34
1.2.2. Intuitive Elicitation Task	34
1.2.3. Lightweight Measurement Device	35
1.2.4. Face-to-face vs. Video Calls	35
1.3. Contextualizing CHICA within the Landscape of Developmental Corpora	36
1.4. Recording Methodology	37
1.4.1. Participants	37
1.4.2. Conversational Task	37
1.4.3. Logistics and Equipment	38
1.4.3.1. The video call recording step	38
1.4.3.2. The face-to-face recording step	38
1.4.4. Procedure	39
1.4.4.1. Video call recording	40
1.4.4.2. Face-to-face recording	40
1.5. Data Processing and Annotation	41
1.5.1. Synchronization	42
1.5.2. Transcription, Diarization, and Forced Alignment	42
1.5.3. Non-verbal behavior	43
1.5.3.1. Video Call Data Enrichment	43
1.5.3.2. Face-to-Face Data Enrichment	44
1.6. Analyses	44
1.6.1. Methods	45
1.6.2. Results and Discussion	46
1.7. Conclusion	48

1.1. Introduction

This chapter outlines the theoretical and methodological foundations of the CHICA corpus (for **Child Interpersonal Communication Analysis**). We begin by examining the complexity of conversational competence and the specific developmental gap remaining in middle childhood research. We then identify the methodological limitations of existing datasets (specifically the lack of ecological validity) before introducing the CHICA corpus as a resource designed to bridge these gaps through naturalistic, multimodal data collection.

1.1.1. The Multifaceted Nature of Conversational Competence

Becoming a competent conversationalist in one's native language is a cornerstone of human cognitive and social development. This process involves more than simply mastering the building blocks of language, such as phonology, vocabulary, and syntax. Children must also learn to apply these tools with the dynamic, adaptive skills necessary for successful face-to-face interaction (Clark 1996). To achieve adult-level mastery, children must acquire a complex set of abilities that allow them not only to speak but also to coordinate actions with a partner. Scholars have identified a variety of skills essential for this orchestration, including the fluid management of turn-taking (Levinson 2016; Casillas et al. 2016; Nguyen et al. 2022), the appropriate use of listener feedback and backchannels (Bavelas et al. 2000; Bodur, Nikolaus, Prévot, et al. 2023; Dingemanse, Torreira, et al. 2015), repair strategies to resolve communicative breakdowns (Clark and Brennan 1991; Purver 2004; Dingemanse and Enfield 2024), and the ability to produce contingent and coherent responses (Grice 1975; Keenan et al. 1975; Bloom et al. 1974; Abbot-Smith et al. 2023).

Crucially, these verbal skills do not operate in a vacuum; they are deeply intertwined with a rich repertoire of non-verbal signals that facilitate the dual imperatives of communication: information acquisition and signaling. Among these, eye gaze is particularly vital. It functions as an input channel for information acquisition, allowing speakers to monitor their partner's attention and comprehension. Simultaneously, it serves as an output channel for signaling social intent, such as regulating turn transitions or displaying engagement (Kendon 1967; Argyle and Cook 1976). Mastering this multimodal orchestration allows interlocutors to achieve interactive alignment (Pickering et al. 2021; Fusaroli et al. 2023), a state of mutual synchronization that is crucial for a child's transition from a language learner to a socially competent conversational partner.

1.1.2. A Developmental Lacuna: Middle Childhood

Scholarly attention has primarily focused on the bookends of development—the conversational precursors in early childhood (e.g., Donnelly et al. 2021; Masek et al. 2021; Elmlinger et al. 2023; Nikolaus et al. 2023) and the polished skills of adults. Leaving a significant gap in our understanding of the developmental path between them.

1. The CHICA Corpus: Ecological Validity in Middle Childhood – 1.1. Introduction

This gap is particularly glaring during middle childhood. During this period, children are no longer learning the basic rules of conversation (which are acquired mainly in early childhood, see Casillas et al. 2016). Instead, they are fine-tuning the precise timing and coordination required for adult-level fluency (Hess et al. 1988; Anderson, Clark, et al. 1994; Bodur, Nikolaus, Prévot, et al. 2023). The few existing studies on this period suggest a protracted developmental trajectory for conversational skills, spanning much of the school-age years and sometimes extending into adolescence (Hess et al. 1988; Sehley et al. 1992; Maroni et al. 2008; Baines et al. 2010). This period also marks a profound maturation in the socio-cognitive faculties that enable sophisticated conversation. Core abilities such as mentalizing¹, working memory, inhibitory control, and other executive functions are notably refined during these years (Wang et al. 2016; Matthews et al. 2018). Put all together, this is a critical period for investigating how the development of these cognitive skills maps onto observable improvements in conversational competence. Yet, progress in this area has been hampered by a persistent challenge: a scarcity of rich, naturalistic data from children in this age group.

1.1.3. The Imperative of Ecological Validity

A central challenge in the field of inquiry is the well-known "observer's paradox": the very act of studying a conversation can alter its very nature. This issue is especially pronounced when examining natural face-to-face conversation. As conversation analysts have long maintained (Sacks et al. 1974; Schegloff 1991), and as Schegloff et al. 1996 later argued, highly controlled laboratory experiments may offer a high degree of precision, but risk "compromising the naturally occurring constitution of talk-in-interaction."

This risk is higher when studying non-verbal behaviors such as eye gaze. Traditional eye-tracking paradigms often require participants to maintain a fixed head position (e.g., using a chin rest) or interact primarily with a screen rather than a person. On the other hand, studies that forgo these constraints to allow for unrestricted movement typically rely on external video cameras, which necessitate labor-intensive manual annotation. This methodological bottleneck often limits the scale of analysis (and in some cases the precision). The artificial environment of a lab and the presence of an unfamiliar experimenter does go against the spontaneity of the interaction, the very thing we aim to understand. It is now well-established that social interactions are highly context-sensitive (Kleinke 1986; Risko et al. 2016; Dideriksen et al. 2023). To authentically capture the "fine-tuning" of children's communicative skills, observation must occur in settings where these skills are used freely, comfortably, and naturally, while still striking a balance with the precision of the measurement.

This chapter introduces the CHICA corpus, a novel developmental resource specifically designed to address the lacuna above. The principal objective of its creation was

1. Mentalizing, often used interchangeably with Theory of Mind, refers to the cognitive capacity to attribute mental states (such as beliefs, intents, desires, and knowledge) to oneself and others.

the collection of rich, multimodal data from child-caregiver conversations in a maximally naturalistic environment. By leveraging recent advances in mobile, lightweight sensor technology and automated data processing tools, the CHICA corpus was curated to provide a precise, yet ecologically sound, foundation for examining the nuances of communicative development in middle childhood.

1.2. Design Principles and Rationale

The design of the CHICA corpus was guided by a foundational commitment to ecological validity. The objective was to document interactions that are as optimally representative as possible of the spontaneous conversations that children have on an everyday basis. This principle informed four core methodological decisions regarding: context, task, measurement technology, and medium.

1.2.1. Ecological context

We recorded conversations between school-age children and their parents at home. While conducting recordings in a laboratory setting with an experimenter would have ensured standardized conditions, it would have prioritized experimental control over ecological validity. This is a crucial consideration, as social behaviors are highly context-sensitive (Risko et al. 2016). An interaction with a stranger in a novel setting may not be indicative of a child's spontaneous behavior under more familiar conditions. The home environment, in contrast, is hypothesized to be more conducive to eliciting the full spectrum of a child's communicative repertoire.

1.2.2. Intuitive Elicitation Task

Researchers have traditionally used physical prompts to elicit discussions, such as the maze game, the map task, or the spot-the-difference tasks (Garrod et al. 1987; Anderson, Clark, et al. 1994; Van Engen et al. 2010). However, in our pilot, we realized that such prompts tend to capture children's attention, making the face-to-face multimodal interaction suboptimal. To elicit spontaneous interaction, we moved away from traditional physical prompts. Instead, we opted for a loosely structured word-guessing game. This design was chosen to enforce three critical interactional features:

- **Absence of a Physical Prompt:** Unlike tasks that are centered around an object, the game has no shared physical focus. The relevant information exists solely between the two participants, making the conversational partner the primary focus. This design inherently promotes the face-directed gaze and the multimodal signaling in question.
- **Enforced Reciprocity:** The game's primary rule is that roles must switch after each successful guess. This serves as a structural mechanism to counterbalance the inherent social asymmetry of the child-adult dyad. This ensures that

1. The CHICA Corpus: Ecological Validity in Middle Childhood – 1.2. Design Principles and Rationale

both participants have regular opportunities to initiate and respond, so we can guarantee a more equitable distribution of speaking and listening turns.

- **Conversational Flexibility:** The rules were intentionally minimal. Participants were not limited to yes/no questions and were explicitly encouraged to ask for and provide hints as they see fit. This flexibility enabled a wide range of natural conversational phenomena, such as clarification requests, repairs, and elaborations, making the interaction as close as possible to a spontaneous dialogue.

1.2.3. Lightweight Measurement Device

A central objective was the acquisition of precise, continuous data on eye gaze and head movement: two non-verbal signals critical for regulating social interaction (Kendon 1967; Hale et al. 2020) and known to have a relatively late developmental trajectory (Hess et al. 1988; De Lillo et al. 2021). To maintain a high degree of naturalness with minimal disruption, the study employed a state-of-the-art wearable eye-tracking device, the "Pupil Invisible" eyeglasses (Tonsen et al. 2020), which consist of an eye-tracking system, a gyroscope (measuring angular velocity), and an accelerometer (measuring linear acceleration). This device is lightweight (<50g) and visually indistinguishable from conventional eyeglasses, rendering it unobtrusive for the wearer and minimally distracting for their conversational partner. This technology has proven effective for capturing gazing patterns in naturalistic settings, even with young children (Schroer et al. 2023). This enabled the collection of highly accurate gaze and head motion data with minimal interference with the natural flow of the interaction.



Figure 1.1. – The Pupil Labs Invisible eyewear used to capture eye movements and first-person scene video.

1.2.4. Face-to-face vs. Video Calls

To better understand the specificities of face-to-face conversation, we contrast it with another popular medium of communication: Video calls. We optimize our

ability to draw valid conclusions from this comparison by adopting a within-dyad design: The child and parent played the same conversational game both via video call and face-to-face. Here again, to make the conversation feel more natural, the video call also takes place at home. The caregiver and child use different devices, and they communicate from separate rooms. They were instructed to act as if they were communicating from remote locations.

1.3. Contextualizing CHICA within the Landscape of Developmental Corpora

The creation of the CHICA corpus was motivated by specific gaps identified in the existing landscape of developmental language resources. While many valuable corpora exist, they are often ill-suited for the fine-grained analysis of multimodal, face-to-face interaction in middle childhood.

Most large-scale resources, such as the invaluable CHILDES repository (MacWhinney 2014), primarily rely on audio recordings or video captured from a static, third-person perspective. While instrumental for lexical and syntactic analyses, this recording methodology inherently limits the study of face-to-face interaction, as it obscures the subtle, yet critical non-verbal cues, such as precise gaze direction, mutual gaze, and facial expressions, that are central to conversational studies.

More recent efforts have sought to capture the child's perspective more directly through the use of head-mounted cameras, yielding rich egocentric datasets (e.g., Sullivan et al. 2021). These corpora provide an unprecedented window into the child's visual world and have been particularly fruitful for studying the environmental inputs that support language learning (e.g., Vong et al. 2024). However, their primary focus is on capturing the child's perceptual input; they do not simultaneously capture the child's own facial expressions, gestures, or gaze behavior. Consequently, they are not optimized for analyzing the child's active role as a multimodal communicator within a dyadic exchange.

Other corpora are more aligned with the present work's focus on multimodality but differ in key respects. The Ecolang corpus (Gu et al. 2025), for instance, provides rich multimodal data but focuses on a much younger age group, addressing developmental questions pertinent to early language acquisition rather than the refinement of conversational skills in school-age children. Closer in age range is the corpus introduced by Bodur, Nikolaus, Kassim, et al. 2021; however, the dataset is composed exclusively of video-call interactions. While valuable for studying computer-mediated communication, it does not permit the direct comparison between face-to-face and video call modalities, a key design feature of the CHICA corpus.

Thus, to our knowledge, the CHICA corpus is the first resource specifically designed to fill this methodological niche. Its novelty is fourfold: (1) its profound commitment to ecological validity, capturing interactions in the home environment to ensure the data is as naturalistic as possible; (2) its specific focus on the under-researched period of middle childhood; (3) its collection of high-fidelity, dual-perspective eye-tracking

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

Table 1.1. – The distribution of our 15 dyads across age groups, children’s gender, and parents’ gender. The children’s average age is given in years and months ($\pm$ average deviation from the mean).

gaze data; and (4) its unique within-dyad design, which directly contrasts face-to-face and computer-mediated conversational contexts.

1.4. Recording Methodology

1.4.1. Participants

The sample consists of $N = 30$ French-speaking children and their caregivers, making 15 dyads, 5 per age group (approximately 7, 9, and 11 years old). This comprises a total of approximately 4 hours of face-to-face conversations and a parallel 4 hours of video calls. Procuring datasets of this nature presents considerable challenges, primarily due to the difficulty of recruiting volunteers willing to undertake recordings of their children in their home environments. The aim is to strike a balance between obtaining a feasible sample size and ensuring rich and ecologically valid intra-individual data. See demographic information in Table 1.1.

1.4.2. Conversational Task

The elicitation of interactional data posed a specific methodological challenge: the need to foster a conversation that was both naturalistic enough to be ecologically valid, yet structured enough to be balanced and analyzable. Pilot studies revealed that entirely unstructured conversations with child-caregiver dyads were often dominated by the parent. At the same time, conventional object-focused tasks (e.g., map tasks) drew visual attention away from the conversational partner, thereby suppressing the very gaze behaviors central to our research.

To address this, each dyad was instructed to play a word-guessing game intended to elicit naturalistic, balanced conversation. The rules were simple and loosely structured: one participant thinks of a word, and the other attempts to guess it by asking any question (i.e., not restricted to yes/no questions). Upon a successful guess, the roles would switch. The game was designed to last 10-15 minutes, with parents instructed to conclude the session once both participants had guessed a similar number of words, ensuring a relatively balanced exchange.

Therefore, we have engineered the word-guessing game to be a methodologically grounded tool to elicit a specific type of interaction: one that is reciprocal, face-

focused, and rich in spontaneous, multimodal behaviors relevant to this thesis.

1.4.3. Logistics and Equipment

The two recording modalities, video calls and face-to-face interactions, had distinct logistical and equipment requirements.

1.4.3.1. The video call recording step

Families were required to provide:

- Two network-enabled devices (e.g., two computers, or a computer and a tablet/smartphone), each equipped with a functioning camera and microphone.
- A high-speed internet connection to ensure stable call quality.
- The Zoom video conferencing software is installed on both devices.
- Two separate rooms within the home, sufficiently distant to prevent direct auditory bleed-through.

1.4.3.2. The face-to-face recording step

The logistics required from the family were only a room and two chairs. Additionally, the researcher brought with them the following equipment:

- Two Pupil Invisible eye-tracking eyeglasses (see Figure 1.2), which served as the primary data acquisition devices (their size and shape are very similar to typical eyeglasses: $144mm \times 48mm \times 160mm$.)
- Two Samsung A135 smartphones with a $8160 \times 6120px$ camera resolution and a microphone.
- Two tripods to mount a second set of smartphones for recording fixed, frontal views of each participant.

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

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



Figure 1.2. – Annotated overview of the Pupil Labs Invisible eye-tracking glasses. The device integrates multiple sensing and recording modules, including an inertial measurement unit (IMU) embedded in the frame for head-motion tracking; an RGB indicator for system status feedback; infrared (IR) illuminators and inward-facing eye cameras for pupil detection; optical lenses; a USB-C port for data transfer and charging; a unique serial identifier; and a forward-facing scene camera mounted on the left temple to capture the scene.

1.4.4. Procedure

Interested families filled out an online form. The form linked to documents that explained the procedure in detail (including the required hardware and software), as well as to the documents related to the consent and data protection forms. Parents had to agree to the procedure and provide their consent before they could proceed with the registration.

Data collection was done in two steps: 1) video call recording, and 2) face-to-face recording, as follows (see also Figure 1.3).



Figure 1.3. – The recording procedure involved (1) video-call recording where the child and caregiver communicated from different rooms at home, and (2) face-to-face recording at home using mobile eye-gaze and head movement detection, in addition to fixed frontal view using smartphone cameras on tripods.

1.4.4.1. Video call recording

Parents booked an online Zoom appointment with the researcher. During the online appointment, the researcher explained the procedure for the video recording step. He made sure that: (a) both the child and parents were well positioned (fully visible on the screen), (b) there were no sound issues or echoes (in case the child and parent's devices were not distant enough from each other), and (c) both have checked “hide selfview” and pinned the other interlocutor's window on full-screen mode (so the child only sees the parent and vice versa). The researcher remained in the Zoom conversation, ensuring he was entirely out of view. This was achieved by the researcher turning his camera and microphone off and asking both participants to check “hide non-video participants.” Once the participants are ready, the researcher starts the recording, and the dyad starts playing the word-guessing game. After they are done, the researcher stops the recording and turns on his camera so that he can reappear in view. He then congratulates the child and organizes the next step (face-to-face recording session) with the parents.

1.4.4.2. Face-to-face recording

At the end of the Zoom recording, the researcher and the parent agreed to a future in-person appointment at the family's home. The recording procedure during the appointment was as follows. The researcher first makes sure the lighting and chair arrangement are adequate. Once the interlocutors are seated, the researcher installs the smartphones on tripods, ensuring they capture a clear, frontal view of the other interlocutor. The researcher assists the participants in wearing the Pupil Invisible device and, if necessary, uses dedicated head straps (also manufactured by the same company) to secure it, especially for children. The researcher calibrated the device before each use. This was done by asking each participant to fixate on different objects in the room (without moving their head) and then adjusting the glasses' placement

accordingly. This process was facilitated by the proprietary Pupil software, which enabled real-time streaming of gaze data on a dedicated smartphone, overlaid on the participant's egocentric view (see Figure 1.4).



Figure 1.4. – Eye-tracking glasses were calibrated before each use by asking participants to fixate on an object in the room. The gaze marker (indicated by the red circle) was then adjusted to match the target object in real time via streaming from the camera and the gaze signal.

Note that the eyeglasses needed to be cable-connected to a smartphone for all computation and storage of the recorded data. While this choice from the manufacturer can be understood as making the device itself less cumbersome, the cable adds some constraints on movement. In our case, however, both speakers were in a sitting position, and the cable was long enough to allow free head movement. Finally, just before the start of the game, the researcher made a clap to serve as an audiovisual marker to facilitate later synchronization. The researcher then retreated to a corner (or a different room when possible), informing the participants he would be busy with a different work-related activity, to minimize interference or the feeling of being observed by a third person.

1.5. Data Processing and Annotation

To analyze the simultaneous audio-visual data streams from the CHICA corpus, we first established a reliable processing workflow. Like all naturalistic corpora, we faced the primary challenge of the annotation bottleneck: manually coding continuous behaviors, such as gaze and speech, is prohibitively time-consuming. Therefore, we developed a hybrid processing pipeline to handle this task. This approach combined (1) automation, using advanced computational tools for speed and scalability, with (2) manual correction and annotation. The manual step was crucial, where automated tools proved unreliable on our data. This section outlines the multi-stage workflow of

this pipeline, from initial data synchronization to the final extraction of verbal and non-verbal data streams that form the basis for the following chapters.

1.5.1. Synchronization

The video call data, recorded via Zoom, did not require synchronization because Zoom already provides separate, synchronized audio channels for each participant.

The face-to-face data involved aligning four independent audio-visual streams for each dyad: two egocentric views (from the Pupil Invisible devices) and two frontal views (from tripod-mounted smartphones). Rather than relying solely on a manual clap marker (which the researcher signaled moments before the conversation began), we developed a custom automated synchronization pipeline based on audio cross-correlation:

- Signal Extraction: Audio streams were extracted from the video files.
- Pairwise Cross-Correlation: The tool accepts two video files as input. It computes the cross-correlation function between their respective audio signals. This mathematical operation slides one signal over the other to identify the specific time lag (τ) that maximizes the similarity between their waveforms.
- Alignment: The calculated offsets were applied to the video streams, effectively locking them to a shared timeline based on the audio's shared acoustic form.

While the initial clap provided a coarse check, this cross-correlation method ensured precise, sample-level synchronization across all four video streams. The accuracy of this automated synchronization was subsequently verified manually using ELAN (Wittenburg et al. 2006).

1.5.2. Transcription, Diarization, and Forced Alignment

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

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

The way we did the transcription differed between the video calls and face-to-face conditions. For video calls, Zoom allows recording in a separate audio channel for each speaker, thereby bypassing speaker diarization². A manual investigation of the

2. Diarization is the process of identifying who is speaking and when; a classification crucial for

transcription confirmed its high quality and low error rate.³

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

The overall transcription of face-to-face data, using only one channel for both interlocutors, was, as one would expect, lower in quality than video-call recordings with two separate channels. In addition, some turns were missing, such as short responses (e.g., "yes" and "no"), especially when overlapping with the interlocutor's speech. Thus, the automatic transcription of the face-to-face recording was entirely corrected by hand, including the addition of any missing turns. Crucially, we performed manual speaker labeling for every transcribed turn. This labor-intensive step was essential to ensure gold-standard speaker attribution, particularly for analyzing rapid turn-transitions where automated diarization errors typically cluster.

1.5.3. Non-verbal behavior

We extracted both continuous signals (time series) and categorical events (behavioral codes) for gaze and head movement. Due to the distinct recording technologies used in each condition, the enrichment pipelines differed significantly. Accordingly, we detail the specific data processing strategies employed for the video call and face-to-face settings separately below.

1.5.3.1. Video Call Data Enrichment

- Continuous Signals: We extracted time-continuous measures of gaze and head orientation indirectly using OpenFace (Baltrušaitis et al. 2016). Since these measures are derived from 2D video rather than physical sensors, they represent estimated pose and gaze vectors. To facilitate the detection of nods and shakes, we computed rotation velocities along the x-axis (roll) and y-axis (pitch) within a sliding, fixed-time window of 20 frames (video is at 25 frames per second).
- Categorical Annotation: While recent advances in computer vision allow for automated gaze-behavior categorization in constrained settings (e.g., Erel et al. 2023), these tools proved less reliable for our naturalistic, unconstrained video

accurate transcription of dialog, especially when there is speech overlap between interlocutors

3. Note, however, that while Whisper is good at capturing the semantic content, its transcription systematically filtered out several forms of disfluency, e.g., "uh" and "um" (see Radford et al. 2023). Such units can be important when analyzing dialog; their study in our corpus would require inserting them back into the transcript manually and/or using specific detection techniques.

1. The CHICA Corpus: Ecological Validity in Middle Childhood – 1.6. Analyses

call data due to high between-subject variability (citegoumri2023automatic, chapter 2). Consequently, we resorted to complete manual annotation for gaze categorization. We adopted the coding scheme from Bodur, Nikolaus, Kassim, et al. 2021, categorizing gaze into "looking at screen" (a proxy for "gaze on face") versus "looking away" (a proxy for "averted gaze"). Note that this distinction does not precisely map onto face-to-face gaze due to the webcam offset, but it reflects the constraints inherent to computer-mediated communication. Mainly because the webcam is not aligned with the face; however, this is a constraint inherent to most current commercial video call software, and people have had to adapt to it. In fact, one of the many goals of the current corpus is to enable future research to investigate how such constraints may influence gaze dynamics in online conversation compared to face-to-face conversation.

- Inter-Rater Reliability: To ensure reliability, a human annotator was first trained on 80% of the videos from the Bodur, Nikolaus, Kassim, et al. 2021 dataset. Then, a second independent annotator coded 20% of the remaining video of the CHICA corpus to estimate inter-rater reliability. Since this task required not only assessing agreement on identification (whether a gaze was detected) but also agreement on segmentation (start-time and end-time boundaries), we cannot use standard measures like Cohen's Kappa. Instead, we used the Staccato algorithm Bodur, Nikolaus, Prévot, et al. 2023 implemented in ELAN. Running 1000 Monte Carlo Simulations with a 10ms granularity ($p < 0.05$), we achieved an agreement score (degree of organization) of 0.66, indicating substantial reliability. After reaching this relatively good agreement score, our annotator then coded all videos in our corpus.

1.5.3.2. Face-to-Face Data Enrichment

- Continuous Signals: In the face-to-face condition, we obtained direct, high-precision time series for gaze and head movement via the sensors integrated into the Pupil Invisible eyeglasses.
 - Gaze: The device's internal eye cameras map eye movements to the scene camera's coordinate system at 200Hz.
 - Head Movement: An integrated Inertial Measurement Unit (IMU) provided gyroscope and accelerometer data (200Hz) to track head orientation and translational acceleration.
- Categorical Annotation: Unlike the video call data, we used a specific pipeline to estimate "gaze-on-face" from raw time-series gaze data. A comprehensive description of this gaze-on-face estimation method is provided in Chapter 3.

1.6. Analyses

In this section, we provide preliminary analyses to demonstrate the richness of the corpus for studying child-caregiver conversational dynamics across interlocutor (child

or caregiver), developmental age in middle childhood (7, 9, 11 years old), modality (e.g., verbal and non-verbal), and medium of communication (i.e., face-to-face vs. video call). The goal is not to provide a thorough scientific study or test specific hypotheses, but rather, to summarize the main characteristics of our data using, mainly, high-level measures.

1.6.1. Methods

For the verbal modality, we quantified four primary metrics, utilizing distinct processing stages for each:

- Word Count: Calculated from the manually corrected automatic transcripts.
- Turn Count & Turn Duration: Derived from the automatic turn segmentation boundaries provided by Whisper.
- Word Duration: Estimated using the word-level timestamps generated by the forced alignment module of WhisperX (see Section 1.5.2).

For the nonverbal modality, we focused our analysis on the video call recordings, quantifying the proportion of "gazing at the screen" (as a proxy for looking at the interlocutor) versus "looking away" using the manual annotations described in Section 1.5.3.1. Detailed analysis of gaze behavior in the face-to-face condition is reserved for subsequent chapters.

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

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

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

1.6.2. Results and Discussion

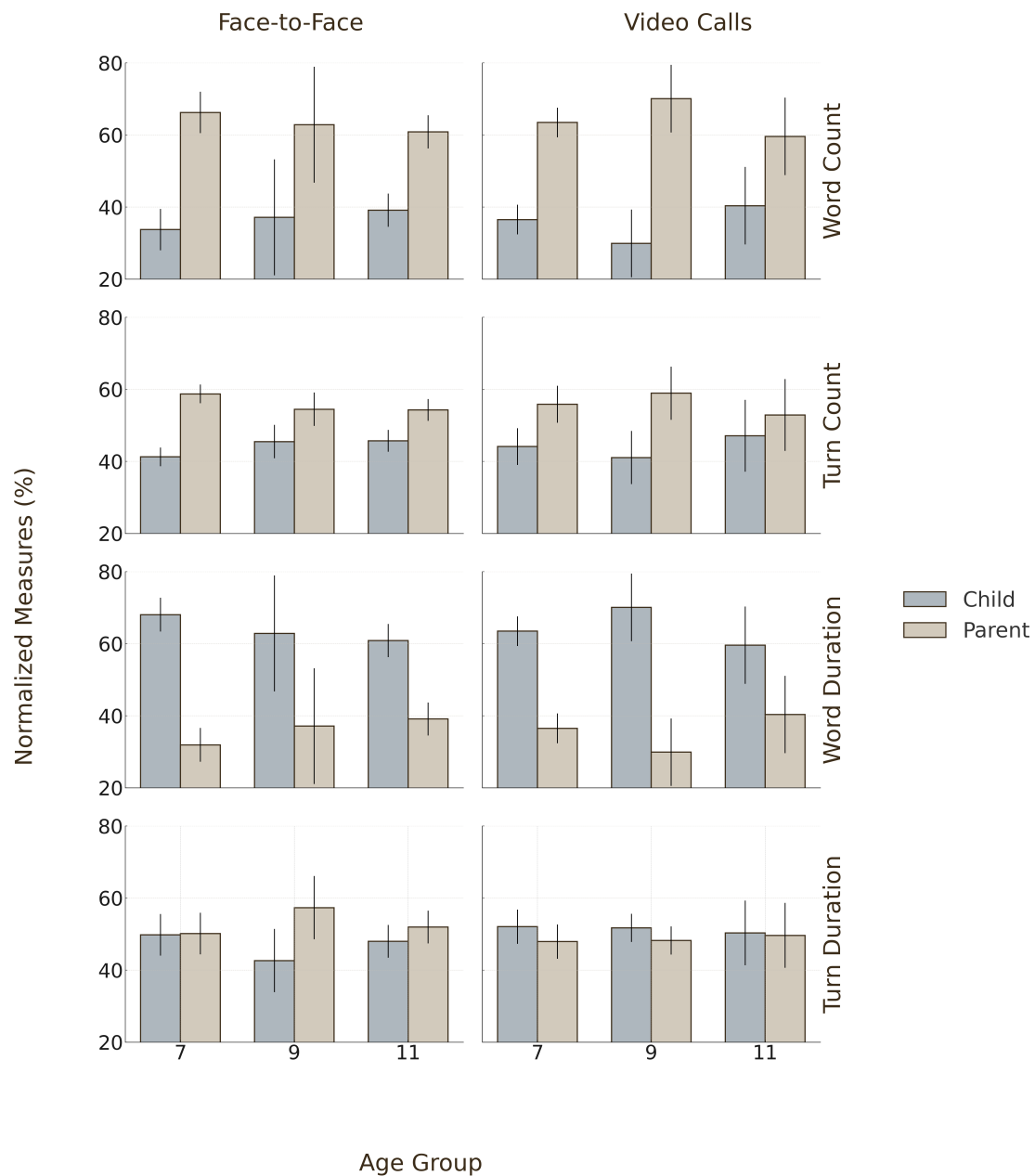


Figure 1.5. – Mean normalized contributions of children and parents to Word Count, Turn Count, Word Duration, and Turn Duration for three age groups (7, 9, 11), shown separately for Face-to-Face and Video-Call modalities. Normalization was performed per dyad by dividing each speaker's measure by the total measure for that dyad. Error bars represent the standard deviation of normalized values across participants.

Figure 1.5 shows the results for the verbal modality. We can make several observations. First, regarding the total number of words uttered in a conversation, children produced much fewer words; almost half the number of words produced by their parents. Second, regarding the total number of turns in a conversation, children also took fewer turns than the parents, although the difference is not as significant as in the case of words. As for the duration of a single turn, children generally took as much time as their parents. Finally, we observe a reliable pattern in which children's word utterances took longer than their parents', suggesting that children generally spoke at a slower rate.

These verbal measures are intrinsically linked. Since children produced significantly fewer words yet maintained turn durations comparable to adults (with only a marginally lower turn count), the mathematical result is a slower speech rate. Essentially, children occupy the floor for the same amount of time but contribute less linguistic content. This aligns with established psycholinguistic models of speech production (e.g., Levelt 1993), which posit that children require longer planning phases for lexical retrieval and syntactic formulation compared to adults. The preservation of turn duration despite lower word counts suggests that children in middle childhood prioritize maintaining the conversational floor even as they navigate higher cognitive loads during production.

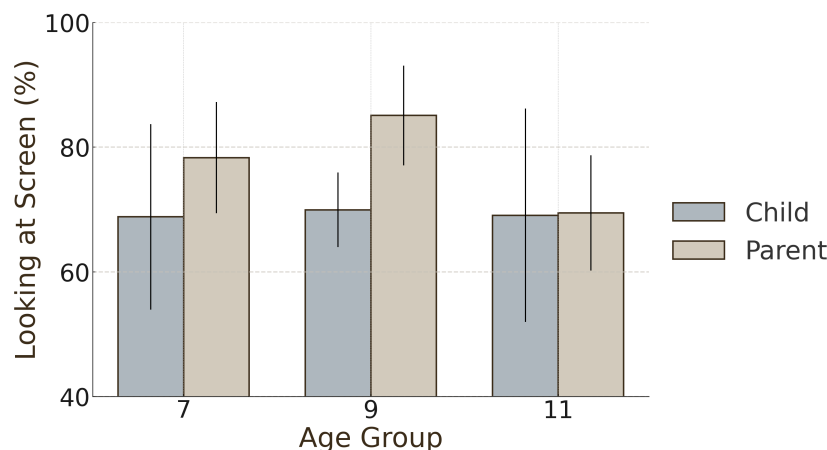


Figure 1.6. – Average percentage of time that children and parents spent looking at the screen during video-call sessions, grouped by child age (7, 9, 11). For each session, looking-at-screen proportions were computed separately for the child and the parent, then averaged within age groups. Error bars represent standard deviations.

Figure 1.6 shows the results for the non-verbal modality in video calls, specifically the proportion of "gazing at the screen" vs. "looking away", a binary variable averaged across parent/child and age groups. Figure 1.6 shows similar behavior in children and adults, with no significant developmental trend observed in this preliminary analysis.

One important observation, for both verbal and non-verbal measures, is the absence of significant developmental change across our age groups for these high-level

measures. This suggests a form of structural stability: the broad parameters of interaction, such as turn dominance and gaze distribution, appear well-established by age 7. This finding is consistent with developmental literature indicating that the fundamental "rules" of turn-taking are acquired in early childhood (Casillas et al. 2016). It reinforces the hypothesis that the developmental work of middle childhood lies not in learning how to take a turn or where to look, but in the qualitative fine-tuning of these behaviors, such as the precise timing of transitions, which these coarse-grained distributional measures are not designed to capture.⁴

Another important observation is the striking similarity in verbal behavior across face-to-face and video-call settings. The patterns regarding word count, turn duration, and speech rate remained consistent across the medium. This result is methodologically significant. While younger children often exhibit a "video deficit" that hampers interaction quality (Anderson and Pempek 2005), our results indicate that by middle childhood, children have adapted sufficiently to computer-mediated communication. This validates the use of video calls in the CHICA corpus as a robust proxy for studying the verbal structure of turn-taking, even if the non-verbal dynamics require more specialized handling.

1.7. Conclusion

This chapter established the CHICA corpus as a methodological answer to a long-standing challenge in developmental science: the tension between ecological validity and experimental precision. By deploying wearable sensors in the home environment, we have created a resource that captures the nuances of middle childhood communication without the distortions of the laboratory.

Beyond the data itself, this work validates a reproducible methodological framework for naturalistic inquiry. We demonstrated that by combining state-of-the-art automated tools (Whisper, OpenFace, Pupil Invisible) with targeted manual intervention, it is possible to navigate the "annotation bottleneck" effectively. This hybrid pipeline proves that high-fidelity multimodal signals can be extracted at scale from unconstrained environments, offering a blueprint for future research that seeks to move beyond the static confines of the recording booth.

The availability of high-fidelity, multimodal data in a naturalistic setting has significant implications for the field of developmental pragmatics. By integrating precise gaze tracking with naturalistic speech, the CHICA corpus enables a shift in analytical focus: from observing discrete linguistic events to understanding the continuous orchestration of interaction. This moves the field beyond static measures of competence toward a dynamic model of communicative performance, allowing researchers to investigate how non-verbal signals function not merely as accompaniments to speech, but as fundamental regulatory mechanisms in the acquisition of adult-like fluency.

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

1. The CHICA Corpus: Ecological Validity in Middle Childhood – 1.7. Conclusion

Finally, our preliminary analyses revealed a remarkable structural stability across ages 7 to 11. The observation that broad interactional parameters, such as turn-taking dominance and gaze distribution, are well-established by age 7 suggests that the developmental work of this period is not about learning what to do, but about refining how to do it. This finding provides the essential baseline for this thesis. With the "macro" structure of conversation proven robust, the following chapters will drill down into the "micro" dynamics, investigating how the precise orchestration of gaze evolves to achieve adult-level fluency.

2. Automatic Annotation of Smiling and Gaze in Unconstrained Video Call Data

Table of contents

2.1. Introduction	51
2.2. Methodological Positioning and Related Work	52
2.2.1. Automatic Detection of Facial Expressions in Video	52
2.2.2. Automatic Gaze Estimation from Video	53
2.2.3. Challenges of Unconstrained and Developmental Video Data	54
2.2.4. Summary and Positioning of this work	54
2.3. Methods	54
2.3.1. Data and Annotation	54
2.3.1.1. Interaction Corpus	54
2.3.1.2. Manual Annotation	55
2.3.2. Pre-processing and Face Extraction	55
2.3.2.1. Face Detection	55
2.3.2.2. Face Cropping and Normalization	56
2.3.2.3. Feature Extraction with OpenFace	56
2.3.3. Modeling Approaches	56
2.3.3.1. Feature-based Models	57
2.3.3.2. End-to-End Models	58
2.3.3.3. Transfer Learning for Smile Detection	58
2.3.3.4. Summary of Modeling Design	58
2.3.4. Experimental Design and Evaluation	59
2.3.4.1. Training and Testing Protocol	59
2.3.4.2. Class Imbalance and Data Availability	59
2.3.4.3. Evaluation Metric	59
2.3.4.4. Aggregation and Interpretation of Predictions	59
2.4. Results	60
2.4.1. Automatic Detection Performance	60
2.4.2. Smile Detection	60
2.4.3. Gaze Detection	61
2.4.4. Reproducing Aggregate Behavioral Patterns	61

2.5. Discussion	62
2.5.1. Smile Detection: Robustness Under Cross-Domain Transfer . . .	63
2.5.2. Gaze Direction: Structural Limits Imposed by Video-Call Geometry	63
2.5.3. Implications for Automatic Annotation in Ecological Interaction Research	64
2.5.4. Methodological Consequences for the Present Thesis	64

2.1. Introduction

The proliferation of digital recording devices has led to an unprecedented availability of video data capturing human interaction in naturalistic settings. From remote education and telehealth to large-scale developmental corpora, researchers now have access to vast amounts of “in-the-wild” video footage. However, this abundance of data has revealed a central methodological bottleneck: while raw recordings are plentiful, extracting meaningful behavioral signals from them remains difficult, time-consuming, and costly.

Traditionally, to annotate behavioral data, researchers have resorted to either highly controlled laboratory settings using specialized sensors or manual frame-by-frame coding by trained annotators. The former prioritizes measurement precision at the cost of ecological validity, while the latter preserves contextual elements of the conversation but severely limits scalability. As a result, we believe that neither approach alone is sufficient for large-scale studies of naturalistic interaction.

Automatic annotation from video offers a potential third path: (1) scalable: automatic annotation makes it possible to work with hours of video, not just small samples. Once a model is trained, it can be applied consistently across large datasets without the time and cost constraints of manual coding. (2) reproducible: The same input processed with the same model yields the same annotations every time. This reduces variability introduced by human annotators and makes the analysis easier to replicate. (3) applicable to ecologically valid data.

Recent advances in computer vision and machine learning have demonstrated impressive performance on tasks such as facial expression recognition and gaze estimation. Yet, these advances have been mainly validated under controlled or semi-controlled conditions. Their robustness in unconstrained, real-world video remains an open empirical question.

This chapter addresses this question by investigating the automatic detection of two foundational social signals (smiling and gaze direction) from unconstrained video recordings of interaction. These signals were selected not only for their central role in social coordination, but because they instantiate two qualitatively different detection problems. Smiling is primarily a facial configuration, determined mainly by muscle activations visible in the face. Gazing toward an interaction partner, on the other hand, is a relational behavior whose interpretation depends on the spatial relationships among the individual, the camera, and the environment.

The goal of this chapter is not to propose a definitive solution for automatic gaze

estimation or affect annotation, but to empirically assess the feasibility and limits of video-based approaches under realistic recording conditions. By comparing feature-based and end-to-end modeling paradigms across these two signal types, this chapter delineates which aspects of social behavior can be robustly extracted from unconstrained video and which remain fundamentally constrained by the absence of explicit spatial information.

2.2. Methodological Positioning and Related Work

Automatic annotation of non-verbal behavior from video is often presented as a promising approach for scalable analysis in computer vision and multimodal interaction research. However, we believe that whether this promise holds in practice actually depends on a set of conditions that are not always made explicit: the properties of the signal to be estimated, the constraints imposed by the recording setup, and the degree of similarity between the data used to train a model and the context in which it is applied. This section approaches the literature with that question in mind, reviewing prior work on automatic facial expression recognition and gaze estimation with particular attention to their applicability to unconstrained, developmental interaction data.

2.2.1. Automatic Detection of Facial Expressions in Video

Automatic facial expression recognition has been extensively studied over the past decades. Early approaches relied on handcrafted features inspired by the Facial Action Coding System (FACS) (Ekman et al. 1978), which decomposes facial expressions into anatomically grounded Action Units corresponding to specific muscle activations. Recently, deep learning approaches (particularly convolutional neural networks (CNNs)) have become dominant. CNNs can learn representations directly from raw pixel data and achieve strong performance on a wide range of facial expression recognition benchmarks (Zhang et al. 2015; Jung et al. 2015; Chen et al. 2016).

Among facial expressions, smiling is often considered one of the most tractable signals for automatic detection. Numerous studies report high accuracy in classifying smiling from neutral facial configurations, primarily in adult populations and under controlled or semi-controlled conditions (Cui et al. 2013; Chen et al. 2016; Li et al. 2020). These results are commonly interpreted as evidence that smiling can be reliably captured by local facial features, such as lip-corner movements and cheek raising.

Toolkits such as OpenFace (Baltrušaitis et al. 2016) operationalize FACS by providing frame-level estimates of Action Unit activation and related facial features, and have been widely adopted in behavioral research due to their interpretability and ease of integration into analysis pipelines. By explicitly encoding facial muscle configurations, feature-based representations reduce the dimensionality of the learning problem. For that reason, they usually can generalize well when training data are limited (an essential consideration for developmental datasets).

However, it is essential to note that much of the reported success in smile detection relies on training and evaluation protocols that are tightly coupled to specific datasets. Widely used benchmarks such as SMILES (Hromada et al. 2010) or GENKI (MPLab n.d.) primarily consist of curated still images, featuring adult faces or public figures captured under favorable conditions. Models trained and evaluated on such datasets may achieve high performance by exploiting dataset-specific regularities that do not necessarily transfer beyond the benchmark.

In particular, the generalization of smile detection models from static images to video data collected in unconstrained environments remains underexplored. Video recordings introduce additional sources of variability, including motion blur, compression artifacts, and partial occlusion. Moreover, we argue that in naturalistic data, facial expressions are often subtle and transient, rather than posed or exaggerated for the camera. As a result, strong performance on image-based benchmarks does not guarantee robustness when the same models are applied to “in the wild” video.

2.2.2. Automatic Gaze Estimation from Video

Automatic gaze estimation from video raises a qualitatively different set of challenges. A large body of work has focused on estimating gaze direction from monocular cameras, often under constrained conditions with fixed camera placement, controlled lighting, and known geometry of the scene. Recent surveys document substantial progress in gaze estimation using deep learning, while also emphasizing the strong dependence on calibration assumptions (Cheng et al. 2021; Ghosh et al. 2022).

While such methods can achieve high accuracy in laboratory settings, their applicability to naturalistic interaction data is limited. In unconstrained recordings, camera position, screen size, distance from the camera, and head pose vary widely across participants. Consequently, identical gaze vectors may correspond to different functional interpretations depending on the recording configuration. This ambiguity is particularly problematic when the goal is not to estimate gaze direction per se, but to classify the gaze vector into categories such as gaze toward an interaction partner versus gaze aversion, or looking at the screen versus looking away.

From an interactional perspective, gaze is inherently relational. Classic work in social interaction has emphasized its role in regulating attention, turn-taking, and interpersonal coordination, rather than treating it as a purely geometric signal (Kendon 1967; Duncan and Fiske 2015). More recent theoretical accounts similarly situate gaze within the multimodal coordination systems that depend on shared spatial and contextual information (Hamilton et al. 2023). Feature-based gaze representations that abstract away from this relational structure with the environment, may therefore fail.

Very few studies have addressed the automatic classification of partner-directed gaze in conversation. Closest existing work has focused on infants’ looking toward predefined stimuli in semi-controlled experimental setups (Erel et al. 2023). While informative, such paradigms limit the room for variability in the experimental setup.

2.2.3. Challenges of Unconstrained and Developmental Video Data

The challenges associated with automatic annotation are further amplified in unconstrained, real-world video recordings. Such data typically exhibit high variability in lighting conditions, background clutter, camera angles, image resolution, motion blur, occlusions, and abrupt head movements.

Developmental interaction data introduce additional sources of variability. Children’s facial morphology differs from that of adults. Recording conditions in child-centered studies are also less standardized, as children may not consistently remain within the camera’s field of view or maintain a stable orientation toward the recording device. These factors raise concerns about the generalization of models trained predominantly on adult datasets (Xia et al. 2017).

Recent multimodal corpora of child–caregiver interaction have emphasized the importance of capturing behavior in ecologically valid contexts while acknowledging the methodological challenges this poses for automatic analysis (Bodur, Nikolaus, Kassim, et al. 2021; Bodur, Nikolaus, Prévot, et al. 2023). Rather than treating this variability as noise to be eliminated, the present work treats it as an intrinsic property of the nature of data. Thus, we evaluate automatic annotation methods under these conditions.

2.2.4. Summary and Positioning of this work

Building on these insights, the present chapter adopts a comparative methodological approach, to evaluate feature-based and end-to-end models for smiling and gaze detection in naturalistic video data. Rather than aiming to optimize performance for a specific application, the goal is to empirically delineate the boundaries of what can be reliably inferred from unconstrained video alone, thereby informing methodological choices in subsequent chapters.

2.3. Methods

We adopt a comparative experimental approach to evaluate the feasibility of automatically detecting smiling and gaze behavior from unconstrained interaction video. Rather than proposing a single optimized pipeline, the methods are designed to contrast two broad modeling paradigms: feature-based and end-to-end approaches, across two signal types that are fundamentally different.

2.3.1. Data and Annotation

2.3.1.1. Interaction Corpus

The experiments reported in this work are based on the CHICO corpus (Bodur, Nikolaus, Kassim, et al. 2021), a multimodal dataset of child–caregiver interactions

collected to study conversational development in middle childhood. We used a subset that comprises eight dyads of school-age children (6–12 years old) interacting with a caregiver during a remote conversational task. Each interaction lasted between 10 and 15 minutes and followed a weakly structured guessing game designed to elicit balanced participation from the child.

At this point, one might reasonably ask why this study relies on CHICO rather than on CHICA, which was introduced in the previous chapter. The short answer is that the two corpora serve different purposes within the thesis. CHICO is an earlier corpus that offers more extensive, carefully curated annotations, including full manual gaze annotation. For this purpose, CHICO provides a more suitable testbed, while later chapters return to CHICA to examine gaze behavior in fully ecological interaction.

The recordings were collected using consumer-grade devices (usually a PC or a tablet) in participants’ home environments and mediated by standard video communication software. As a result, the corpus exhibits many characteristics of unconstrained video data, including heterogeneous camera placement, variable lighting conditions, varying image resolutions, and frequent head and body movement. These properties make CHICO a stringent testbed for evaluating the robustness of automatic annotation methods.

The CHICO corpus was manually annotated for a range of non-verbal behaviors, including facial expressions and gaze direction. These annotations serve as the ground truth for evaluating the automatic methods examined in this work.

2.3.1.2. Manual Annotation

All video recordings were manually annotated at the frame level by trained annotators. Two behavioral categories were annotated:

- Smiling, coded as smile vs. no-smile, based on visible facial expression.
- Gaze direction, coded as looking at the screen vs. looking away.

For gaze, partner-directed gaze was operationalized as gaze directed toward the screen. We do acknowledge that this operationalization constitutes an approximation rather than a direct equivalent of face-to-face gaze-on-face. Inter-annotator agreement was assessed in the original corpus documentation.

2.3.2. Pre-processing and Face Extraction

Automatic detection of smiling and gaze direction from unconstrained video critically depends on the reliability of the upstream pre-processing steps, particularly face detection and normalization.

2.3.2.1. Face Detection

Face detection was performed frame by frame using RetinaFace (Deng et al. 2020), a state-of-the-art single-stage face detector that performs robustly under challenging conditions. We have entertained different face detection models (ranging from

classical Haar cascade-based detectors and HOG-SVM approaches to more recent CNN-based detectors), but RetinaFace was selected over other detectors because of its demonstrated robustness in “in-the-wild” scenarios that closely resemble the recording conditions of the present corpus.

Across the CHICO recordings, RetinaFace successfully detected faces in the vast majority of frames. Detection failures occurred primarily in cases of extreme head rotation, strong motion blur, or when participants temporarily moved out of the camera’s field of view. Frames in which no face was detected were excluded from analyses.

2.3.2.2. Face Cropping and Normalization

For each detected face, a bounding box was extracted and used to crop the face region from the frame. Cropped faces were converted to grayscale and resized to a fixed resolution of 64×64 pixels.

The choice of a fixed, low-resolution input was motivated by two considerations. First, it ensured consistent input dimensions across all models. Second, it reflects realistic constraints encountered in large-scale behavioral corpora, where high-resolution facial imagery is often unavailable due to camera grade and placement.

2.3.2.3. Feature Extraction with OpenFace

For feature-based models, facial and gaze-related features were extracted using OpenFace 2.0, a widely used open-source toolkit for facial behavior analysis (Baltrušaitis et al. 2016). OpenFace provides frame-level estimates of facial Action Unit, eye-gaze direction, and head-pose parameters.

OpenFace was applied to the pre-processed face images on a frame-by-frame basis. For smile detection, Action Units associated with lip corner pulling and cheek raising were selected. For gaze detection, eye-gaze direction estimates for both eyes, along with head-pose parameters, were extracted and used jointly as predictors.

OpenFace successfully produced feature estimates for the majority of detected faces. However, feature extraction failures occurred in a non-negligible subset of frames, particularly under low illumination, partial occlusion, or extreme facial orientation. These failures resulted in the exclusion of additional frames from feature-based analyses.

Table 2.1 summarizes the distribution of frames across smile and gaze categories, as well as the number of frames after face detection and feature extraction. These differences define an upper bound on the performance achievable by feature-based models and must be taken into account when interpreting classification results.

2.3.3. Modeling Approaches

To assess the feasibility of automatically detecting smiling and gaze direction from unconstrained interaction video call data, two broad modeling paradigms were evaluated: feature-based models that rely on explicit facial and gaze-related descriptors,

Table 2.1. – Distribution of video frames in the ChiCo corpus across behavioral categories (smile and gaze) and preprocessing stages. The table reports the number and proportion of frames in the original recordings, after face detection with RetinaFace, and after feature extraction with OpenFace. Differences across stages reflect detection and feature extraction failures.

Behavior	Original	RetinaFace	OpenFace
Smile	51,909 (12.00%)	51,418 (12.02%)	51,324 (12.28%)
No smile	380,523 (88.00%)	376,225 (87.98%)	366,574 (87.72%)
Gaze	314,297 (72.68%)	312,173 (73.00%)	304,254 (72.81%)
No gaze	118,135 (27.32%)	115,470 (27.00%)	113,644 (27.19%)
Total	432,432	427,643	417,898

and end-to-end models that learn directly from raw image data. This comparative design was motivated by the distinct assumptions these paradigms make about representation, generalization, and robustness under domain shift.

Notably, the goal was not to optimize performance for a specific application but to evaluate how each modeling strategy behaves under the constraints of naturalistic, developmental video call data.

2.3.3.1. Feature-based Models

Feature-based models first extract interpretable descriptors from video frames and then learn a mapping from these descriptors to behavioral categories. This approach has two key advantages: it reduces the dimensionality of the learning problem. It provides a degree of interpretability, allowing performance limitations to be traced back to specific feature representations.

For smile detection, feature-based models relied on facial Action Units extracted using OpenFace. These features were selected to reflect facial muscle activations rather than holistic facial appearance, thereby abstracting away from contextual variability in lighting or background.

For gaze direction detection, feature-based models used a combination of eye-gaze direction estimates and head-pose parameters. These features approximate where an individual is looking and are commonly used in gaze estimation pipelines. However, as discussed in the previous section, their interpretability depends critically on assumptions about camera geometry that are violated in unconstrained recordings.

In both cases, extracted feature vectors were used to train Support Vector Machine (SVM) classifiers. Separate classifiers were trained for smiling and gaze detection.

Feature-based models were trained and evaluated only on frames for which feature extraction succeeded. As shown in Table 2.1, feature availability differed across behaviors and preprocessing stages, particularly for gaze-related features.

2.3.3.2. End-to-End Models

End-to-end models bypass explicit feature extraction and instead learn a direct mapping from raw image data to behavioral labels. This approach allows models to exploit any visual cues present in the input, including contextual information that handcrafted features may not capture.

End-to-end models were implemented using Convolutional Neural Networks (CNNs) that take as input a normalized grayscale face image, as described in Section 2.3.2. All models shared the same architecture, consisting of three convolutional layers, each followed by a max-pooling operation, and a fully connected layer with 128 units, followed by a sigmoid output. We chose this architecture as a compromise between capacity and the limited dataset size.

Models were trained using the Adam optimizer with a binary cross-entropy loss function. Training was performed for a fixed number of epochs using mini-batches of equal size. The same architecture and training protocol were used for both smile and gaze detection.

Unlike feature-based models, end-to-end models were trained on all frames for which a face was detected, regardless of whether OpenFace successfully extracted features or not. This distinction is crucial for gaze detection, as it enables end-to-end models to leverage visual information even when explicit gaze estimates fail.

2.3.3.3. Transfer Learning for Smile Detection

For smile detection, we additionally explored the effect of domain shift by using a publicly available image-based smile dataset consisting of curated still images. Three experimental regimes were considered:

- Within-domain training, where models are trained and tested exclusively on the interaction corpus.
- Direct transfer, where models are trained on the external image dataset and tested directly on the interaction corpus.
- Fine-tuning, where models are first trained on the external dataset and then fine-tuned on a subset of interaction data.

These scenarios were designed to evaluate whether performance achieved on curated image datasets carries over to unconstrained video data, and whether limited exposure to in-domain data can mitigate domain mismatch. Transfer learning was applied only to smile detection, as no comparable large-scale annotated dataset exists for gaze.

2.3.3.4. Summary of Modeling Design

Together, these modeling approaches allow for a principled comparison between explicit, interpretable representations and end-to-end learning under realistic recording conditions.

2.3.4. Experimental Design and Evaluation

This experimental design aims to evaluate the robustness and generalization of automatic annotation methods under realistic conditions, rather than to maximize performance within a single dataset. To this end, we present here our choices on model training and evaluation.

2.3.4.1. Training and Testing Protocol

All models were evaluated using a leave-one-participant-out cross-validation scheme. In each fold, data from one participant (child or caregiver) is held out as the test set, while models are trained on the remaining participants' data. This procedure ensures that reported performance reflects generalization to unseen individuals.

For smile detection, additional experiments involved training on an external dataset and testing on the CHICO corpus, with and without fine-tuning. In these cases, participant-wise separation was maintained during fine-tuning.

For gaze detection, all training and testing were performed exclusively on the corpus, as no comparable external dataset annotated for partner-directed gaze in conversational contexts is currently available.

2.3.4.2. Class Imbalance and Data Availability

Both smiling and gaze behaviors exhibited substantial class imbalance, as shown in Table 2.1. In particular, smiling constituted a relatively small proportion of frames compared to neutral facial configurations, while gaze behavior was dominated by looking at the screen class.

2.3.4.3. Evaluation Metric

Model performance is evaluated using balanced accuracy, defined as the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate). We chose balanced accuracy because it provides an interpretable performance measure under class imbalance and avoids inflated scores arising from majority-class predictions.

2.3.4.4. Aggregation and Interpretation of Predictions

In addition to frame-level classification performance, the outputs of the best-performing models were aggregated to estimate higher-level behavioral measures, such as the proportion of time participants spent smiling or averting gaze during an interaction. These derived estimates were compared with manual annotations to assess whether model predictions preserve meaningful behavioral patterns when aggregated.

This analysis was not intended to validate automatic annotation as a replacement for manual coding at the individual level, but rather to evaluate its potential utility for large-scale, aggregate analyses, where fine-grained precision may be less critical than overall trends.

2.4. Results

This section reports the performance of feature-based and end-to-end models for the automatic detection of smiling and gaze. Results are presented separately for the two behavioral signals.

All reported performance metrics reflect generalization to unseen participants, using balanced accuracy as defined in Section 2.3.4.3.

2.4.1. Automatic Detection Performance

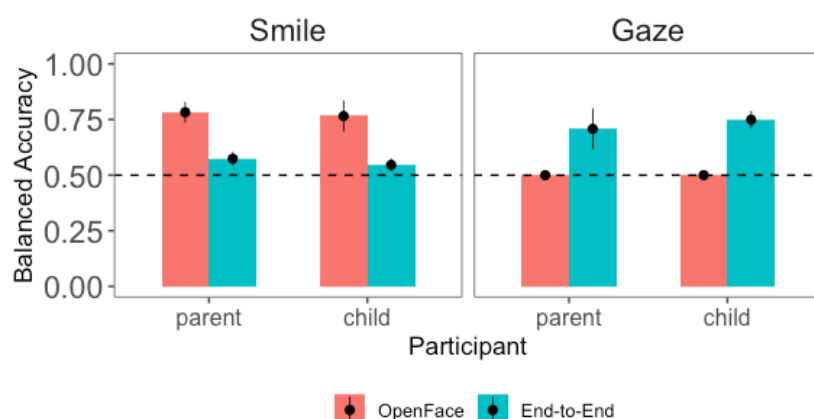


Figure 2.1. – The (balanced) accuracy score for Smile and Gaze comparing Feature-based models to End-to-End models. Ranges indicate 95% confidence interval over 8 unique participants in the CHICO corpus, both parents and children. Scores for each individual were obtained using models that had not seen that individual during training. The dashed lines represent chance.

Figure 2.1 summarizes the balanced accuracy scores for smiling and gaze detection, comparing feature-based and end-to-end modeling approaches. Scores are averaged across participants, with error bars indicating 95% confidence intervals. Dashed lines indicate chance-level performance.

2.4.2. Smile Detection

The smile detection results reported here correspond to a specific transfer-learning scenario in which models were trained on an external image-based smile dataset (SMILE) and tested directly on the CHICO interaction corpus, without fine-tuning on in-domain video data. This scenario was selected because within-domain training on CHICO alone yielded poor generalization due to severe class imbalance and limited diversity.

For smile detection, feature-based models relying on Action Unit representations substantially outperformed end-to-end convolutional models. This pattern was consistent across participants and was observed despite the strong performance of end-to-end models when trained and evaluated on curated image-based datasets.

In contrast, end-to-end models exhibited poor generalization to unconstrained video. Although the same architecture achieved high balanced accuracy when evaluated within the image-based dataset, this performance did not transfer to data involving children and caregivers. Fine-tuning on a subset of in-domain data did not yield reliable performance improvements either.

Overall, these results indicate that feature-based representations generalize more robustly than end-to-end models for smile detection under domain shift from curated images to naturalistic video.

2.4.3. Gaze Detection

A reversed pattern is observed for gaze direction detection. Feature-based models relying on explicit gaze and head pose estimates performed at or near chance level, indicating a failure to distinguish between screen-directed and averted gaze.

End-to-end models achieved balanced accuracy above chance for gaze detection and consistently outperformed feature-based approaches. Importantly, this advantage should not be interpreted as evidence that end-to-end models recover gaze direction *per se*. Instead, it reflects their ability to partially adapt to variability in camera placement relative to the face and the screen, which differs substantially across participants in the corpus recordings.

Feature-based gaze representations assume a stable mapping between eye or head orientation and the target of gaze. In the present data, this assumption is violated because the webcam position varies across setups. End-to-end models can implicitly learn participant-specific visual configurations associated with screen-directed gaze, yielding above-chance performance despite the absence of explicit spatial information.

2.4.4. Reproducing Aggregate Behavioral Patterns

Beyond frame-level classification accuracy, we evaluated whether automatic predictions could reproduce aggregate behavioral patterns derived from manual annotation. Using the best-performing model for each behavior, frame-level predictions were aggregated to estimate the proportion of time participants spent smiling or averting gaze during each interaction.

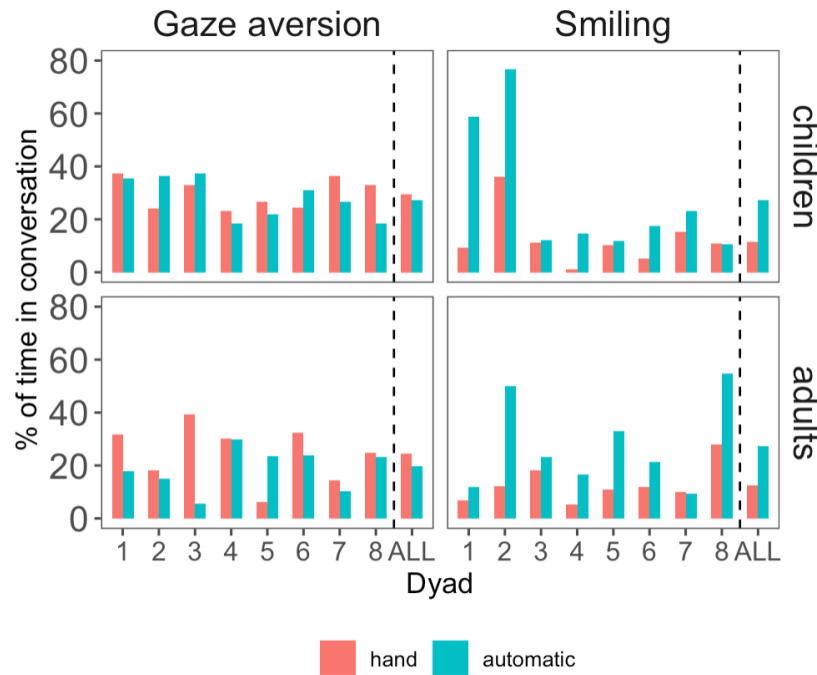


Figure 2.2. – The proportion of frames in a video where children (top) or adults (bottom) were smiling (right) or/and averting gaze (left). We show the numbers obtained for each of the 8 dyads in CHICO and the average ('ALL'), using both hand-coded and automatically coded frames.

Figure 2.2 compares these automatically derived estimates to those obtained from manual annotations, both at the level of individual dyads and when aggregated across all interactions. At the individual level, discrepancies between automatic and manual coding varied substantially across dyads.

However, when data were aggregated across participants, automatic coding reproduced overall trends observed in the manual annotations, particularly for gaze behavior. Although smile frequency was systematically overestimated, the relative patterns across participants were preserved.

These results suggest that, despite limitations at the granular level, automatic annotation can approximate aggregate behavioral measures when applied cautiously and interpreted at an appropriate level of analysis.

2.5. Discussion

This chapter investigated the feasibility of automatically detecting smiling and gaze direction from video-call recordings of child–caregiver interaction. By comparing feature-based and end-to-end modeling approaches across two distinct behavioral signals, the goal was not to optimize annotation performance, but to empirically delineate the limits of what can be reliably inferred from video-mediated interaction data collected under ecologically valid conditions.

The results reveal a principled dissociation between smiling and gaze detection that reflects fundamental differences in how these signals are expressed, represented, and constrained by the geometry of video calls.

2.5.1. Smile Detection: Robustness Under Cross-Domain Transfer

For smile detection, feature-based models that rely on Action Unit representations consistently outperformed end-to-end models on the ChiCo corpus. Notably, the results reported in this chapter correspond to a specific and stringent evaluation scenario in which models were trained on an external, image-based smile dataset (SMILE) and tested directly on ChiCo, without fine-tuning on in-domain video data.

By explicitly encoding facial muscle activations, Action Unit-based features abstract away from many sources of contextual variability introduced by video calls, such as differences in lighting, background, image compression, and camera quality.

In contrast, end-to-end models trained on curated image datasets appear to rely, at least in part, on dataset-specific regularities that do not transfer reliably to continuous, naturalistic video recordings. Although such models achieve strong performance when training and evaluation conditions are closely matched, this performance does not carry over to new data. These findings underscore a broader methodological concern: high benchmark performance on static image datasets does not guarantee robustness when models are applied to unconstrained video data.

2.5.2. Gaze Direction: Structural Limits Imposed by Video-Call Geometry

The results for gaze direction detection present a fundamentally different picture. Feature-based models relying on explicit gaze-direction and head-pose estimates performed at or near-chance levels. End-to-end models consistently outperformed feature-based approaches and achieved balanced accuracy above chance, but performance remained moderate and highly variable across participants.

The results reflect a structural property of video-call geometry. In video calls, the webcam is spatially separated from the interlocutor's face displayed on the screen, and its position varies substantially across devices and user setups (e.g., laptop bezel, external webcam, screen tilt). As a result, "looking at the screen" does not correspond to a single, stable configuration of eye or head orientation across participants.

End-to-end models can partially adapt to this variability by implicitly learning participant-specific visual configurations associated with screen-directed gaze, thereby achieving above-chance performance. However, performance remains limited and variable, revealing a ceiling that cannot be overcome through architectural refinements alone.

2.5.3. Implications for Automatic Annotation in Ecological Interaction Research

The findings of this study demonstrate that the feasibility of automatic annotation from webcam-based video-call recordings depends critically on the nature of the behavioral signal under investigation. Signals intrinsic to facial morphology, such as smiling, are amenable to scalable automation using feature-based representations, even under substantial domain shifts. In contrast, signals that are inherently relational and spatially defined, such as gaze, are on a higher level of difficulty when it comes to automatic annotation.

Taken together, I argue that current automatic annotation tools should not be treated as universally applicable solutions. Inference on new corpus data should be done with caution. Performance on public benchmarks provides useful information, but it does not by itself establish reliability in unconstrained video.

2.5.4. Methodological Consequences for the Present Thesis

The empirical limits identified in this chapter directly motivate the methodological choices adopted in the remainder of the thesis. We do believe that automatic annotation from video-call recordings provides a viable approach for certain facial behaviors. However, it does not offer sufficient reliability for studying fine-grained gaze dynamics in interaction.

For that reason, subsequent lines of work shift away from fully automated gaze inference in first-person videos and instead rely on eye-tracking data to capture gaze behavior with higher spatial and temporal fidelity. This methodological pivot is not driven by convenience, but by the structural constraints revealed here. We find it necessary that we have established these constraints here to ensure that later analyses rest on measurements appropriate to our theoretical questions.

3. Gaze on Face in Middle Childhood: An Emergent Adult-like Preference

Table of contents

3.1. Introduction: The "Gaze-on-Face" Phenomenon	65
3.2. The Literature and the Middle Childhood Gap	66
3.3. Quantifying "Gaze-on-Face"	67
3.3.1. Estimating "Gaze-on-Face"	68
3.3.1.1. Distance Calculation	69
3.3.1.2. Distribution Analysis	69
3.3.1.3. Gaze Aversion Threshold Heuristic	70
3.4. A Preference for "Gaze-on-Face"?	71
3.5. Results: A Robust and Early Adult-like Preference	72
3.5.1. Analysis 1: The Spatial Distribution of Gaze	72
3.5.2. Analysis 2: The Proportion of "Gaze on Face"	74
3.6. Discussion: Re-evaluating the Developmental Trajectory	75

3.1. Introduction: The "Gaze-on-Face" Phenomenon

Having established the methodological foundations of the CHICA corpus, we now turn to our first substantive analysis. This chapter focuses on face-to-face data and investigates the most fundamental component of gaze in conversation: the tendency to look at the interlocutor's face.

Gaze in social interaction is understood to serve a dual function: information acquisition and information signaling (Argyle and Cook 1976; Gobel et al. 2015; Risko et al. 2016; Hessels 2020). We look at our partner's face to acquire crucial information, such as monitoring their comprehension and reactions to adapt our communication accordingly (e.g., Clark and Krych 2004, or enhancing speech comprehension, especially in noisy environments (Schwartz et al. 2004). Simultaneously, our gaze (or lack thereof) acts as a powerful signal. Looking at a partner can signal attention and engagement, while averting gaze is a key component in managing conversational

3. Gaze on Face in Middle Childhood: An Emergent Adult-like Preference – 3.2. The Literature and the Middle Childhood Gap

dynamics, such as coordinating turn-taking (e.g., Kendon 1967; Duncan and Fiske 2015; Ho et al. 2015).

Both of these functions are overwhelmingly concentrated on the interlocutor's face. The face is the primary source of visual and social information, and the target of most signaling behaviors. Therefore, quantifying the development of "gaze-on-face" is the foundational first step in understanding gaze dynamics in conversation.

3.2. The Literature and the Middle Childhood Gap

While we believe the existing literature on gaze preference is extensive, it does present two significant gaps that this thesis aims to address.

First, a large body of research, traditionally conducted in laboratory settings, has established that humans have a strong preference for gazing at faces over other objects (Yarbus 1967). This preference is a cornerstone of human social bias and emerges in early infancy (Johnson et al. 1991). However, we argue that findings from these non-interactive settings (e.g., looking at static images or videos) cannot be directly translated to real-world, face-to-face conversations. In natural scenarios, the "information signaling" function of gaze becomes active; the awareness that one is being watched fundamentally alters gaze behavior (Gobel et al. 2015; Laidlaw et al. 2011).

Second, developmental research on gaze has overwhelmingly focused on infancy and early childhood (e.g., Carpenter et al. 1998; Brooks et al. 2005; Schroer et al. 2023). Their work has provided invaluable insights into the role of mutual gaze and gaze-following in early language acquisition, including the use of a caregiver's gaze to identify word referents. However, we believe that this focus on gaze as a learning tool does not align directly with its functions in mature, adult-like conversation, where it is primarily used for interaction management (e.g., turn-taking, feedback monitoring). This leaves a significant gap in our understanding of how gaze, as an interactional skill, develops after infancy and through middle childhood.

A relevant study is that of De Lillo et al. 2021, which used mobile eye-tracking to examine gaze in naturalistic face-to-face conversations. They reported a surprising finding: adolescents (ages 10-19) paid less attention to their interlocutor's face than young adults. This finding is complex to interpret. As the authors themselves suggest, this result may not solely reflect a developmental difference in gaze preference. The social context of the study, where participants conversed with an adult experimenter (a stranger) in a lab setting, is a big factor. As a matter of fact, it is well-established that gaze behavior is highly modulated by factors such as social rank, familiarity, and social anxiety Kleinke 1986; Gobel et al. 2015. It is therefore within reason to interpret the reported "dip" in gaze as a context-dependent social response to an unfamiliar social situation, rather than a fundamental developmental immaturity. This particular ambiguity highlights the importance of considering social context when studying gaze development.

3.3. Quantifying "Gaze-on-Face"

The face-to-face data collection for the CHICA corpus was explicitly designed to address this confound. As detailed in Chapter 1, by observing children in middle childhood (ages 7, 9, and 11) interacting in a highly familiar, ecologically valid setting at home, playing a simple game with their own parents, we aimed to minimize the social and contextual pressures that might suppress natural gaze behavior. This design allows us to isolate, as much as possible, the underlying developmental patterns of gaze.

While recording, the Pupil Invisible Companion device calculates gaze data in real-time at 120 frames per second. After a recording is uploaded to Pupil Cloud, gaze data is automatically re-computed at the full 200 Hz framerate. Gaze data is output in the scene camera's pixel space, with a resolution of $1088 \times 1080px$. The origin is in the top-left corner of the picture. The gaze estimation algorithm is based on end-to-end deep learning and provides robust gaze data without calibration. A high-level description and a thorough evaluation of the algorithm's accuracy are provided in Tonsen et al. 2020. Fixations and saccades are calculated automatically in Pupil Cloud after uploading a recording. During fixations, the eyes are directed at a specific point in the scene. A saccade is a rapid movement where the gaze moves from one fixation to the next. The fixations are identified using a velocity threshold algorithm, which classifies all samples with gaze velocities below a given threshold as fixations. However, for our studies, we use the raw gaze coordinates mapped on the scene image.

A primary technical challenge in this chapter is the robust quantification of "gaze-on-face" from the Pupil Invisible raw gaze data. While the manufacturer provides a proprietary estimate, preliminary analyses indicate it is highly unreliable for our data, yielding implausibly low results (e.g., 2% gaze-on-face in some recordings). Upon investigation, we identified two primary vectors of errors:

- Proprietary Face Detection vs. Post-Hoc Analysis Needs: The Pupil Invisible system is engineered for real-time processing on a mobile device, and its built-in face detector is optimized for speed. While this approach is practical for immediate, in-field applications, it involves an inherent trade-off between speed and robustness. For this thesis, however, real-time analysis is not required; high-fidelity data is definitely the priority. We found that the proprietary detector was not sufficiently robust to handle the high variability in our naturalistic, in-home data, including non-frontal head poses, partial occlusions, and sub-optimal lighting conditions. This led to an unacceptable rate of detection failures, which would have systematically biased the system's final gaze-on-face estimations.
- Systematic Gaze Offset: A more significant problem arises from the device's "calibration-free" nature. While this feature enhances usability, it can, when combined with slight differences in how each participant wears the glasses, result in a systematic, constant offset in the reported gaze coordinates (e.g., the gaze point may be consistently reported 5 degrees to the left of the actual target).

3.3.1. Estimating "Gaze-on-Face"

To mitigate these problems, we developed a two-step gaze processing pipeline.

First, to address the face detection issue, we replaced the proprietary detector with the state-of-the-art RetinaFace algorithm (Deng et al. 2020). This provided a highly accurate bounding box for the interlocutor's face in each participant's egocentric video feed. This algorithm was applied to each frame of the 30Hz egocentric video. For each frame, RetinaFace outputs a list of all detected face candidates, each with a corresponding bounding box ($[[x_{min}, y_{min}, x_{max}, y_{max}]]$), a confidence score, and an ID to track the face throughout the session. This presented a new challenge: in our "in-the-wild" recordings, the egocentric camera often captured multiple faces (see Figure 3.1), including siblings in the background or even faces in photographs on the wall. To isolate the actual interlocutor, we applied a two-part heuristic: (1) we filtered out all detections with a confidence score below a high threshold (below 0.66), and (2) from the remaining high-confidence candidates, we selected the one with the largest bounding box area. This heuristic is based on the reliable assumption that in a seated, dyadic conversation, the interlocutor's face will always be the largest and most prominent in the field of view. This process yielded a robust, frame-by-frame time series of the interlocutor's face location, thereby solving the detection problem.



Figure 3.1. – Illustration of the RetinaFace detection and interlocutor selection process in a single egocentric frame from the Pupil Invisible glasses. The algorithm detects two faces (green bounding boxes), but only the larger one (on the right) is selected as the interlocutor based on our heuristic. The red circle indicates the most recent fixation event, and the blue circles represent previous fixations with IDs: 1149 and 1150, respectively. (Note: Faces are redacted for privacy.)

Second, simply checking if the gaze coordinate falls within this new bounding box

3. Gaze on Face in Middle Childhood: An Emergent Adult-like Preference – 3.3. Quantifying "Gaze-on-Face"

was insufficient, as it did not solve the offset problem. A systematic offset could mean the reported gaze point never falls within the true face box, even when the participant is looking directly at it. To solve this, we developed a novel heuristic based on the distribution of gaze coordinates relative to the face.

3.3.1.1. Distance Calculation

For every frame, we computed the Euclidean distance between the reported gaze coordinate and the center of the RetinaFace-detected bounding box. We measured face-directed gaze by calculating the distance between the gaze coordinates and the center of the bounding box of the interlocutor's face (the output of RetinaFace), as follows. For each time frame t , the gaze is characterized with two projected coordinates $(x_G(t), y_G(t))$ (G for gaze), indicating the point in the scene the participant is looking at. The face was also detected dynamically: For each time frame t , RetinaFace outputs a face bounding box whose center is characterized with the coordinates $(x_F(t), y_F(t))$ (F for face). Using these two dynamic points, we characterize the spatial location of gaze relative to the interlocutor's face using the Euclidean Distance:

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

3.3.1.2. Distribution Analysis

We then generated a histogram of these distances for the participant's entire recording. This distribution shows a sharp, dominant peak at smaller distances (representing the cluster of gaze points on or near the face) and subsequent, smaller peaks representing gaze at other objects in the scene.

3. Gaze on Face in Middle Childhood: An Emergent Adult-like Preference – 3.3. Quantifying "Gaze-on-Face"

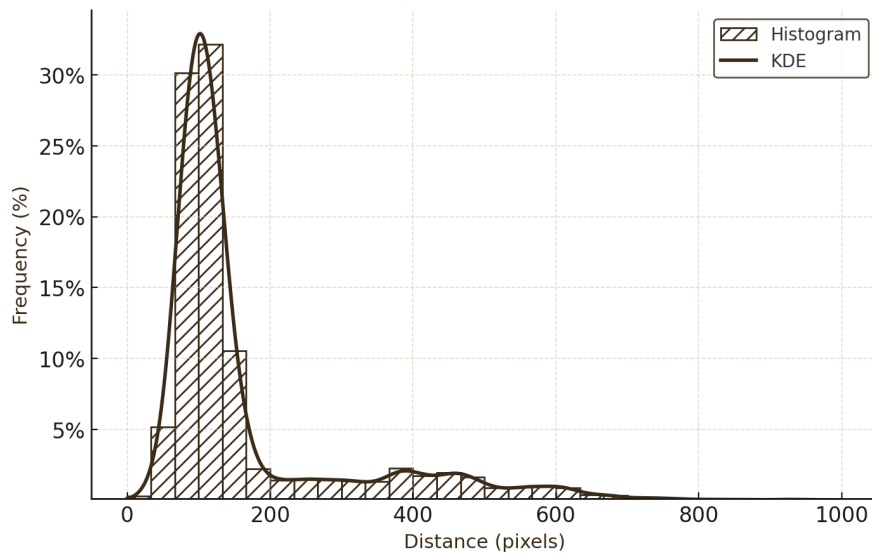


Figure 3.2. – Histogram of gaze distances to the face center for a single participant during a session. The striped bars represent the relative frequency (in %) of gaze distances within 0–1000 pixels, while the smooth curve indicates the Kernel Density Estimation (KDE) of the distribution.

3.3.1.3. Gaze Aversion Threshold Heuristic

We leveraged the fact that the boundary between the "on-face" and "off-face" gaze clusters is visible as a dip in the distribution. We localized this boundary as the first local minimum in the histogram immediately following the dominant "face" peak. This local minimum was used as a participant-specific threshold. Any frame where the gaze distance was smaller than this threshold was classified as Gaze-on-Face (GOF), and any frame greater than it was classified as Averted Gaze (AG).

3. Gaze on Face in Middle Childhood: An Emergent Adult-like Preference – 3.4. A Preference for "Gaze-on-Face"?

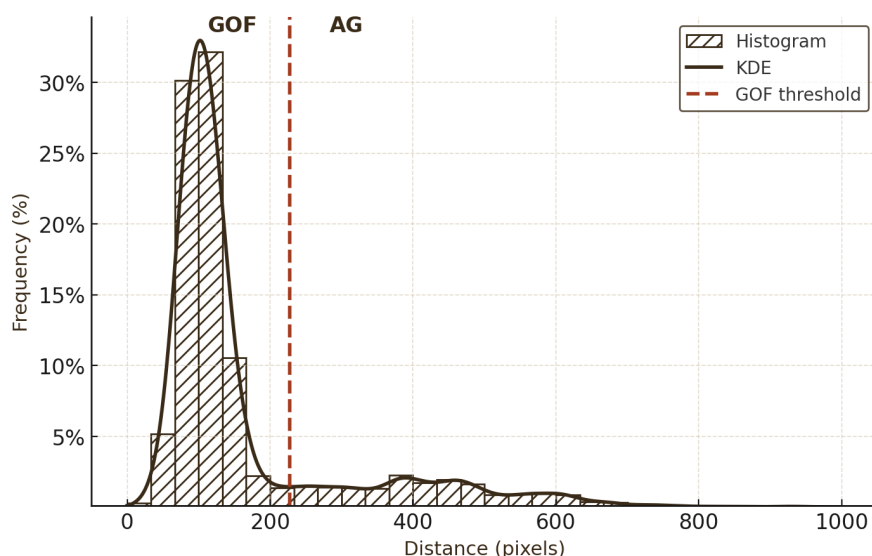


Figure 3.3. – Histogram of gaze distances to the face center for a participant during a session. The dashed vertical line marks the GOF threshold separating two gaze states: GOF (Gaze On Face) on the left, and AG (Averted Gaze) on the right.

This method provides a continuous, frame-by-frame time series of "gaze-on-face" that is robust to both detection failures and systematic calibration offsets.

3.4. A Preference for "Gaze-on-Face"?

Before we can analyze the complex, dynamic signaling functions of gaze, we must first answer a more fundamental, descriptive question. The existing literature, as discussed, provides a clear picture of gaze in infancy (as a learning tool) and in adulthood (as an interaction management tool), but a less clear picture of the period in between.

Our query is to establish a clear developmental baseline. We seek to provide a quantitative answer to the most basic question: in a comfortable, naturalistic setting, how much do school-age children actually look at their conversational partner?

Do children in middle childhood (ages 6-12) exhibit a consistent, preferential gaze towards their interlocutor's face during a naturalistic conversation, and how does this preference compare to that of their adult caregivers?

Answering this question is a prerequisite for all subsequent analyses. If children in our study show a high, adult-like proportion of "gaze-on-face," it would suggest that the fundamental ability and preference to direct gaze on an interlocutor are mature skills by this age. This would imply that the key developmental changes to look for are not in the quantity of gaze, but in its quality, which we explore in Chapter 4.

Conversely, if children show significantly less "gaze-on-face" than their parents, even in a familiar setting, it would suggest a more protracted developmental trajectory for this foundational skill.

3.5. Results: A Robust and Early Adult-like Preference

3.5.1. Analysis 1: The Spatial Distribution of Gaze

First, we examined the spatial distribution of all gaze coordinates relative to the center of the interlocutor's face across the entire 15-minute conversations. The kernel density plot (Figure 3.4) shows the average distribution for all children and all parents across all sessions. We present the average distribution across the whole corpus in Figure 3.5. Since the distances are expressed in pixels, we first normalized them by the threshold to average across all sessions.

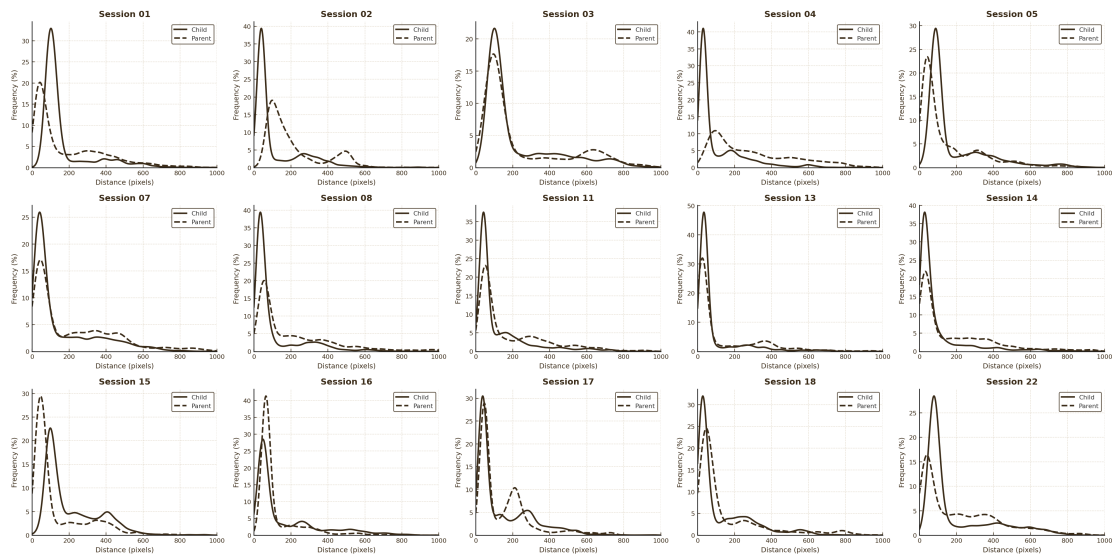


Figure 3.4. – Kernel density estimates (KDEs) of gaze distances to the face center across 15 sessions. Each subplot compares the gaze behavior of the child (solid line) and parent (dashed line) within the same session. The x-axis represents gaze distance in pixels (0–1000), and the y-axis shows the relative frequency (%).

3. Gaze on Face in Middle Childhood: An Emergent Adult-like Preference – 3.5. Results: A Robust and Early Adult-like Preference

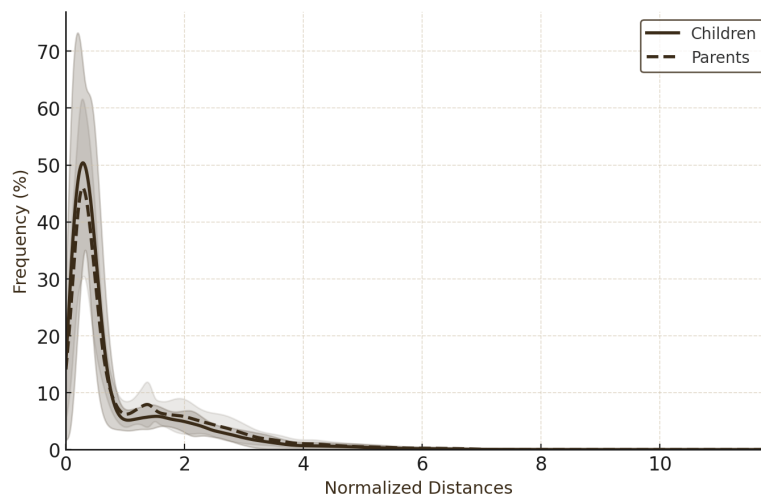


Figure 3.5. – Averaged Kernel Density Estimates (KDEs) of normalized gaze distances across 15 child–caregiver sessions. Distances were normalized by each session’s gaze threshold ($distance/threshold$). The solid line represents the mean distribution for children, and the dashed line represents the mean distribution for parents. Shaded areas indicate the standard deviation across sessions, reflecting inter-session variability in gaze behavior.

The analysis of this distribution reveals a clear and consistent pattern. For both children and parents, the distribution of gaze distances is dominated by a sharp, high-density peak at short distances. This peak graphically represents the massive concentration of gaze points falling directly on or in the immediate vicinity of the interlocutor’s face. This finding strongly indicates that the interlocutor’s face serves as the primary spatial anchor for gaze throughout the entire interaction, for both children and their adult caregivers.

Furthermore, the distribution is not purely unimodal. Smaller, secondary peaks are visible farther from the face. These secondary modes can correspond to other consistent points of fixation in the scene that participants occasionally scanned (e.g., objects in the room, a painting on the wall, or a toy). It demonstrates that participants were not simply "locked" onto a single visual target, but were actively moving their eyes and heads, with their gaze preferentially and repeatedly returning to the interlocutor’s face.

While Figure 3.5 presents a compelling average, it could mask significant individual variation. However, a separate analysis of individual participant data confirms the robustness of this finding. This pattern was not an artifact of averaging; every participating child and adult exhibited a qualitatively identical distribution, with their gaze most densely clustered around their partner’s face. This spatial analysis thus provides an unequivocally affirmative answer to the first part of the research question: children in middle childhood exhibit a consistent and dominant preferential gaze towards their interlocutor’s face, a pattern indistinguishable from that of adults.

3.5.2. Analysis 2: The Proportion of "Gaze on Face"

Next, we quantified the total proportion of time spent gazing at the interlocutor's face. Using our gaze-on-face estimation pipeline to determine where participants look. Our second analysis quantified how long they spend looking at the face. Using the participant-specific thresholds derived from our heuristic, we converted the continuous time series into a binary variable: gaze-on-face (1) or averted-gaze (0) for each frame. This allowed us to calculate the total proportion of conversation time each participant spent gazing at their partner's face. Figure 3.6 presents the proportion of GOF calculated for each participant (child and parent) across the 15 dyads. This visualization reveals considerable variability between dyads, with proportions ranging from just under 50% to over 80%. However, in most dyads, the proportions of child and parent are relatively similar.

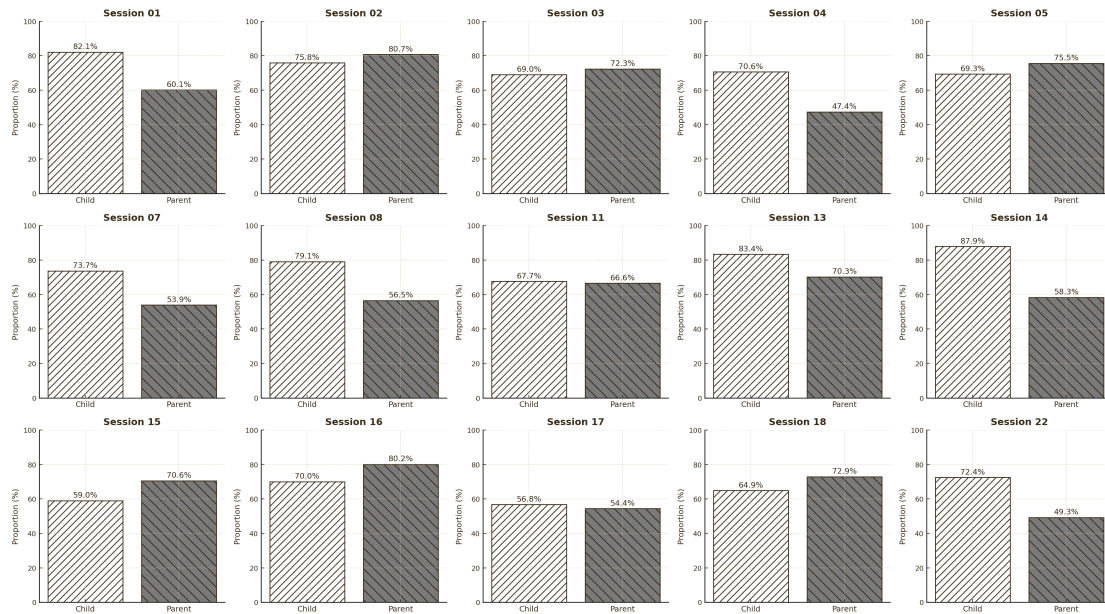


Figure 3.6. – Proportion of gaze samples where the gaze distance was below the session-specific threshold (i.e., gaze-on-face) for children and parents across 15 sessions. Each subplot shows the percentage of gaze-on-face instances for the corresponding session, with children represented by hatched white bars and parents by lightly hatched grey bars. This comparison illustrates inter-session variability and relative differences in gaze-on-face behavior between children and parents.

Figure 3.7 aggregates these individual data points to show the overall average proportion of gaze-on-face for children compared to parents. The bars represent the mean proportion across all participants in each group, with error bars indicating the variability. This figure shows that, on average, children (72.1%) spent slightly more time gazing at their parents' faces than parents (64.6%).

3. Gaze on Face in Middle Childhood: An Emergent Adult-like Preference – 3.6. *Discussion: Re-evaluating the Developmental Trajectory*

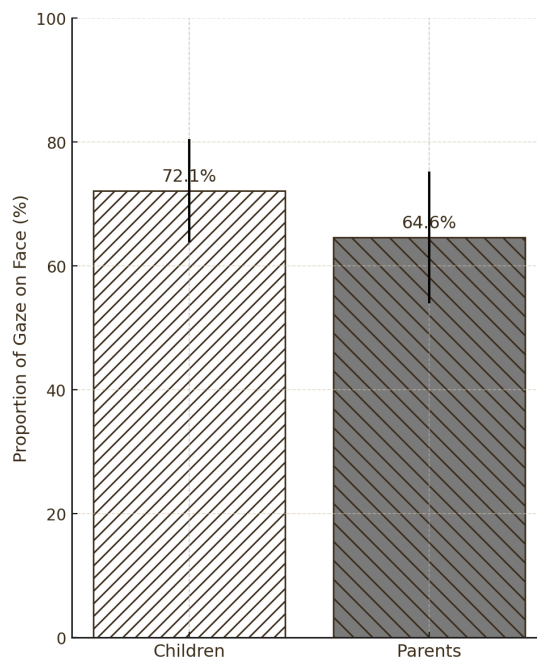


Figure 3.7. – Mean proportion of gaze samples where the gaze distance was below the session-specific threshold (gaze-on-face) averaged across 15 sessions. Error bars represent standard deviation across sessions.

3.6. Discussion: Re-evaluating the Developmental Trajectory

The findings from this chapter offer a clarification to the field. Our data show that children in middle childhood, when observed in a familiar and comfortable social context, exhibit a "gaze-on-face" preference that is unequivocally adult-like in its consistency and magnitude.

This strongly supports the hypothesis that the reduced gaze findings in prior work were largely context-dependent (De Lillo et al. 2021). Our findings, gathered in an ecologically valid context, demonstrate that the fundamental drive and ability to "anchor" one's gaze on an interlocutor's face during conversation are well-established by at least age 7.

This conclusion has significant implications. It suggests that development beyond middle childhood may not be about learning to look at the face, but rather about learning how and when to modulate this gaze for complex social signaling.

We have established that, overall, children's gaze-on-face behavior is adult-like. This, however, does only address the "information acquisition" coin side of gaze. It does not tell us whether children use this gaze in an adult-like way for "information signaling."

An adult's gaze is not static; their conversational role dynamically modulates it. Adults typically gaze more when listening (to gather information) and less when

3. *Gaze on Face in Middle Childhood: An Emergent Adult-like Preference – 3.6.*
Discussion: Re-evaluating the Developmental Trajectory

speaking (to manage cognitive load or signal they are holding the floor) (Kendon [1967](#)). Is this crucial modulation also present in middle childhood?

Having established that children do look at the face, we now turn to this more complex question. In Chapter 4, we will analyze the fine-grained, moment-to-moment dynamics of this gaze. We will investigate how it is modulated by speech and whether children's gaze patterns differ when speaking versus listening.

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling

Table of contents

4.1. From Prevalence to Function: Posing the Central Problem	77
4.2. Methodological Note: Segmentation of Speech and Gaze Data	79
4.2.1. Speech Segments	79
4.2.2. 'Speech Segment' vs. 'Conversational Turn': A Critical Distinction	81
4.2.3. Gaze Segments	82
4.3. Analysis 1: Role-Based Gaze Modulation (Listening vs. Speaking)	83
4.4. Analysis 2: Turn-Coordination Gaze (Turn Taking vs. Yielding)	86
4.5. Discussion	89
4.5.1. Role-Based Gaze Modulation as a Stable Component of Interac- tional Competence	89
4.5.2. Gaze and Turn Coordination: Graded Precision Rather Than Ab- sent Competence	90
4.5.3. Developmental Continuity and Refinement of Gaze Use	90
4.5.4. Implications for Modeling Turn-Taking Behavior	90
4.5.5. Summary	91

4.1. From Prevalence to Function: Posing the Central Problem

For decades, the functional mechanisms of gaze in adult face-to-face interactions have been well understood. Foundational work established that gaze is not random, but a highly structured, functional signal used to regulate conversational flow (Kendon 1967) and manage turn-taking (Ho et al. 2015). However, the developmental trajectory of these skills has remained a significant gap in the literature, marked by conflicting data.

On one hand, some studies suggest adolescents exhibit less face-directed gaze than adults, implying a protracted, non-linear developmental curve (De Lillo et al. 2021). On the other hand, the foundational question of when children master the prevalence of adult-like gaze has been confounded by non-naturalistic, lab-based settings.

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.1. From Prevalence to Function: Posing the Central Problem

Chapter 3 established a foundational premise: children in middle childhood (ages 6-12) are not gaze-averse in natural conversations. On the contrary, they demonstrate a robust, adult-like preference for gazing at their interlocutor's face.

This high prevalence of gaze, however, presents a critical research problem. We argue that the sole act of "looking" is not synonymous with the functional use of gaze. This distinction, between gaze as a passive state and as an active, coordinated social signal, has been a central theme in conversational analysis since the foundational work of Kendon 1967. Gaze is not simply "on" or "off"; it is a dynamic and multipurpose tool. As we framed it in earlier chapters, gaze serves a dual purpose: "information acquisition" (e.g., monitoring a partner's comprehension) and "information signaling" (e.g., coordinating turns).

The findings from Chapter 3, that children's gaze prevalence is adult-like presents a critical puzzle. To mistake this high prevalence for complete functional competence would be to adopt an incomplete and potentially misleading model of conversational development. It raises the question: Does this high prevalence correlate with an equivalent level of functional maturity? Or does a developmental divergence exist, where the high quantity of gaze has yet to be refined into a functional tool?

This chapter addresses this problem by presenting a two-part analysis of gaze function. This approach enables a separate investigation of two distinct types of gaze skills: one based on static, role-based knowledge, and the other on dynamic, high-paced signaling.

The first part of this analysis examines the functional role of gaze in a relatively static domain: modulation of gaze by conversational role (Listening vs. Speaking). This examines whether children, like adults, alter their gaze patterns in response to the distinct cognitive demands of speech production versus during listening. This line of inquiry assesses the child's mastery of the broad, rule-based social knowledge of conversational gaze.

The second part of the analysis transitions from this static, role-based function to a highly dynamic, time-critical one: the use of gaze as a precise signal for turn-coordination. This investigates the acts of "Taking" and "Yielding" the floor (Ho et al. 2015). This second inquiry moves beyond social knowledge to assess the child's procedural fluency in using gaze as a split-second tool for managing the conversation.

This two-part analysis will be used to construct a central argument regarding the relationship between these two functional skills. To build this argument, Section 4.3 will analyze the role-based modulation findings, followed by Section 4.4, which examines the turn-coordination signals. Finally, Section 4.5 will synthesize these findings and discuss the implications of this developmental divergence for our understanding of socio-cognitive maturation.

This two-pronged investigation will demonstrate a clear developmental divergence, forming the core argument of this chapter. To build this argument, Section 4.2 will first outline the key data processing and methodological distinctions that form the basis of the study. Section 4.3 will then present the analysis of role-based gaze, followed by Section 4.4, which will present the analysis of gaze in turn-coordination. Finally, will synthesize these findings into a unified discussion of this developmental divergence

and its cognitive implications.

4.2. Methodological Note: Segmentation of Speech and Gaze Data

Before proceeding to the two main analyses, we will be presenting the data processing that transformed the raw data into the analytical units required for this chapter.

4.2.1. Speech Segments

The first analysis (Listening vs. Speaking) requires the entire conversational interaction to be segmented into discrete segments of "speech" and "silence" (absence of speech) and correctly attributed to either the child or the parent (diarization). As detailed in Chapter 1 methodology, this was achieved through a combination of automated and manual processing. An initial pass was made using the WhisperX automatic segmentation tool (Bain et al. [2023](#)), but this contained notable inaccuracies. This initial draft was manually corrected using the ELAN interface. This rigorous manual pass, which included adjusting boundaries and adding missing segments, ensured a gold standard dataset. Additionally, speaker diarization was also manually adjusted to correct for microphone cross-talk. This process yielded a reliable set of speech segments that serve as the basis for analyzing the "Speaking" versus "Listening" roles.

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.2. Methodological Note: Segmentation of Speech and Gaze Data

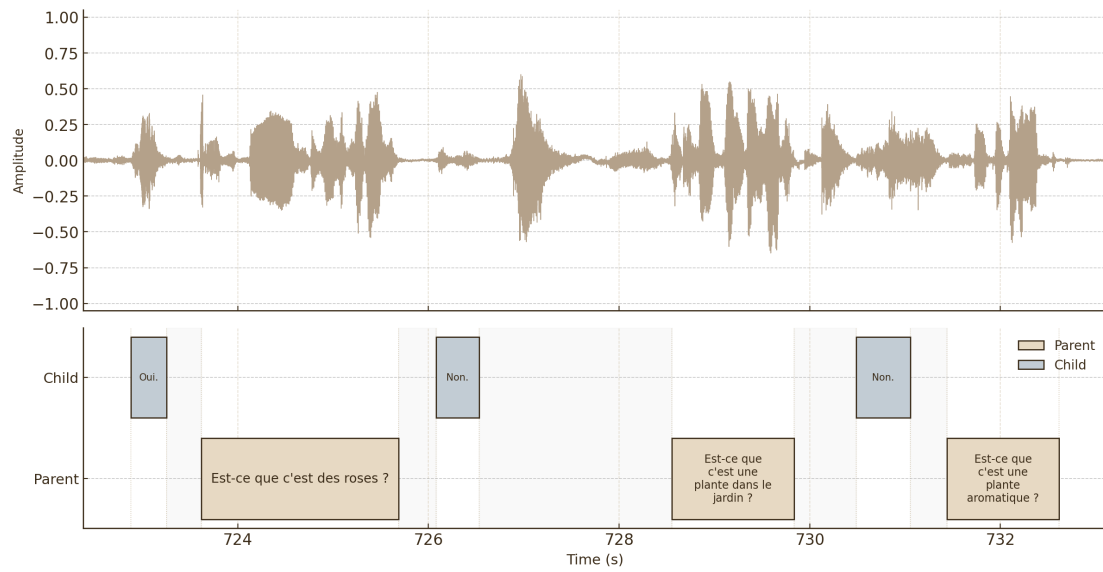


Figure 4.1. – Alignment between the audio waveform and transcribed speech segments for the Parent and Child. Each colored bar represents a speech turn, positioned according to its start and end times. The waveform above corresponds to the acoustic signal. Dotted vertical lines mark segment boundaries, and light gray shaded areas indicate pauses in the conversation. Together, these elements illustrate the temporal structure and alternation of turns between speakers.

The automatic pass with WhisperX, based on the VAD approach, has a significant, known limitation for conversational analysis: it identifies the presence of any speech in the channel, resulting in a single "speech" track that, by its nature, cannot represent overlapping speech. For that, the manual pass also served to locate all the overlaps that were absent from the initial VAD output.

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.2. Methodological Note: Segmentation of Speech and Gaze Data

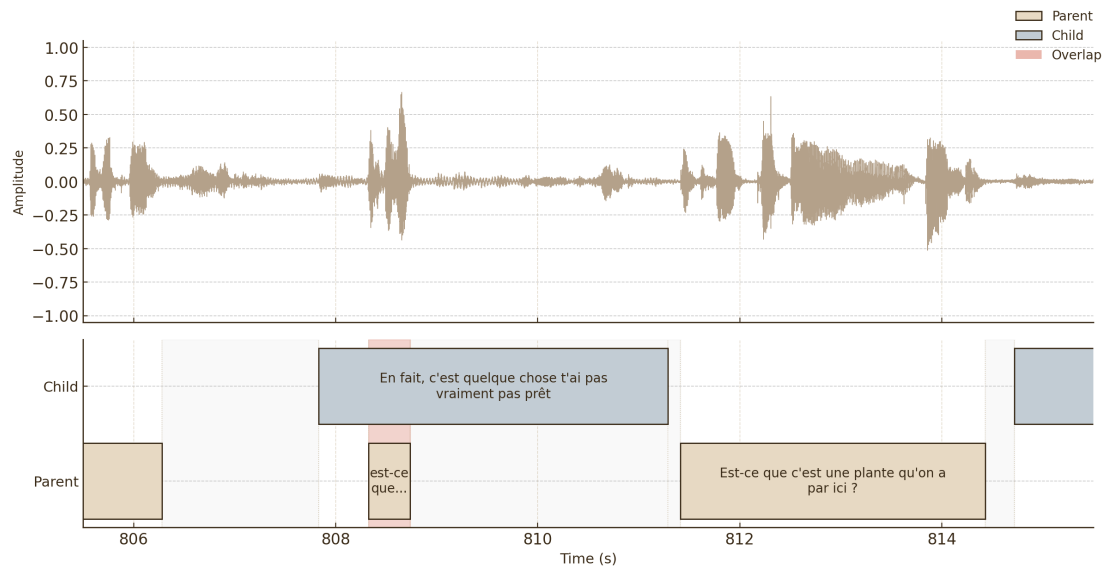


Figure 4.2. – Example of overlapping parent–child speech segments. The figure illustrates a short exchange where both speakers talk simultaneously, with the overlapping region shaded in light coral. The parent’s utterance (“est-ce que...”) begins as the child continues speaking (“En fait, c’est quelque chose t’ai pas vraiment pas prêt”), before the child completes their turn.

4.2.2. ‘Speech Segment’ vs. ‘Conversational Turn’: A Critical Distinction

This thesis draws a sharp, formal distinction between the “speech segments” defined in 4.2.1 and the “conversational turns” that are the subject of foundational Conversation Analysis in the literature (e.g., Sacks et al. 1974), a “turn” is a high-level functional unit, referring to a speaker’s entire contribution to a conversation before the next speaker begins. This definition is powerful for Conversation Analysis but operationally ambiguous. For example, I might ask, does a 3-second pause for thought constitute the end of a ‘turn’? Are brief, overlapping “mhm” backchannels their own turns, or part of the listener’s role?

The analyses in this chapter require a more precise, low-level unit of measurement. The “speech segment,” by contrast, is a micro-level, operational, and acoustic-phonetic unit (as opposed to a functional one). It is defined by the physical, measurable presence of a single speaker’s voice, as captured by a microphone and annotated by a hybrid pipeline that combines a WhisperX baseline with manual correction using ELAN. It is simply a continuous (or near-continuous) burst of speech from a single speaker.

This distinction is crucial for both of this chapter’s analyses:

- For Analysis 1 (Role-Based): A “turn” model would ambiguously categorize long pauses (where a speaker is still “holding the floor”). The “speech segment” model, by contrast, enables a precise categorization of all data into either “Speak-

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.2. Methodological Note: Segmentation of Speech and Gaze Data

ing" (when an active segment is present) or "Listening" (when an active segment from the other interlocutor is present).

- For Analysis 2 (Turn-Coordination): The cross-correlation method (Ho et al. 2015) is a mathematical operation. It requires two binary continuous-time series: "speech on/off" and "gaze on/off." A high-level functional "turn" cannot be fed into this analysis. The "speech segment" provides the precise speech timeseries required to measure the micro-temporal alignment between gaze and speech at turn boundaries.

We believe that it is an error to treat these two concepts as interchangeable. The signals this thesis investigates are precisely what complicate a simple one-to-one mapping of segment to turn. This is not a simplification; it is a deliberate methodological choice. A "turn" is a high-level social accomplishment, a complete, functional unit that participants jointly negotiate. Our research quest, however, is not "what is a turn?" but rather "how are the micro-level signals (gaze, speech) that build that turn develop in middle childhood?

Therefore, this chapter focuses on "speech segments", enabling a level of quantitative precision that would be impossible with the more abstract, high-level concept of "conversational turn."

4.2.3. Gaze Segments

The two analyses in this chapter require gaze data to be prepared in two distinct forms, derived from the same raw eye-tracking source.

The first data form is a binary, frame-by-frame timeseries. As detailed in Chapter 3, a gaze aversion threshold was calculated for each participant. This threshold allowed every frame of the recording to be classified as either "gaze-on-face" (coded as 1) or "averted-gaze" (coded as 0). This un-smoothed, high-resolution binary time series is the data used to calculate the average proportion of face-directed gaze during the "Speaking" and "Listening" speech segments.

However, raw eye-tracking data is often noisy, containing transient, millisecond-level flickers that do not represent a meaningful change in gaze. To reduce the influence of noise and transient gaze changes, the binary time series required an additional smoothing step using a 50-ms sliding window. This process transformed the frame-by-frame gaze coordinates into largely stable, continuous segments.

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.3. Analysis 1: Role-Based Gaze Modulation (Listening vs. Speaking)

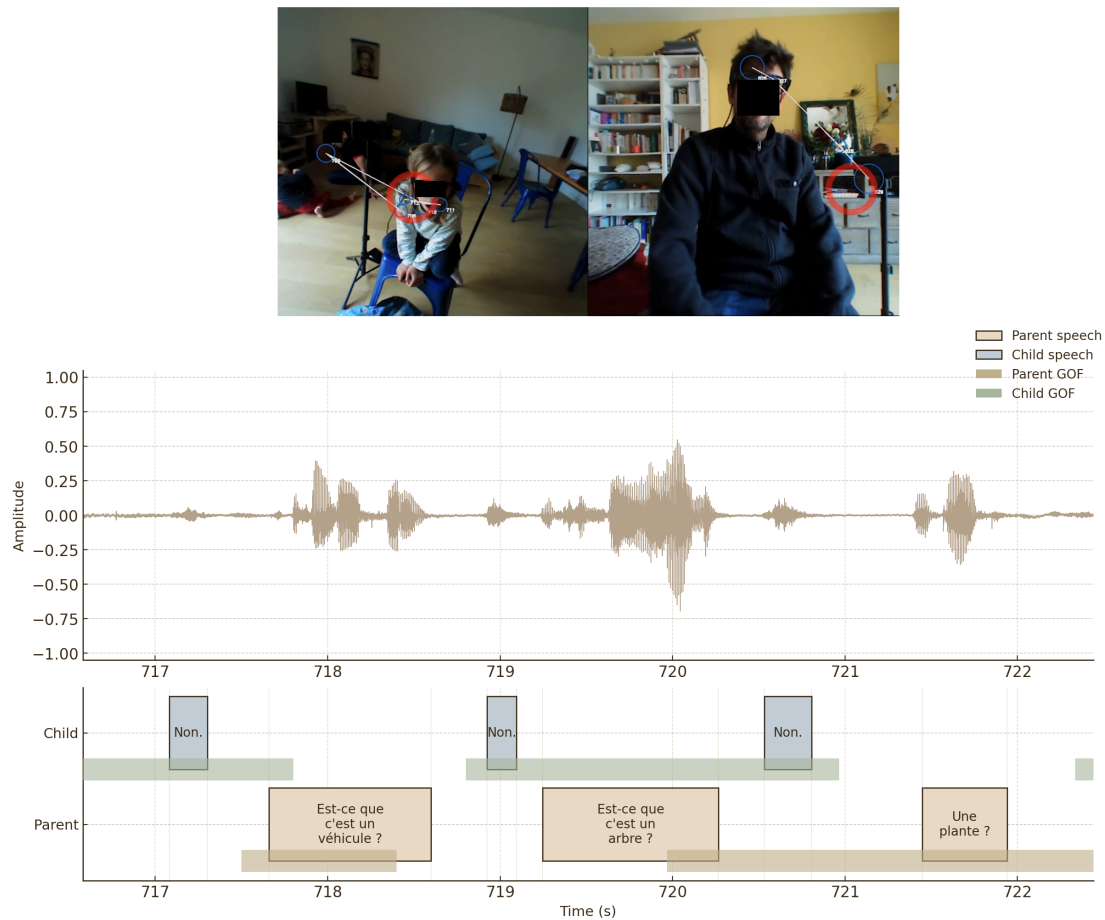


Figure 4.3. – Synchronized example of gaze and speech. Top: paired egocentric frames with overlaid gaze vectors from child (left) and parent (right). Bottom: The figure illustrates a brief segment of interaction, featuring alternating speech turns between the parent and the child. Green and beige horizontal bands represent moments when each participant's gaze was directed toward the other's face (GOF). Time in seconds.

4.3. Analysis 1: Role-Based Gaze Modulation (Listening vs. Speaking)

The first analysis addresses the more static, role-based function of gaze. The theoretical premise, grounded in cognitive load theory, states that gaze behavior is functionally split: speakers avert gaze to manage the high cognitive demand of speech planning (Glenberg et al. 1998), while listeners maintain gaze to monitor social signals (Kendon 1967). The question is whether children in middle childhood have successfully acquired and operationalized this functional dichotomy.

This is tested by binning the speech segments (4.2.1) against the gaze timeseries (4.2.3) to determine the average percentage of face-directed gaze during roles: (1)

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.3. Analysis 1: Role-Based Gaze Modulation (Listening vs. Speaking)

speaking, and (2) listening. We modeled this relationship using a logistic mixed-effects model (GLMM) to account for the hierarchical nature of the data.

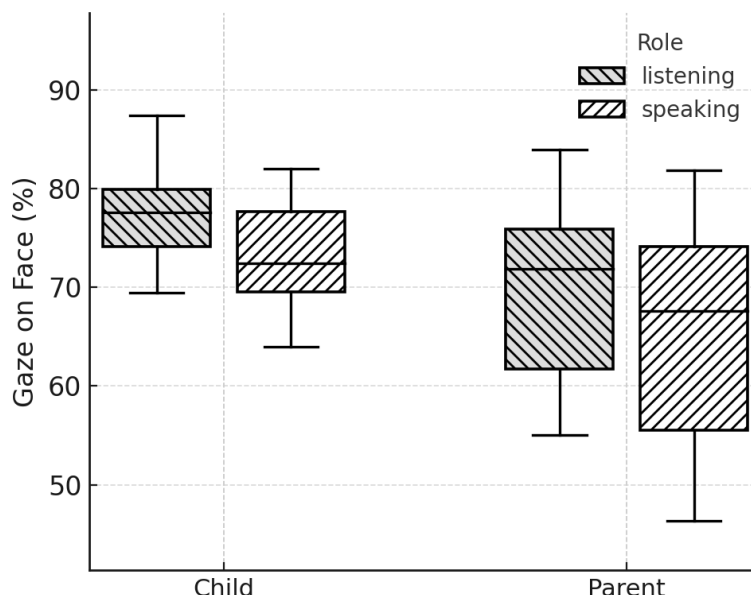


Figure 4.4. – Gaze-on-face proportions for children and parents while listening and speaking. Each boxplot reflects the distribution across 15 sessions.

As illustrated in Figure 4.4, both children and their adult caregivers exhibited the same functional pattern: both children and parents generally display higher face-directed gaze while listening than while speaking.

To quantify these patterns, we used a logistic mixed-effects model predicting gaze_at_face from Interlocutor (child vs. parent) and Role (speaking vs. listening), along with their interaction. In addition, because we are analyzing speech segments within each conversation/dyad, we included a random effect for dyad and for segment (nested within dyad).¹

As in our earlier analysis, we added an auto-regressive process to the glmmTMB model.²

As expected, we found a main effect of Interlocutor (child vs. adult) on gaze whereby children gaze more at the face: $\beta = 0.46$ ($SE = 0.10$, $p < 0.001$) as well as a main effect of Role (listening vs. speaking) whereby participants gazed more when listening than when speaking: $\beta = 0.21$ ($SE = 0.1$, $p < 0.05$). We did not observe an interaction between Interlocutor and Role: $\beta = 0.001$ $SE = 0.14$, $p = 0.99$, indicating that the modulation of gaze between listening and speaking does not differ significantly between children and adults.

1. Specified in glmmTMB as: `gaze_at_face ~ Interlocutor * Role + (1 | dyad/segment)`

2. Specified in the same glmmTMB as: `dispformula = ~ ar1(timestamp | dyad:segment)`

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.3. Analysis 1: Role-Based Gaze Modulation (Listening vs. Speaking)

We conducted post hoc comparisons using the "emmeans" R library applied to the glmmTMB model. It revealed higher odds of looking at the interlocutor's face when listening rather than speaking, both for children $\beta = 1.23$ ($SE = 0.12$, $p < 0.05$) and adults $\beta = 1.23$ ($SE = 0.12$, $p < 0.05$). Consistent with the non-significant interaction in the main model, the magnitude of this effect does not differ between children and adults.

The findings from our analysis are definitive. Statistically, we found a significant main effect of Role ($\beta = 0.21$, $SE = 0.1$, $p < 0.05$). This confirms that for both groups, the act of speaking triggers a measurable shift in visual attention away from the partner's face.

Crucially, our model revealed no significant interaction between Interlocutor (Parent vs. Child) and Role ($\beta = 0.001$, $SE = 0.14$, $p = 0.99$). This null result is a robust developmental finding. It indicates that the magnitude of gaze modulation—the extent to which one looks away when speaking—is statistically indistinguishable between school-age children and adults. Post-hoc comparisons further supported this, showing higher odds of looking at the face when listening for both children ($\beta = 1.23$, $SE = 0.12$, $p < 0.05$) and adults ($\beta = 1.23$, $SE = 0.12$, $p < 0.05$).

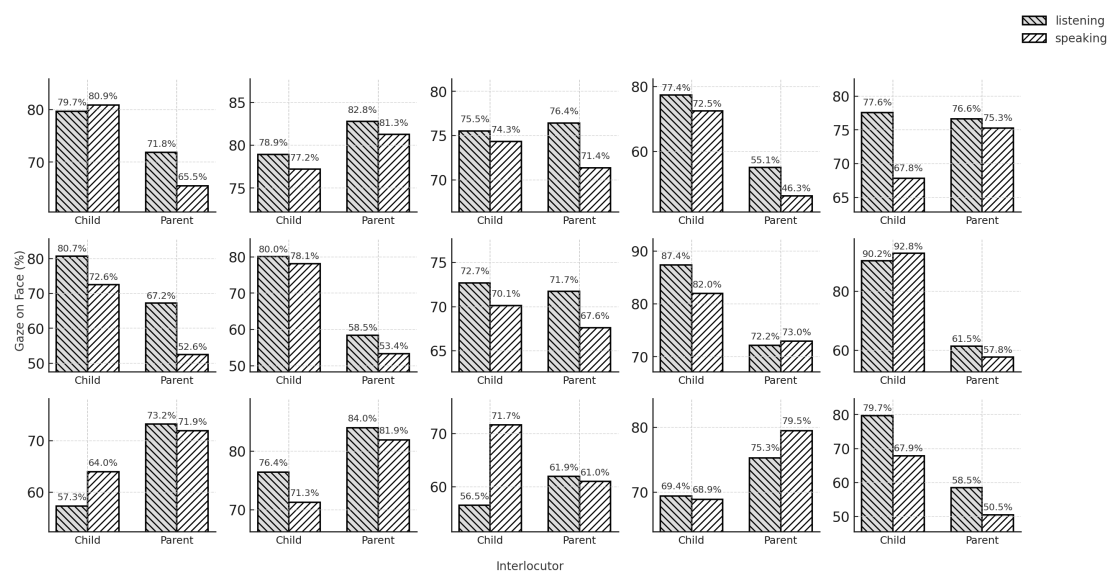


Figure 4.5. – Gaze-on-face percentages for children and parents while listening and speaking across 15 recorded sessions. Each subplot represents one session and includes two bars per interlocutor.

However, while the aggregate pattern is robust, naturalistic data is inherently noisy. To validate that this is a generalizable competence and not an artifact of averaging, Figure 4.5 breaks down this modulation for each of the 15 dyads.

The dyad-level analysis in Figure 4.5 reveals significant individual heterogeneity; the absolute baseline of gaze varies considerably from person to person. Yet, despite this variability, the modulation signal remains consistent: the pattern of "Listening Gaze >

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.4. Analysis 2: Turn-Coordination Gaze (Turn Taking vs. Yielding)

Speaking Gaze" holds for a clear majority of participants (11 out of 15 children and 13 out of 15 adults). This consistency across individuals, despite the varying dynamics of each unique conversation, confirms that role-based modulation is a stable, underlying competence.

Taken together, these results indicate that children in middle childhood possess a stable, role-sensitive gaze organization that closely parallels that of adult interlocutors. The absence of an interaction between Interlocutor and Role is particularly informative in this respect: it suggests that the functional modulation of gaze between speaking and listening is not just present in children but is deployed in a quantitatively comparable manner to that of adults.

4.4. Analysis 2: Turn-Coordination Gaze (Turn Taking vs. Yielding)

We test whether gaze is used effectively as a signal for turn coordination, focusing on two previously documented patterns in adults (Kendon 1967; Duncan and Fiske 2015): (1) **turn-taking**: when the participant intends to take a turn, they begin speaking with eye gaze averted, a way to signal that they now have the floor, and (2) **turn-yielding**: prior to the end of speaking, the participant will signal turn-transition by gazing at their partner.

We follow closely the methodology outlined in Ho et al. 2015, adopting their operationalization of both turn-taking and turn-yielding signals, using cross-correlation. In this approach, speech and gaze are treated as binary timeseries. For Speech timeseries, sequences of 1s indicate speech segments while sequences of 0s indicate silence. For Gaze timeseries, sequences of 1s indicate gaze-on-face segments while sequences of 0s indicate gaze aversion. Cross-correlation measures the alignment between these timeseries across different temporal shifts (or lags). The correlation is high when speech and gaze are synchronized (e.g., speaking while looking at the interlocutor) and low when they are not (e.g., speaking first, then gazing later). By systematically shifting one timeseries relative to the other, cross-correlation identifies the lag at which the alignment is maximum. This reveals the extent to which one timeseries tends to follow or precede the other.

For the turn-taking signal, we test whether own-gaze *follows* own-speech. When using the own-gaze timeseries as the reference, we expect higher alignment (i.e., stronger correlation) when own-speech is shifted forward in time, meaning that the maximum correlation should occur at positive lags.

For the turn-yielding signal, we test whether own-gaze *precedes* other-speech. When using the own-gaze time series as the reference, we expect higher alignment when other-speech is shifted backward in time, meaning that the maximum correlation should occur at negative lags.

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.4. Analysis 2: Turn-Coordination Gaze (Turn Taking vs. Yielding)

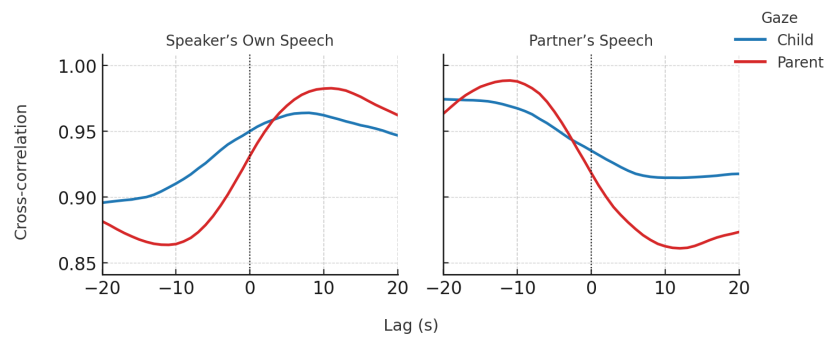


Figure 4.6. – Mean cross-correlation functions (-20 to +20 s) between gaze-on-face and speech activity, averaged across all dyads and sessions. Each plot shows separate curves for child gaze and parent gaze. The left panel corresponds to moments when the speaker is producing speech (“Speaker’s Own Speech”), and the right plot corresponds to moments when the partner is speaking (“Partner’s Speech”). Vertical dashed lines indicate zero lag.

We computed cross-correlations separately for each dyad and, for visualization purposes, averaged the results across all dyads at each lag. Figure 4.6 presents these aggregated cross-correlation patterns.

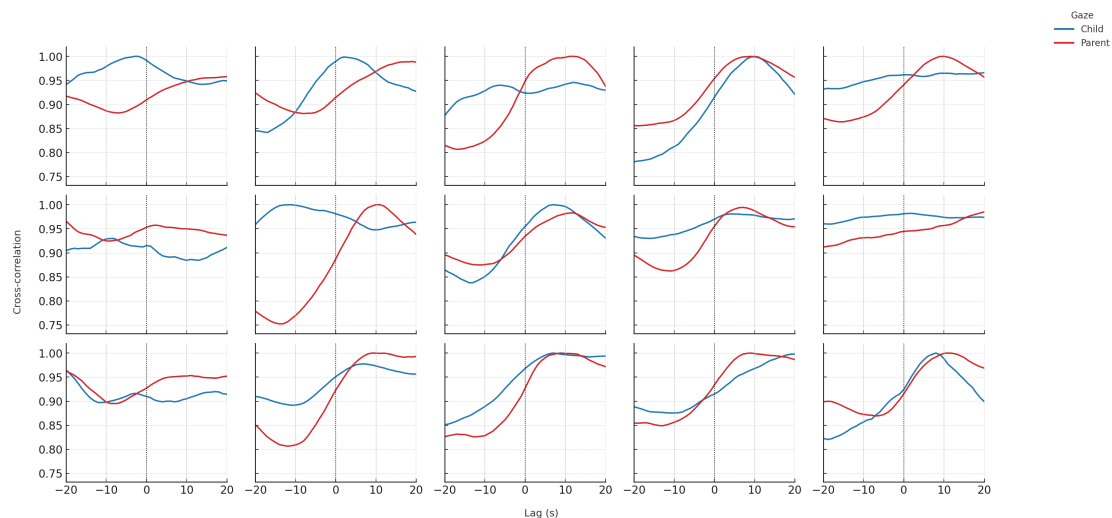


Figure 4.7. – Mean cross-correlation functions (-20 to +20 s) between gaze-on-face and speech activity, computed across dyads. The speaker is producing speech (“Speaker’s Own Speech”).

In Figure 4.6 (left), which reports results on turn-taking signaling, adults exhibit the highest cross-correlation values at positive lags, indicating that their gaze tends to lag behind the onset of their own speech. This supports the turn-taking signaling hypothesis. Children also show higher cross-correlation values at positive lags. An examination of individual dyadic data (Figure 4.7) further supports this pattern: in the

4. Gaze Dynamics in Middle Childhood: Role-Based Modulation and Turn-Signaling – 4.4. Analysis 2: Turn-Coordination Gaze (Turn Taking vs. Yielding)

majority of parents and children, the maximum correlation occurs at positive lags. To quantify these qualitative observations, a binomial test was conducted to determine whether the maximum cross-correlation occurred significantly more at positive lags. For adults, the max correlation occurred at positive lags in 13 out of 15 cases (86.7%, $p < 0.01$). For children, the max correlation occurred at positive lags in 11 out of 15 cases (73.3%, $p = 0.059$).

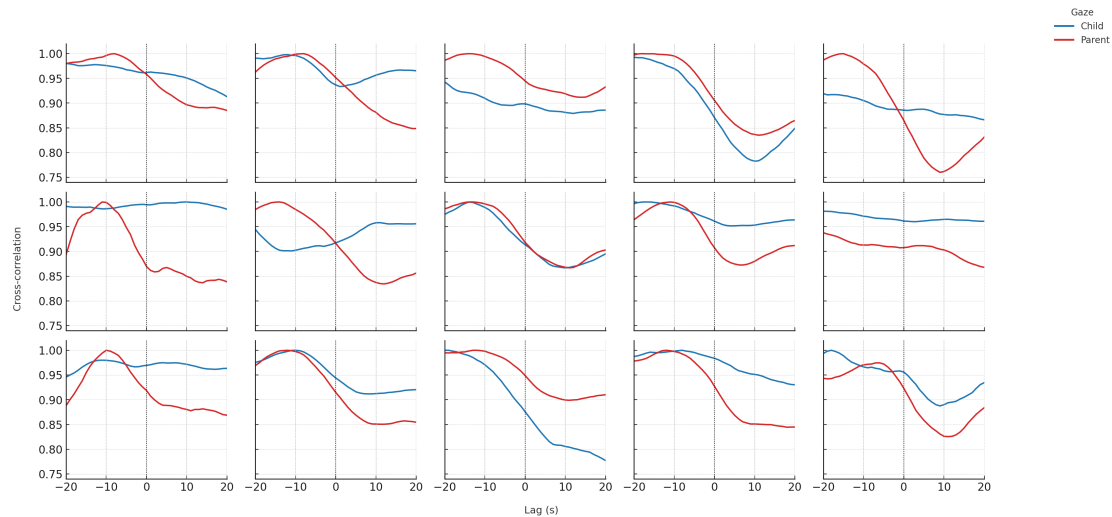


Figure 4.8. – Mean cross-correlation functions (-20 to +20 s) between gaze-on-face and speech activity, computed across dyads. The partner is speaking (“Partner’s Speech”).

In Figure 4.6 (right), which reports results on turn-yielding signaling, adults exhibit the highest cross-correlation values at negative lags, indicating that their gaze precedes the interlocutor’s speech onset, consistent with the turn-yielding hypothesis. Children also show higher cross-correlation at negative lags. Individual data (Figure 4.8) confirm that most dyads mirror this general pattern observed in the aggregate results. This was the case for all parents and the majority of children. We conducted a binomial test to determine whether the maximum cross-correlation occurred significantly more at negative lags. For adults, the max correlation occurred at negative lags in 15 out of 15 cases (100%, $p < 0.001$). For children, the max correlation occurred at negative lags in 13 out of 15 cases (86.7%, $p < 0.01$).

Our findings replicate previous results observed in adults regarding turn-taking and turn-yielding signals (Ho et al. 2015). Moreover, we demonstrate that these signaling patterns are already present in middle childhood, though with a reduced magnitude, as we examine next.

Figure 4.6 shows that adults have higher cross-correlation values in both the case of turn-taking and turn yielding. This pattern was overall mirrored across individual dyads (Figures 4.7 and 4.8). To statically test these qualitative observations, we, first, computed the peak cross-correlation coefficient (i.e., the maximum value of the cross-correlation function) for both parents and children within each dyad. We then tested

whether the mean difference of these peak values (parent - child) was significantly different from zero using a paired permutation test, providing a fully non-parametric evaluation.³ In the turn-taking condition (i.e., own gaze vs. own speech), the observed difference was statistically significant ($M = 0.47$, $p = 0.02$). In the turn-yielding condition (i.e., own gaze vs. others' speech), the observed difference was marginally significant ($M = 0.51$, $p = 0.05$). We also conducted a Wilcoxon signed-rank test, which produced p-values identical to those from the paired permutation test.

Taken together, these results indicate that children in middle childhood do deploy gaze in relation to turn structure, but that this deployment is less strongly and less consistently time-locked to turn transitions than in adults. While adults show clear modulation of gaze during turn yielding and turn taking, children show more variable patterns. These findings do not imply that children are unable to manage turn-taking. Instead, they point to a developmental refinement in the integration of gaze as a coordination tool.

4.5. Discussion

This chapter examined gaze behavior in child–caregiver interaction with the aim of moving beyond prevalence measures toward an analysis of gaze as a functional and an interactional signal. Using first-person eye-tracking data, the analyses focused on two complementary aspects of gaze use: its modulation by conversational role (speaking vs. listening) and its temporal deployment around speech. Together, these analyses provide a nuanced picture of gaze competence in middle childhood.

4.5.1. Role-Based Gaze Modulation as a Stable Component of Interactional Competence

The first analysis demonstrated that children modulate their gaze behavior as a function of their role. They exhibit higher levels of face-directed gaze while listening than while speaking. This pattern closely mirrors well-established findings in adult conversation and was observed consistently across dyads. The absence of an interaction between interlocutor type and role indicates that children and adults deploy this modulation in a quantitatively comparable manner.

From a developmental perspective, this finding suggests that the role-based organization of gaze constitutes a stable and relatively early-acquired competence. Children do not just look at their interlocutor; instead, they adjust their gaze to meet the cognitive demands of speaking and listening. This form of gaze regulation can be understood as a rule-based and relatively static foundation for using gaze in conversation.

3. We randomly shuffled the “parent” and “child” labels within each dyad across 10,000 permutations (a number that led to stable results across runs) to generate a null distribution. We calculated the p-value as the proportion of permuted differences that were as extreme or more extreme than the observed difference.

4.5.2. Gaze and Turn Coordination: Graded Precision Rather Than Absent Competence

The second analysis addresses a more demanding aspect of the conversation: the use of gaze in relation to the fine-grained timing of speech. Here, both children and adults exhibited systematic changes in gaze behavior around moments of turn yielding and turn-taking. However, the patterns observed in children were weaker, and more variable than those observed in adults.

Importantly, these differences should not be interpreted as evidence that children lack turn-taking competence. We argue that children in middle childhood are clearly capable of managing turns, as evidenced by the smooth progression of interaction and gaze modulation with turn structure. Instead, the results point to differences in the strength and temporal precision. In other words, gaze supports turn-taking in children, but it is less systematic and more probabilistic than in adults.

This distinction highlights the value of analyzing interactional behavior at multiple levels. As shown in Chapter 3, children in middle childhood spend a substantial proportion of conversational time gazing at their interlocutor's face. We have demonstrated that children are not gaze-averse in natural interaction. However, prevalence alone does not capture how gaze is functionally deployed. The present findings suggest that the dynamic form of gaze coordination continues to be refined during middle childhood. We believe that our findings underscore the limits of approaches (both in this thesis and in the broader literature) that focus exclusively on aggregate gaze proportions without considering their function, modulation, and dynamics.

4.5.3. Developmental Continuity and Refinement of Gaze Use

Taken together, the results of this chapter support a view of gaze development characterized by continuity and gradual refinement. Children in middle childhood already possess core competencies related to gaze, including sensitivity to their role in the conversation and to turn-taking structure. What distinguishes their behavior from that of adults is not the presence or absence of these competencies, but the consistency and strength.

This perspective aligns with developmental accounts that emphasize incremental tuning of social signals as children age. Rather than acquiring entirely new mechanisms, children are refining the competencies they already have.

4.5.4. Implications for Modeling Turn-Taking Behavior

By establishing that gaze contributes to turn coordination in children, albeit with reduced temporal precision. These findings motivate the modeling approach adopted in the next chapter.

4.5.5. Summary

In summary, this chapter demonstrates that children in middle childhood use gaze in a functional manner during conversation. Role-based modulation of gaze is already adult-like and stable, while the use of gaze for coordinating turns is present but under refinement.

5. Predicting Turn-Taking Dynamics in Middle Childhood

Table of contents

5.1. Introduction	93
5.2. Task Definition: Hold vs. Shift	95
5.2.1. Temporal Properties of Turn Transitions	95
5.3. Modeling Approaches	96
5.3.1. Method 1: Feature-Based Prediction	97
5.3.1.1. Feature Extraction	97
5.3.1.2. Models and Baselines	97
5.3.2. Method 2: LSTM-Based Prediction	98
5.3.2.1. Sequence Modeling	98
5.3.3. Method 3: Wav2Vec2 Representation	98
5.3.3.1. Deep Acoustic Embeddings	98
5.3.3.2. Ensemble Model: Late Fusion of Wav2Vec2 embeddings and Gaze Predictions	99
5.4. Results	99
5.4.1. Overview of Model Performance	99
5.4.2. Baselines	99
5.4.3. Feature-Based Models	100
5.4.4. Sequence-Based Models	100
5.4.5. Representation-Learning Models	100
5.4.6. Summary of Results	101
5.5. Discussion	101
5.5.1. The task is structurally asymmetric: Shift is learnable; Hold is not as much	101
5.5.2. Acoustic-only and gaze-only are not competing explanations; they fail in different ways	102
5.5.3. Naive multimodal fusion does not solve the hard part — and that is informative	102
5.5.4. Why LSTMs did not help: the bottleneck is not temporal modeling but available information	102
5.5.5. Deep acoustic representations improve the right thing when combined with gaze	103
5.5.6. What the results say about turn-taking rather than about models	103

5.1. Introduction

Throughout this thesis, we have been developing an account of gaze-on-face behavior in middle childhood. Chapter 3 established that children in this age range exhibit a robust and stable preference for gazing at their interlocutor’s face in natural conversation. Chapter 4 then demonstrated that gaze-on-face is not static, but dynamically modulated as a function of conversational role and temporally coordinated with turn-taking events. Together, these findings indicate that gaze-on-face is systematically integrated into the organization of children’s conversational behavior.

While these analyses provide strong evidence for a functional relationship between gaze-on-face, speech, and turn structure, they remain fundamentally descriptive. In daily conversation, interlocutors must continuously anticipate whether a speaker is about to continue speaking or yield the floor in order to time their own actions appropriately (Levinson and Torreira 2015; Levinson 2016). From this perspective, we believe that turn-taking is inherently predictive.

Early interactional and phonetic studies identified acoustic cues associated with turn completion, including pitch declination, final lengthening, amplitude reduction, and the temporal organization of pauses (Duncan 1972; Ford 1996; Gravano et al. 2011). In parallel, research on visual behavior demonstrated that gaze plays a central regulatory role in turn-taking, with speakers typically averting gaze while holding the floor and re-establishing gaze toward the interlocutor as they approach turn completion (Kendon 1967; Goodwin 1981). Subsequent multimodal studies showed that combining acoustic and visual information improves the prediction of turn transitions relative to acoustic cues alone (Skantze et al. 2014).

More recent computational approaches treat turn-taking as a short-term prediction problem. Rather than modeling turn transitions as discrete events. The objective is to estimate the probability of future speech activity within a short temporal window, based on dense acoustic representations and, in some cases, visual input (Skantze 2017). Recent extensions incorporate representation learning and large language models to improve speech activity forecasting (Ekstedt et al. 2022). While such approaches achieve strong performance in predicting speech onset and offset, their outputs typically consist of continuous probability estimates rather than discrete interactional decisions.

From an interactional perspective, this formulation of the task poses a fundamental limitation. Turn-taking cannot be reduced to local speech activity alone, because the distinction between holding and yielding the floor is not determined by whether speech momentarily stops, but by whether the current contribution is considered to be complete (Sacks et al. 1974). This prediction relies on the configuration of multiple cues produced across the turn, including prosodic contours that project completion, the placement of pauses relative to syntactic structure, and visual behaviors such as gaze toward the interlocutor (Kendon 1967; Local et al. 1986; Ford 1996).

5. Predicting Turn-Taking Dynamics in Middle Childhood – 5.1. Introduction

These cues only make sense in relation to one another and to where they occur within an unfolding turn. A brief silence after a syntactically incomplete unit is usually considered as hesitation and treated as part of the same turn. However, a silence that follows a prosodically complete unit and coupled with the speaker's gaze toward the listener is commonly understood as an opportunity for the other participant to take the floor (Schegloff et al. 1996; Levinson and Torreira 2015). Frame-based models struggle with this distinction because they focus on what happens locally in time rather than on how an action fits into the larger turn. Example: a hesitation mid-turn (a pause following an incomplete syntactic unit) and a silence that follows a completed turn look the same to the model (both predict silence), even though participants in conversation treat them very differently. From the model's point of view, a pause is simply a moment where speech energy drops, regardless of what came before or what the speaker is doing visually.

This limitation is particularly glaring in naturalistic interaction, where speech is incrementally constructed, pauses are frequent, and turn boundaries are continuously negotiated rather than discretely marked (Schegloff 2007). In such data, pauses of similar duration may serve radically different interactional functions. Distinguishing between pauses that maintain the floor and those that invite for a shift requires an analytical framework that explicitly targets turn transitions, rather than inferring turn-taking indirectly from local speech activity alone.

Against this backdrop, the present work adopts an event-based prediction framework based on turn transitions. We formulate a binary prediction task in which, for each conversational turn, the model must anticipate whether the subsequent state will be a Hold, in which the current speaker continues speaking, or a Shift, in which the floor is transferred to the interlocutor.

This formulation serves both methodological and theoretical purposes. Methodologically, it enables a controlled comparison between different signal types and modeling strategies, ranging from interpretable acoustic and gaze-based features to sequence-based and representation-learning approaches. Theoretically, it provides a test of functional relevance: if gaze-on-face is genuinely integrated into the turn-taking system, it should contribute to predictions beyond chance.

The central research question of this chapter is: to what extent does gaze-on-face, alone or in combination with acoustic cues, contribute to predicting turn transitions in naturalistic conversation? It is important to note tho, that the objective of this chapter is not to maximize predictive accuracy or to propose a comprehensive computational model of turn-taking. Rather, prediction is used here as an analytical tool to probe the structure and limits of multimodal coordination in naturalistic conversation. By quantifying what can (and can not) be predicted from gaze-on-face and speech, this chapter provides a quantitative complement to the descriptive analyses in the previous chapters.

5.2. Task Definition: Hold vs. Shift

The prediction task addressed in this chapter is formulated as a binary classification problem over speech-segment transitions. For each speech segment produced by a speaker, the goal is to predict whether the subsequent speech segment is produced by the same speaker (holding the floor) or by the interlocutor (speaker change). For each turn t produced by a speaker S , we try to predict the speaker of the next turn $t + 1$.

- **Hold (0):** The current speaker S continues to speak in the next turn (i.e., $Speaker_{t+1} = S$).
- **Shift (1):** The conversational floor is yielded to the interlocutor (i.e., $Speaker_{t+1} \neq S$).

This formulation deliberately adopts an operational unit that differs from the high-level notion of “conversational turn” used in Conversation Analysis. As discussed in Chapter 4, the analyses in this thesis rely on speech segments (contiguous stretches of speech attributed to a single speaker) because they provide temporally precise boundaries suitable for aligning gaze and acoustic signals, and because they allow overlaps and short vocal events to be represented explicitly.

The dataset for this analysis consists of 3,678 speech-segment transitions extracted from the face-to-face portion of the CHICA corpus. The class distribution is imbalanced, reflecting the natural structure of these interactions, with 2,655 instances of Shift (approximately 72%) and 1,023 instances of Hold (approximately 28%). This imbalance motivates evaluation metrics that do not rely solely on raw accuracy.

For each instance, prediction is based on multimodal information extracted from a fixed temporal window preceding the end of the current speech segment. This window captures the acoustic and visual cues available prior to the transition point. The feature sets and modeling approaches used to operationalize this prediction task are described in the following sections.

5.2.1. Temporal Properties of Turn Transitions

Before introducing the predictive models, it is informative to examine the temporal properties of turn transitions in the dataset, in particular, the duration of the silent gap separating successive speech segments. Gap duration has been shown to play an important role in conversational coordination, and very short gaps are often associated with smooth turn transitions, whereas longer gaps may reflect hesitation or floor holding. At the same time, interactional research has emphasized that gap duration alone is rarely sufficient to determine the interactional status of a transition.

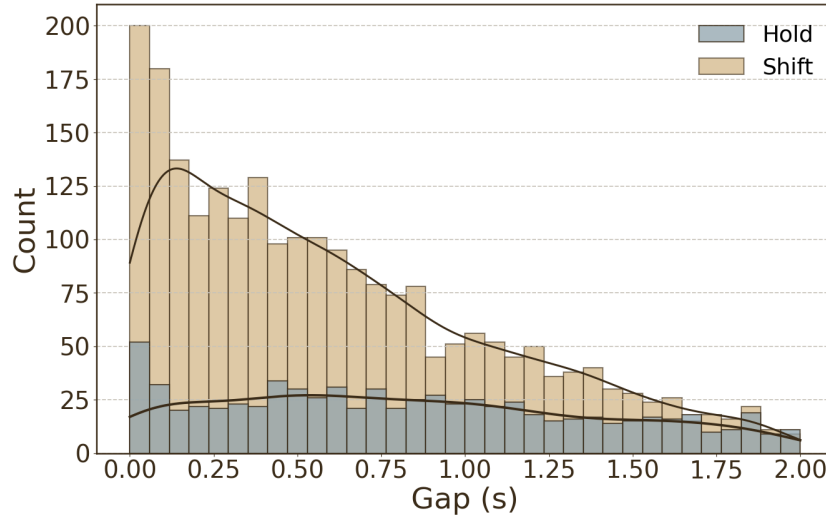


Figure 5.1. – Gap duration distributions for Hold and Shift transitions. Distribution of silent gap durations preceding Hold and Shift. Although Shift transitions tend to be associated with shorter gaps on average, the substantial overlap between the two distributions indicates that gap duration alone is insufficient to reliably distinguish between turn holding and turn yielding in naturalistic conversation.

Figure 5.1 shows the distribution of gap durations preceding Hold and Shift outcomes. The gap is defined as the silent interval between the end of the current speaker’s speech segment and the onset of the next speech segment, irrespective of speaker identity.

As expected, Shift transitions are associated with shorter gaps, with a higher concentration of very short gap durations preceding speaker change. Gap durations preceding Shift and Hold transitions show distinct temporal profiles. Shift transitions are concentrated at very short gap durations. As gap duration increases, Shift transitions become progressively less frequent. In contrast, Hold transitions are observed across a wide range of gap durations, spanning both short and long gaps. Rather than clustering at a specific temporal interval, Hold outcomes remain present across the full distribution.

5.3. Modeling Approaches

To address the turn-transition prediction task defined in Section 5.2, we evaluate multiple modeling strategies under a common experimental framework. The objective is not to optimize performance for a particular architecture, but to compare how different representations of speech and gaze contribute to predicting Hold and Shift outcomes. All models are trained and evaluated on the same dataset, using identical evaluation procedures.

The modeling approaches span three levels of abstraction: (i) feature-based models

using hand-crafted acoustic and gaze descriptors, (ii) sequence-based models that operate on time-series data of speech and gaze, and (iii) representation-learning models based on pretrained speech embeddings.

5.3.1. Method 1: Feature-Based Prediction

Our first approach relies on extracting interpretable, hand-crafted features from the multimodal data streams. We hypothesize that the dynamics of the voice (prosody) and the eyes (gaze) towards the end of a turn contain critical cues for the upcoming transition.

5.3.1.1. Feature Extraction

For every turn, we extracted features from the **last 1 second** of the segment, enabling the model to focus on the final cues most relevant for transition signaling (Duncan 1972).

- **Acoustic Features (OpenSmile):** We utilized the opensmile toolkit to extract the **eGeMAPSv02** feature set (Eyben et al. 2010). This standard set includes 88 acoustic parameters designed for affective and paralinguistic analysis, covering frequency (F0), energy (loudness), spectral, and cepstral characteristics. We specifically used the `FeatureLevel.Functionals` to obtain statistical summaries (mean, variance, etc.) of these descriptors over the 1-second window.
- **Visual Features (Gaze):** We calculated the **proportion of gaze-on-face (GOF)** for both the speaker and the listener during the same 1-second window.

5.3.1.2. Models and Baselines

We evaluated the predictive power of these features using both a Logistic Regression (LR) and a feedforward neural network using a multilayer perceptron (MLP). The network consists of two hidden layers with 256 and 64 units, respectively, followed by a sigmoid output layer for binary classification. This architecture was chosen as a moderate-capacity model: large enough to capture non-linear interactions between acoustic and gaze features, but shallow enough to limit overfitting given the size of the dataset. To establish a performance floor, we first computed two baselines:

1. **Majority Baseline:** A model that always predicts the most frequent class (“Shift”).
2. **Random Baseline:** A model that predicts classes randomly based on the training set distribution.

We then trained separate models for different feature subsets to isolate the contribution of each modality:

- **F0-based Models:** Using only acoustic features.
- **Gaze-based Models:** Using only the gaze proportion features.
- **Combined Model:** Using both acoustic and visual features.

5.3.2. Method 2: LSTM-Based Prediction

While the feature-based approach collapses temporal dynamics into static statistical functionals (mean, variance), conversational signals unfold over time. To capture these temporal dependencies, our second approach employs Long Short-Term Memory (LSTM) networks (Greff et al. 2016), a type of Recurrent Neural Network (RNN) well-suited for sequence modeling.

5.3.2.1. Sequence Modeling

Instead of a single summary vector, the input to the model is a time-series sequence extracted from the **last 2 seconds** of the turn. This window was chosen to capture the immediate trajectory of prosodic and visual cues.

We implemented a sequence-based classifier using a unidirectional LSTM network. The model consists of a single LSTM layer with 50 hidden units, followed by a single fully connected classification layer.

We investigated three input configurations:

- **LSTM on F0:** The input is a univariate time series of fundamental frequency (F0), expressed in semitones and sampled at 100Hz . Each input sequence covered a fixed 2-second window preceding the end of the segment, resulting in sequences of length $T = 200$ with input shape 200×1 .
- **LSTM on Gaze:** At each time step, the input vector contained two values corresponding to the gaze of the child and the caregiver, respectively. Gaze was sampled at the same resolution as F0, yielding sequences of length $T = 200$ with input shape 200×1 .
- **LSTM Combined:** Acoustic and visual information were combined at the frame level. At each time step, the input to the LSTM was a three-dimensional feature vector containing F0 (in semitones) as well as the gaze of both the child and the caregiver. This resulted in sequences of length $T = 200$ with input shape 200×3 . This early-fusion representation preserves the alignment between the two modalities.

5.3.3. Method 3: Wav2Vec2 Representation

Our third method leverages modern transfer learning techniques for speech processing. We use **Wav2Vec 2.0**, a large-scale self-supervised model trained on vast amounts of unlabeled audio data (Baevski et al. 2020). Wav2Vec2 learns acoustic representations directly from the raw audio waveform, it captures phonetic and prosodic features that may be missed by traditional engineered features such as F0.

5.3.3.1. Deep Acoustic Embeddings

For this analysis, we process the **whole speech segment** of the turn using a pre-trained Wav2Vec2 model. Speech representations were extracted using Wav2Vec 2.0 Large (XLSR-53), specifically the `jonatasgrosmann/wav2vec2-large-xlsr-53-french`

model. The model is a language-specific fine-tuning of the Wav2Vec 2.0 Large architecture and retains the original 1024-dimensional hidden representation.

5.3.3.2. Ensemble Model: Late Fusion of Wav2Vec2 embeddings and Gaze Predictions

Wav2Vec 2.0 embeddings are high-dimensional, learned representations optimized for speech embeddings. On the other hand, gaze-on-face is a hand-crafted low-dimensional feature. We do not combine gaze and Wav2Vec embeddings at the feature level. Instead, we adopt a late-fusion ensemble that operates at the level of decisions rather than representations. Two independent base models are reused for this ensemble: the MLP trained on Wav2Vec2 embeddings and the MLP trained on gaze-on-face features introduced in 5.3.1. Rather than retraining or modifying these models, we take their output predictions as given. Each base model produces an estimate of the probability of Shift/Hold. These predictions are then combined using a logistic regression node.

5.4. Results

Models performance was evaluated on a held-out test set of 736 speech-segment transitions, unless otherwise specified. We report overall Accuracy, Macro Average F1-score, and Weighted Average F1-score for each model. Given the imbalance between Shift and Hold outcomes, interpretation focuses primarily on macro and weighted-balanced metrics rather than accuracy alone. Detailed class-wise classification reports for all models are provided in the Appendix A.

5.4.1. Overview of Model Performance

Table 5.1 summarizes the performance of all evaluated models. Several consistent patterns emerge across modeling approaches. First, we can clearly observe that high accuracy does not necessarily correspond to balanced performance across classes. In multiple cases, models achieve strong accuracy just by majorly predicting Shift, while performing poorly on Hold transitions. Second, Shift transitions are systematically easier to predict than Hold transitions across modeling approaches. Third, gaze-on-face features contribute mostly to balanced performance.

5.4.2. Baselines

The majority baseline achieves a high accuracy (0.74) by always predicting the most frequent class (Shift), but yields very low macro-average performance. This demonstrates that accuracy alone is not a meaningful indicator of success for this task. The random baseline performs substantially worse across all metrics. Turn-transition prediction is non-trivial.

Table 5.1. – Summary of model performance on the Hold/Shift prediction task.

Model	Accuracy	Macro Avg F1	Weighted Avg F1
Majority Baseline	0.74	0.42	0.63
Random Baseline	0.47	0.45	0.51
Acoustic (LR)	0.53	0.52	0.56
Gaze (MLP)	0.63	0.59	0.65
Acoustic + Gaze (MLP)	0.74	0.54	0.69
LSTM (F0)	0.54	0.52	0.60
LSTM (Gaze)	0.54	0.52	0.60
LSTM (Combined)	0.55	0.54	0.61
Wav2Vec2 (LR)	0.73	0.53	0.69
Wav2Vec2 (MLP)	0.73	0.60	0.72
Wav2Vec2 + Gaze (Ensemble)	0.73	0.62	0.72

5.4.3. Feature-Based Models

Feature-based models reveal clear differences between acoustic and visual cues. Models relying solely on acoustic features show limited overall performance, with moderate macro-average scores and relatively low accuracy. Models based on gaze-on-face features achieve comparatively higher macro-average F1 Scores. Combining acoustic and gaze features yields high accuracy. However, macro-average performance remains limited, reflecting a persistent class-wise performance asymmetry. This suggests that simple feature fusion primarily reinforces cues associated with speaker change, while cues for floor holding remain more difficult to capture.

5.4.4. Sequence-Based Models

Sequence-based models using LSTM do not substantially outperform feature-based approaches. Across acoustic-only, gaze-only, and combined configurations, macro-average and weighted F1-scores remain close to baseline models. These results indicate that, for the present dataset, explicitly modeling short-term temporal dynamics does not yield above baseline results, possibly due to noise, and the extremely limited size of the data.

5.4.5. Representation-Learning Models

Models based on pretrained Wav2Vec2 acoustic representations achieve high accuracy and weighted-average performance. Further incorporating gaze-on-face features alongside Wav2Vec embeddings using a late-fusion ensemble model yields the highest macro-average F1-score observed, while maintaining stable weighted-average performance. This suggests that gaze provides complementary information when strong

acoustic representations are available.

5.4.6. Summary of Results

Across all modeling approaches, Shift transitions are consistently easier to predict than Hold transitions. While accuracy and weighted-average scores can be high, macro-average performance reveals persistent difficulty in identifying Holds. Gaze-on-face features systematically contribute to improved balance between classes, particularly for Hold prediction, whereas increasing model complexity alone does not eliminate the asymmetry.

5.5. Discussion

This chapter addressed a narrowly defined question: Does gaze-on-face, alone or in combination with acoustic cues, contribute to predicting turn transitions (Hold vs. Shift) in naturalistic interaction? Turn-taking was operationalized as an event-based Hold/Shift prediction task over speech-segment transitions. It allowed a direct evaluation of how different representations support the distinction between floor holding and speaker change.

5.5.1. The task is structurally asymmetric: Shift is learnable; Hold is not as much

The results show that the Hold/Shift prediction task is fundamentally asymmetric. Across all modeling approaches, Shift transitions are substantially easier to predict than Hold transitions. This asymmetry is made clear by differences between accuracy and macro-average F1 performance. Models can achieve high accuracy and weighted-average F1 scores while performing poorly on the Hold class. The majority baseline illustrates this clearly: by always predicting Shift, it reaches high accuracy without ever predicting hold.

We interpret this asymmetry as a structural property of the turn-transition events themselves rather than a consequence of the features used for prediction. The temporal analysis presented earlier in 5.2.1 serves as a diagnostic illustration of this structure: very short gaps are concentrated around a subset of Shift outcomes, whereas Hold transitions are distributed across a broad and stable range of gap durations. What does this mean? It highlights an underlying asymmetry in how turn transitions unfold. While certain temporal configurations make speaker change more likely, no corresponding temporal configuration provides positive evidence for floor holding. The absence of a turn-yield signal does not amount to a turn-hold signal. Hold is not simply the complement of Shift, but an outcome that is intrinsically harder to predict from local cues.

5.5.2. Acoustic-only and gaze-only are not competing explanations; they fail in different ways

The feature-based results indicate that acoustic and gaze cues do not behave as interchangeable predictors. Models relying exclusively on acoustic features reach moderate overall performance, but exhibit a marked trade-off between classes (see [A](#)). This pattern indicates that acoustic hand-crafted features alone do not provide a reliable basis for distinguishing between speaker hold or shift under class imbalance.

By contrast, models based on gaze-on-face features achieve more balanced performance. The gaze-based neural network attains higher macro-average scores and improved Hold F1 relative to acoustic-only models. This indicates that gaze carries information that is particularly relevant to floor holding. Importantly, this improvement is not reflected solely in accuracy (see [A](#)).

5.5.3. Naive multimodal fusion does not solve the hard part — and that is informative

The combined acoustic and gaze feature model achieves high overall accuracy, matching the majority baseline. However, this performance is accompanied by a sharp decline in Hold recall. Rather than resolving the asymmetry between classes, simple feature concatenation exacerbates it. The model appears to prioritize the strongest and most frequent cues associated with speaker change, while the weaker and more context-dependent cues relevant to floor holding are ignored.

This outcome is consistent with the structure of the task. When one class is associated with clearer and more reliable signals and is also more frequent, the classifier is naturally encouraged to optimize performance on that class. The resulting model becomes a more confident predictor of Shift without improving its prediction of Hold. The failure of this naive combination of features is therefore informative. It indicates that Hold prediction cannot be recovered simply by adding more features, but requires a representation that preserves how cues are deployed over time and coordinated within an unfolding turn. Collapsing prosodic and gaze information into static averages removes precisely the temporal and interactional structure that interlocutors rely on to manage turn-taking, motivating the move toward sequence-based models.

5.5.4. Why LSTMs did not help: the bottleneck is not temporal modeling but available information

The sequence-based results further constrain the interpretation of the findings. Despite explicitly modeling temporal dynamics, LSTM-based models achieve performance only marginally above the random baseline. Accuracy remains close to chance level, and macro-average scores do not exceed those obtained with simple feature-based models, indicating that modeling short-window temporal dynamics alone does not provide meaningful additional information for distinguishing Hold from Shift transitions in this dataset.

This suggests that the limitation is not the absence of temporal modeling per se, but the information available in the modeled sequences. Given the size of the dataset and the variability across dyads, short-window temporal dynamics of pitch and gaze may be too noisy or underspecified to support robust learning of floor-holding behavior.

5.5.5. Deep acoustic representations improve the right thing when combined with gaze

The representation-learning results provide a more precise view of where additional information becomes usable. The logistic regression classifier trained on Wav2Vec2 embeddings achieves high accuracy and weighted-average performance, but remains weak on Hold transitions. This mirrors the behavior of several simpler models and indicates that deep acoustic representations alone do not solve the asymmetry.

The multi-layer perceptron improves macro-average performance, primarily by increasing recall and F1-score for the Hold class. This shows that information relevant to floor holding is present in the acoustic embeddings. The further improvement observed when gaze features are combined with Wav2Vec embeddings supports the conclusion that gaze contributes complementary information even in the presence of strong acoustic representations. Importantly, this gain appears in macro-average metrics rather than accuracy, reinforcing the interpretation that gaze primarily affects the difficult Hold cases.

5.5.6. What the results say about turn-taking rather than about models

Taken together, the results support a clear interpretation of turn-taking dynamics. Shift transitions are often marked by positive cues that are comparatively easy to detect, whereas Hold transitions are not consistently associated with a distinctive signal. This makes Hold inherently less predictable from local acoustic and gaze-on-face features alone.

From this perspective, the limited predictability of Hold transitions should not be treated as a modeling failure, but as an empirical result. These signals are clearly informative, but learning how the floor is sustained may require more data or longer stretches of interaction than are available in this corpus.

5.5.7. Contribution

This study makes three contributions. First, it shows that gaze-on-face provides measurable, non-redundant information for turn-transition prediction, particularly for the minority Hold class. Second, it demonstrates that high accuracy does not equate to successful modeling of turn-taking, as performance can be driven by class imbalance and Shift-focused prediction. Third, it identifies Hold prediction as the

5. Predicting Turn-Taking Dynamics in Middle Childhood – 5.5. Discussion

limiting case for event-based turn-taking models, motivating the need for richer contextual and interactional representations.

These conclusions provide the basis for the General Discussion, where the predictive findings of this chapter are integrated with the descriptive analyses of gaze-on-face in earlier chapters and considered in relation to theories of conversational development and multimodal coordination in conversation.

6. Conclusions

Table of contents

6.1. From Feasibility to Function: Revisiting the Central Questions	105
6.2. Establishing an Ecological Baseline for Middle Childhood Conversation	106
6.3. Automation and the Annotation Bottleneck Revisited	106
6.4. Gaze on Face as an Early-Stabilized Conversational Anchor	107
6.5. Role-Based Modulation of Gaze: Functional Deployment in Conversation	107
6.6. Gaze Dynamics	107
6.7. Multimodal Prediction of Turn-Taking: Insights and Limits	108
6.8. Implications for Theories of Conversational Development	108
6.9. Limitations	108
6.10. Future Directions	109
6.11. Conclusion	110

6.1. From Feasibility to Function: Revisiting the Central Questions

During this thesis, we set out to investigate how children in middle childhood deploy gaze in natural conversation. We ask whether their gaze behavior already exhibits the functional properties observed in adults. Motivated by a persistent gap in the developmental literature, this work focused on a period that is theoretically central yet empirically under-documented: the transition from early, scaffolded interaction to conversational competence.

To address this gap, the thesis adopted a deliberately ecological perspective. It did not ask how children behave under experimental constraints; it asked how they coordinate gaze with speech when interacting with a familiar partner, in a familiar environment, and within a task that preserves the reciprocity and contingency of everyday conversation. This methodological stance shaped both the design of the CHICA corpus and the analytical choices presented throughout the manuscript.

Across five chapters, the thesis progressed from methodological validation to increasingly fine-grained analyses of gaze behavior. The objective was not just to describe gaze patterns, but to understand their function within the interactional system.

This chapter synthesizes the findings in light of the six research questions introduced in Chapter 0. Specifically, it (i) evaluates the methodological contributions of the thesis, (ii) integrates the empirical results on gaze prevalence, modulation, and dynamics,

(iii) discusses their implications for theories in child development, and (iv) outlines limitations and directions for future research.

6.2. Establishing an Ecological Baseline for Middle Childhood Conversation

A first endeavor was to determine whether high-quality multimodal data could be collected and analyzed outside the laboratory without compromising analytical rigor (RQ1). Chapter 1 demonstrated that this is feasible. We used lightweight mobile eye-tracking and a carefully designed elicitation task. The CHICA corpus documented several hours of child–caregiver interaction with synchronized audio, video, gaze, and head-movement data in home environments.

These data proved not only descriptive but analytically exploitable. Despite the absence of strict experimental control, the recordings supported quantitative analyses at multiple scales, from global structure of the conversation to fine-grained coordination of gaze and speech. This directly challenges the assumption that ecological validity has to come at the expense of measurement precision.

Beyond that, and beyond feasibility, the corpus provides what been largely missing from the literature: an ecological baseline for conversation in middle childhood. Preliminary analyses already suggested a striking structure across ages 6 to 12, both in verbal participation and in coarse measures of gaze behavior. CHICA provided the necessary foundation for subsequent analyses addressing the subsequent research questions.

6.3. Automation and the Annotation Bottleneck Revisited

A second methodological question was how far non-verbal behavior can be reliably annotated using automated tools (RQ2). Chapter 2 addressed this challenge directly, revisiting the “annotation bottleneck” identified in the Introduction as a central obstacle to ecologically valid research.

The results point to a clear asymmetry across behaviors. Signals that are closely tied to facial morphology, such as smiling, remain reasonably prone to automatic annotation even outside controlled settings. Gaze, by contrast, proved far more difficult to estimate under such conditions in a reliable way.

The thesis advanced with a hybrid methodological position: we argue against treating automation as a general solution to the problem of annotation in ecological research. Instead, our work motivates a more selective and pragmatic use of automated tools.

This hybrid approach is not a compromise, but a methodological choice grounded in what the data actually allow, and it shaped how I carried out the remainder of the thesis.

6.4. Gaze on Face as an Early-Stabilized Conversational Anchor

With the methodological foundations established, the thesis turned to its central developmental questions. Chapter 3 examined the prevalence of gaze-on-face in middle childhood (RQ3). The analyses revealed a robust and consistent preference for face-directed gaze across all age groups studied.

This finding challenges accounts that posit a delayed or disrupted development of social gaze beyond early childhood. In ecologically valid settings, children as young as 7 reliably direct their gaze toward the interlocutor's face. The implication does not mean that development is complete at this age, it means that the baseline architecture of gaze is already in place.

Importantly, this result helps reconcile conflicting findings in the literature. Reduced face-directed gaze that laboratory studies have reported may reflect context in novel environments, unfamiliar interlocutors, or asymmetric relations, rather than intrinsic developmental behavior. When these constraints are eliminated, children's gaze behavior appears stable.

6.5. Role-Based Modulation of Gaze: Functional Deployment in Conversation

While Chapter 3 addressed the prevalence of gaze, Chapter 4 examined gaze as a resource modulated by role (RQ4). The analyses revealed apparent role-based differences: children gazed more at their interlocutor while listening than while speaking, mirroring well-established adult patterns.

This modulation supports theoretical accounts that treat gaze as serving a dual function in conversation. As listeners, children use gaze to acquire information and to monitor their partner for different communicative signals. As speakers, they reduce gaze to manage cognitive load associated with speech planning. The presence of this pattern in middle childhood indicates that children already integrate gaze into the organization of conversation.

The fact that this pattern appears in spontaneous interaction, rather than in an experimental task in a lab, makes it difficult for anyone to view it as a learned laboratory pattern. For that, we argue that it is a part of how children normally conduct conversation.

6.6. Gaze Dynamics

Beyond role-based differences, Chapter 4 further examined the temporal coordination between gaze and turn-taking events (RQ5). The analyses revealed systematic relationships between gaze behavior and moments of turn holding and turn yielding.

Children tended to avert gaze while maintaining the floor and to re-establish gaze when approaching turn completion.

These findings indicate that children do not rely solely on verbal or prosodic cues, they use gaze as part of a multimodal signaling system to manage the structure of the conversation.

This result has important theoretical implications. It suggests that the integration of gaze into turn-taking is not a late-acquired skill, but a core competence that is already operational in middle childhood.

6.7. Multimodal Prediction of Turn-Taking: Insights and Limits

Chapter 5 extended the thesis from descriptive analyses to prediction, using prediction not as an end in itself, but as a tool for probing the structure of multimodal coordination in conversation. We introduced an event-based prediction task centered on turn transitions, asking whether a completed speech segment would be followed by a continuation by the same speaker (Hold) or by a speaker change (Shift).

We evaluated 3 approaches, ranging from feature-based to sequence-based and representation-learning. Across these approaches, a consistent pattern emerged. Acoustic cues provided reliable information for predicting speaker change, while the inclusion of gaze-on-face systematically improved performance. This confirms that gaze contributes information that is relevant to turn management, beyond what is available from speech alone.

Taken together, these findings do clarify the role of gaze in conversation. Gaze contributes to making upcoming transitions legible, but not on its own.

6.8. Implications for Theories of Conversational Development

We adopt a view of middle childhood as a period of functional consolidation rather than foundational acquisition. The core mechanisms of gaze: face-directed, role-based modulation and turn-signaling are already in place by age 7. Development during this period concerns a gradual refinement.

We invite a shift in how middle childhood is understood in theories of conversational development. Competence is not simply acquired and then applied; it is refined by age.

6.9. Limitations

We acknowledge several limitations that should be considered when interpreting the findings of this thesis.

First, regarding sample size and demographic scope. Given the demands of collecting and annotating fully ecological data, the sample size (while appropriate for the depth of analysis pursued) remains modest at 15 dyads. This limits the statistical generalization of group-level comparisons, particularly across the three age groups (7, 9, and 11 years), individual variation may not be fully captured. Moreover, the corpus is composed exclusively of French-speaking, child–caregiver dyads recruited from a relatively homogeneous sociolinguistic context. I firmly believe that whether the gaze patterns documented here generalize across languages, cultures, or socioeconomic backgrounds remains an open question.

Second, regarding context. All recordings involve parent–child dyads, a configuration in which the adult partner naturally scaffolds the interaction. I do view this asymmetry as ecologically authentic because it reflects a common setting for children of this age. But it also means that the gaze behaviors observed may partly reflect the structure imposed by the caregiver. Peer interactions on the other hand, where participants have more symmetric roles would likely elicit different gaze coordination patterns.

Third, regarding the annotation pipeline. Although the hybrid approach combining automated tools (RetinaFace, WhisperX, Pupil Invisible gaze estimation) with manual correction proved effective, each component introduces its own set challenges. Voice Activity Detection for instance, identifies the presence of speech but does not distinguish between backchannels, feedbacks, or laughter. Similarly, face detection accuracy varies with head orientation and lighting conditions inherent to home environment. While we tried to mitigate these errors through manual verification, I acknowledge that they cannot be entirely eliminated.

Fourth, regarding the computational modeling. The predictive approaches explored in Chapter 5 were deliberately constrained, to serve as analytical tools rather than an engineering benchmark. The feature-based, sequence-based, and representation-learning models each captured complementary aspects of the gaze–speech relationship, but they do not exhaust the space of possible architectures and models. Furthermore, the models were trained and tested on pre-segmented speech, and did not attempt real-time, continuous predictions of turn transitions, that could be a closer proxy to the online demands of conversation.

Finally, this thesis focused on gaze-on-face as its primary unit of analysis. We ground this choice in the literature on social attention. However, it leaves aside other dimensions of gaze behavior, such as gaze toward objects, shared referents, and the broader visual environment.

6.10. Future Directions

The limitations outlined above naturally suggest several directions for future work.

A first extension concerns the context. I believe that peer interactions offer a greater symmetry to study gaze coordination, where turn-taking is negotiated more dynamically. Such data could reveal whether the patterns observed in parent–child interaction

reflect general conversational competence or are partly shaped by the asymmetric of conversation with an adult.

A second direction involves cross-linguistic and cross-cultural comparisons. The current corpus is based on French-speaking dyads, and one might reasonably ask to what extent the observed patterns reflect language-specific properties versus universal gaze behavior and competence. Extending the CHICA corpus to other languages and cultural settings would help disentangle these factors and establish more robust developmental baselines.

Third, longitudinal designs would strengthen the developmental claims of this work. The current thesis cannot track within-subject change over time. Following the same children over months or years would allow to observe the trajectory of gaze refinement more directly.

Fourth, the computational pipeline developed in this thesis could be extended in several ways. On the annotation side, advances in end-to-end speech processing—particularly models that jointly segment speech and classify conversational actions—could reduce the reliance on separate VAD and manual correction steps. On the modeling side, future work could explore architectures that operate on continuous streams rather than pre-segmented units, enabling real-time prediction of turn transitions. This would bring the computational framework closer to the online processing demands that actual interlocutors face.

Finally, extending the analysis beyond gaze-on-face to include gaze toward shared objects, gestures, and the broader visual scene would provide a more complete picture of how children coordinate attention in conversation. Such an extension would require richer annotation schemes, but it would align with the increasingly multimodal perspective of interaction research.

6.11. Conclusion

When children are observed in ecological settings, they appear far more conversationally competent than traditional laboratory paradigms suggest. By combining naturalistic data collection with computational analysis, the work fills a critical gap in our understanding of middle-childhood.

We argue for a shift in perspective: middle childhood should not be viewed as a period of incomplete conversational competence, but rather as a subtle refinement within an already functional multimodal system. Understanding this system requires methods that are as dynamic and context-sensitive as conversation itself, a challenge that this thesis met by design.

Bibliography

- [AD65] M. Argyle and J. Dean. “Eye-contact, distance and affiliation”. In: *Sociometry* (1965), pp. 289–304 (cit. on p. 25).
- [Ken67] A. Kendon. “Some functions of gaze-direction in social interaction”. In: *Acta psychologica* 26 (1967), pp. 22–63 (cit. on pp. 22, 23, 32, 35, 53, 66, 76–78, 83, 86, 93).
- [Yar67] A. L. Yarbus. *Eye movements and vision*. Springer, 1967 (cit. on p. 66).
- [Yng70] V. H. Yngve. “On Getting a Word in Edgewise”. In: *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society* (1970), pp. 567–578 (cit. on p. 21).
- [Dun72] S. Duncan. “Some signals and rules for taking speaking turns in conversations.” In: *Journal of personality and social psychology* 23.2 (1972), p. 283 (cit. on pp. 24, 93, 97).
- [Hym72] D. Hymes. “On Communicative Competence”. In: *Sociolinguistics: Selected Readings*. Ed. by J. B. Pride and J. Holmes. Harmondsworth, UK: Penguin, 1972, pp. 269–293 (cit. on p. 21).
- [BHL74] L. Bloom, L. Hood, and P. Lightbown. “Imitation in language development: If, when, and why”. In: *Cognitive psychology* 6.3 (1974), pp. 380–420 (cit. on p. 32).
- [SSJ74] H. Sacks, E. A. Schegloff, and G. Jefferson. “A simplest systematics for the organization of turn-taking for conversation”. In: *language* 50.4 (1974), pp. 696–735 (cit. on pp. 33, 81, 93).
- [Gri75] H. P. Grice. “Logic and conversation”. In: *Speech acts*. Brill, 1975, pp. 41–58 (cit. on p. 32).
- [KK75] E. O. Keenan and E. Klein. “Coherency in children’s discourse”. In: *Journal of psycholinguistic research* 4.4 (1975), pp. 365–380 (cit. on p. 32).
- [AC76] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge, UK: Cambridge University Press, 1976 (cit. on pp. 22, 32, 65).
- [MM76] H. McGurk and J. MacDonald. “Hearing lips and seeing voices”. In: *Nature* 264.5588 (1976), pp. 746–748 (cit. on p. 23).
- [SJS77] E. A. Schegloff, G. Jefferson, and H. Sacks. “The Preference for Self-Correction in the Organization of Repair in Conversation”. In: *Language* 53.2 (1977), pp. 361–382 (cit. on p. 21).

- [EF78] P. Ekman and W. V. Friesen. "Facial action coding system". In: *Environmental Psychology & Nonverbal Behavior* (1978) (cit. on p. 52).
- [Goo81] C. Goodwin. *Conversational Organization: Interaction Between Speakers and Hearers*. New York: Academic Press, 1981 (cit. on pp. 24, 93).
- [Kle86] C. L. Klenke. "Gaze and eye contact: a research review." In: *Psychological bulletin* 100.1 (1986), p. 78 (cit. on pp. 33, 66).
- [LK86] J. Local and J. Kelly. "Projection and 'silences': Notes on phonetic and conversational structure". In: *Human studies* (1986), pp. 185–204 (cit. on p. 93).
- [GA87] S. Garrod and A. Anderson. "Saying what you mean in dialogue: A study in conceptual and semantic co-ordination". In: *Cognition* 27.2 (1987), pp. 181–218 (cit. on p. 34).
- [HJ88] L. J. Hess and J. R. Johnston. "Acquisition of back-channel listener responses to adequate messages". In: *Discourse Processes* 11.3 (1988), pp. 319–335 (cit. on pp. 33, 35).
- [CB91] H. H. Clark and S. E. Brennan. "Grounding in communication." In: (1991) (cit. on p. 32).
- [Joh+91] M. H. Johnson, S. Dziurawiec, H. Ellis, et al. "Newborns' preferential tracking of face-like stimuli and its subsequent decline". In: *Cognition* 40.1-2 (1991), pp. 1–19 (cit. on p. 66).
- [Sch91] E. A. Schegloff. "Reflections on talk and social structure". In: *Talk and social structure* 984 (1991), pp. 44–70 (cit. on p. 33).
- [Sch92] E. A. Schegloff. "In another context". In: *Rethinking Context: Language as an Interactive Phenomenon*. Ed. by A. Duranti and C. Goodwin. Cambridge: Cambridge University Press, 1992, pp. 191–227 (cit. on p. 25).
- [SS92] S. Sehley and C. E. Snow. "The conversational skills of school-aged children". In: *Social Development* 1.1 (1992), pp. 18–35 (cit. on p. 33).
- [Lev93] W. J. Levelt. *Speaking: From intention to articulation*. MIT press, 1993 (cit. on p. 47).
- [ACM94] A. H. Anderson, H. H. Clark, and J. Mullin. "Interactive communication between children: Learning how to make language work in dialogue". In: *Journal of Child Language* 21.2 (1994), pp. 439–463 (cit. on pp. 33, 34).
- [Bar95] S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press, 1995 (cit. on p. 22).
- [Cla96] H. H. Clark. *Using Language*. Cambridge, UK: Cambridge University Press, 1996 (cit. on pp. 21, 32).
- [For96] C. E. Ford. "Interactional units in conversations: Syntactic, intonational, and pragmatic resources for the management of turns". In: *Interaction and grammar* (1996), pp. 134–184 (cit. on p. 93).

- [Sch+96] E. A. Schegloff et al. "Turn organization: One intersection of grammar and interaction". In: *Studies in interactional sociolinguistics* 13 (1996), pp. 52–133 (cit. on pp. 33, 94).
- [Car+98] M. Carpenter, K. Nagell, M. Tomasello, et al. "Social cognition, joint attention, and communicative competence from 9 to 15 months of age". In: *Monographs of the society for research in child development* (1998), pp. i–174 (cit. on pp. 24, 66).
- [GSR98] A. M. Glenberg, J. L. Schroeder, and D. A. Robertson. "Averting the gaze disengages the environment and facilitates remembering". In: *Memory & cognition* 26.4 (1998), pp. 651–658 (cit. on pp. 23, 83).
- [LM99] C. R. Lansing and G. W. McConkie. "Attention to Facial Regions in Segmental and Prosodic Visual Speech Perception Tasks". In: *Journal of Speech, Language, and Hearing Research* 42.3 (1999), pp. 526–539 (cit. on p. 22).
- [Bat+00] A. Batki, S. Baron-Cohen, S. Wheelwright, et al. "Is there an innate gaze module? Evidence from human neonates". In: *Infant Behavior and Development* 23.2 (2000), pp. 223–229 (cit. on p. 24).
- [BCJ00] J. B. Bavelas, L. Coates, and T. Johnson. "Listeners as Co-Narrators". In: *Journal of Personality and Social Psychology* 79.6 (2000), pp. 941–952 (cit. on pp. 21, 32).
- [CK04] H. H. Clark and M. A. Krych. "Speaking while monitoring addressees for understanding". In: *Journal of memory and language* 50.1 (2004), pp. 62–81 (cit. on p. 65).
- [PG04] M. J. Pickering and S. Garrod. "Toward a Mechanistic Psychology of Dialogue". In: *Behavioral and Brain Sciences* 27.2 (2004), pp. 169–190 (cit. on p. 21).
- [Pur04] M. R. J. Purver. "The theory and use of clarification requests in dialogue". PhD thesis. University of London King's College, 2004 (cit. on p. 32).
- [SBS04] J.-L. Schwartz, F. Berthommier, and C. Savariaux. "Seeing to hear better: evidence for early audio-visual interactions in speech identification". In: *Cognition* 93.2 (2004), B69–B78 (cit. on p. 65).
- [AP05] D. R. Anderson and T. A. Pempek. "Television and very young children". In: *American behavioral scientist* 48.5 (2005), pp. 505–522 (cit. on p. 48).
- [BM05] R. Brooks and A. N. Meltzoff. "The development of gaze following and its relation to language". In: *Developmental science* 8.6 (2005), pp. 535–543 (cit. on p. 66).
- [Wit+06] P. Wittenburg, H. Brugman, A. Russel, et al. "ELAN: A professional framework for multimodality research". In: *5th international conference on language resources and evaluation (LREC 2006)*. 2006, pp. 1556–1559 (cit. on p. 42).

- [Sch07] E. A. Schegloff. *Sequence organization in interaction: A primer in conversation analysis I*. Vol. 1. Cambridge university press, 2007 (cit. on p. 94).
- [MGP08] B. Maroni, A. Gnisci, and C. Pontecorvo. “Turn-taking in classroom interactions: Overlapping, interruptions and pauses in primary school”. In: *European Journal of Psychology of Education* 23 (2008), pp. 59–76 (cit. on p. 33).
- [BH10] E. Baines and C. Howe. “Discourse topic management skills in 4-, 6- and 9-year-old peer interactions: Developments with age and the effects of task context”. In: *First Language* 30.3–4 (2010), pp. 508–535 (cit. on p. 33).
- [EWS10] F. Eyben, M. Wöllmer, and B. Schuller. “Opensmile: the munich versatile and fast open-source audio feature extractor”. In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 1459–1462 (cit. on p. 97).
- [Hro+10] D. D. Hromada, C. Tijus, S. Poitrenaud, et al. “Zygomatic Smile Detection: The Semi-Supervised Haar Training of a Fast and Frugal System: A Gift to OpenCV Community”. In: *2010 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*. 2010, pp. 1–5 (cit. on p. 53).
- [Van+10] K. J. Van Engen, M. Baese-Berk, R. E. Baker, et al. “The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles”. In: *Language and speech* 53.4 (2010), pp. 510–540 (cit. on p. 34).
- [GH11] A. Gravano and J. Hirschberg. “Turn-taking cues in task-oriented dialogue”. In: *Computer Speech & Language* 25.3 (2011), pp. 601–634 (cit. on p. 93).
- [Lai+11] K. E. Laidlaw, T. Foulsham, G. Kuhn, et al. “Potential social interactions are important to social attention”. In: *Proceedings of the National Academy of Sciences* 108.14 (2011), pp. 5548–5553 (cit. on p. 66).
- [DRW12] G. Doherty-Sneddon, D. M. Riby, and L. Whittle. “Gaze aversion as a cognitive load management strategy in autism spectrum disorder and Williams syndrome”. In: *Journal of Child Psychology and Psychiatry* 53.4 (2012), pp. 420–430 (cit. on p. 23).
- [Cui+13] Z. Cui, W. Li, D. Xu, et al. “Fusing robust face region descriptors via multiple metric learning for face recognition in the wild”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3554–3561 (cit. on p. 52).
- [YS13] C. Yu and L. B. Smith. “Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination”. In: *PloS one* 8.11 (2013), e79659 (cit. on p. 24).
- [Gib14] J. J. Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014 (cit. on p. 22).

- [Mac14] B. MacWhinney. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press, 2014 (cit. on p. 36).
- [SHO14] G. Skantze, A. Hjalmarsson, and C. Oertel. “Turn-taking, feedback and joint attention in situated human–robot interaction”. In: *Speech Communication* 65 (2014), pp. 50–66 (cit. on p. 93).
- [Ain+15] M. D. S. Ainsworth, M. C. Blehar, E. Waters, et al. *Patterns of attachment: A psychological study of the strange situation*. Psychology press, 2015 (cit. on p. 25).
- [DTE15] M. Dingemanse, F. Torreira, and N. J. Enfield. “Is “Huh?” a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items”. In: *PLOS ONE* 10.11 (2015), e0139804 (cit. on pp. 21, 32).
- [DF15] S. Duncan and D. W. Fiske. *Face-to-face interaction: Research, methods, and theory*. Routledge, 2015 (cit. on pp. 53, 66, 86).
- [GKR15] M. S. Gobel, H. S. Kim, and D. C. Richardson. “The dual function of social gaze”. In: *Cognition* 136 (2015), pp. 359–364 (cit. on pp. 22, 65, 66).
- [HFK15] S. Ho, T. Foulsham, and A. Kingstone. “Speaking and listening with the eyes: Gaze signaling during dyadic interactions”. In: *PloS one* 10.8 (2015), e0136905 (cit. on pp. 23, 66, 77, 78, 82, 86, 88).
- [Jun+15] H. Jung, S. Lee, J. Yim, et al. “Joint fine-tuning in deep neural networks for facial expression recognition”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2983–2991 (cit. on p. 52).
- [KBB15] J. Kennedy, P. Baxter, and T. Belpaeme. “Comparing robot embodiments in a guided discovery learning interaction with children”. In: *International Journal of Social Robotics* 7.2 (2015), pp. 293–308 (cit. on p. 26).
- [LT15] S. C. Levinson and F. Torreira. “Timing in turn-taking and its implications for processing models of language”. In: *Frontiers in psychology* 6 (2015), p. 731 (cit. on pp. 93, 94).
- [Zha+15] Z. Zhang, P. Luo, C.-C. Loy, et al. “Learning social relation traits from face images”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3631–3639 (cit. on p. 52).
- [BRM16] T. Baltrušaitis, P. Robinson, and L.-P. Morency. “Openface: an open source facial behavior analysis toolkit”. In: *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2016, pp. 1–10 (cit. on pp. 43, 52, 56).
- [CBC16] M. Casillas, S. C. Bobb, and E. V. Clark. “Turn-taking, timing, and planning in early language acquisition”. In: *Journal of child language* 43.6 (2016), pp. 1310–1337 (cit. on pp. 32, 33, 48).
- [Che+16] J. Chen, Z. Chen, Z. Chi, et al. “Facial expression recognition in video with multiple feature fusion”. In: *IEEE Transactions on Affective Computing* 9.1 (2016), pp. 38–50 (cit. on p. 52).

- [Gre+16] K. Greff, R. K. Srivastava, J. Koutník, et al. “LSTM: A search space odyssey”. In: *IEEE transactions on neural networks and learning systems* 28.10 (2016), pp. 2222–2232 (cit. on p. 98).
- [Lev16] S. C. Levinson. *Turn-Taking in Human Communication: Origins and Implications for Language Processing*. Cambridge, MA: MIT Press, 2016 (cit. on pp. 21, 32, 93).
- [RRK16] E. F. Risko, D. C. Richardson, and A. Kingstone. “Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze”. In: *Current Directions in Psychological Science* 25.1 (2016), pp. 70–74 (cit. on pp. 33, 34, 65).
- [Wan+16] Z. Wang, R. T. Devine, K. K. Wong, et al. “Theory of mind and executive function during middle childhood across cultures”. In: *Journal of experimental child psychology* 149 (2016), pp. 6–22 (cit. on pp. 24, 33).
- [Ska17] G. Skantze. “Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks”. In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 2017, pp. 220–230 (cit. on p. 93).
- [XHW17] Y. Xia, D. Huang, and Y. Wang. “Detecting Smiles of Young Children via Deep Transfer Learning”. In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2017, pp. 1673–1681 (cit. on p. 54).
- [Bel+18] T. Belpaeme, J. Kennedy, A. Ramachandran, et al. “Social robots for education: A review”. In: *Science robotics* 3.21 (2018), eaat5954 (cit. on p. 26).
- [MBA18] D. Matthews, H. Biney, and K. Abbot-Smith. “Individual differences in children’s pragmatic ability: A review of associations with formal language, social cognition, and executive functions”. In: *Language Learning and Development* 14.3 (2018), pp. 186–223 (cit. on pp. 24, 33).
- [HL19] J. Holler and S. C. Levinson. “Multimodal language processing in human communication”. In: *Trends in cognitive sciences* 23.8 (2019), pp. 639–652 (cit. on p. 21).
- [Bae+20] A. Baevski, Y. Zhou, A. Mohamed, et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: *Advances in neural information processing systems* 33 (2020), pp. 12449–12460 (cit. on p. 98).
- [Bre+20] H. Bredin, R. Yin, J. M. Coria, et al. “Pyannote. audio: neural building blocks for speaker diarization”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7124–7128 (cit. on p. 43).
- [Den+20] J. Deng, J. Guo, E. Ververas, et al. “Retinaface: Single-shot multi-level face localisation in the wild”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5203–5212 (cit. on pp. 27, 55, 68).

- [Hal+20] J. Hale, J. A. Ward, F. Buccheri, et al. “Are you on my wavelength? Interpersonal coordination in dyadic conversations”. In: *Journal of nonverbal behavior* 44.1 (2020), pp. 63–83 (cit. on p. 35).
- [Hes20] R. S. Hessels. “How does gaze to faces support face-to-face interaction? A review and perspective”. In: *Psychonomic bulletin & review* 27.5 (2020), pp. 856–881 (cit. on p. 65).
- [LD20] S. Li and W. Deng. “Deep facial expression recognition: A survey”. In: *IEEE transactions on affective computing* 13.3 (2020), pp. 1195–1215 (cit. on p. 52).
- [TBD20] M. Tonsen, C. K. Baumann, and K. Dierkes. “A high-level description and performance evaluation of pupil invisible”. In: *arXiv preprint arXiv:2009.00508* (2020) (cit. on pp. 35, 67).
- [Bod+21] K. Bodur, M. Nikolaus, F. Kassim, et al. “Chico: A multimodal corpus for the study of child conversation”. In: *Companion publication of the 2021 international conference on multimodal interaction*. 2021, pp. 158–163 (cit. on pp. 36, 44, 54).
- [Che+21] Y. Cheng, H. Wang, Y. Bao, et al. *Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark*. 2021. arXiv: 2104.12668 [cs.CV] (cit. on p. 53).
- [De +21] M. De Lillo, R. Foley, M. C. Fysh, et al. “Tracking developmental differences in real-world social attention across adolescence, young adulthood and older adulthood”. In: *Nature human behaviour* 5.10 (2021), pp. 1381–1390 (cit. on pp. 24, 35, 66, 75, 77).
- [DK21] S. Donnelly and E. Kidd. “The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development”. In: *Child Development* 92.2 (2021), pp. 609–625 (cit. on p. 32).
- [Mas+21] L. R. Masek, B. T. M. McMillan, S. J. Paterson, et al. “Where language meets attention: How contingent interactions promote learning”. In: *Developmental Review* 60 (2021), p. 100961 (cit. on p. 32).
- [PG21] M. J. Pickering and S. Garrod. *Understanding dialogue: Language use and social interaction*. Cambridge University Press, 2021 (cit. on p. 32).
- [Sul+21] J. Sullivan, M. Mei, A. Perfors, et al. “SAYCam: A large, longitudinal audio-visual dataset recorded from the infant’s perspective”. In: *Open mind* 5 (2021), pp. 20–29 (cit. on p. 36).
- [ES22] E. Ekstedt and G. Skantze. “Voice activity projection: Self-supervised learning of turn-taking events”. In: *arXiv preprint arXiv:2205.09812* (2022) (cit. on p. 93).
- [Gho+22] S. Ghosh, A. Dhall, M. Hayat, et al. *Automatic Gaze Analysis: A Survey of Deep Learning based Approaches*. 2022. arXiv: 2108.05479 [cs.CV] (cit. on p. 53).

- [Ngu+22] V. Nguyen, O. Versyp, C. Cox, et al. “A systematic review and Bayesian meta-analysis of the development of turn taking in adult–child vocal interactions”. In: *Child Development* 93.4 (2022), pp. 1181–1200 (cit. on p. 32).
- [Abb+23] K. Abbot-Smith, J. Dockrell, A. Sturrock, et al. “Topic maintenance in social conversation: What children need to learn and evidence this can be taught”. In: *First Language* 43.6 (2023), pp. 614–642 (cit. on p. 32).
- [Bai+23] M. Bain, J. Huh, T. Han, et al. “Whisperx: Time-accurate speech transcription of long-form audio”. In: *arXiv preprint arXiv:2303.00747* (2023) (cit. on pp. 27, 42, 79).
- [Bod+23] K. Bodur, M. Nikolaus, L. Prévot, et al. “Using video calls to study children’s conversational development: the case of backchannel signaling”. In: *Frontiers in Computer Science* 5 (2023), p. 1088752 (cit. on pp. 32, 33, 44, 54).
- [Did+23] C. Dideriksen, M. H. Christiansen, K. Tylén, et al. “Quantifying the interplay of conversational devices in building mutual understanding.” In: *Journal of Experimental Psychology: General* 152.3 (2023), p. 864 (cit. on p. 33).
- [Elm+23] S. L. Elmlinger, J. A. Schwade, L. Vollmer, et al. “Learning how to learn from social feedback: The origins of early vocal development”. In: *Developmental Science* 26.2 (2023), e13296 (cit. on p. 32).
- [Ere+23] Y. Erel, K. A. Shannon, J. Chu, et al. “iCatcher+: Robust and automated annotation of infants’ and young children’s gaze behavior from videos collected in laboratory, field, and online studies”. In: *Advances in methods and practices in psychological science* 6.2 (2023), p. 25152459221147250 (cit. on pp. 43, 53).
- [Fus+23] R. Fusaroli, E. Weed, R. Rocca, et al. “Caregiver linguistic alignment to autistic and typically developing children: A natural language processing approach illuminates the interactive components of language development”. In: *Cognition* 236 (2023), p. 105422 (cit. on p. 32).
- [HH23] A. F. d. C. Hamilton and J. Holler. “Face2face: advancing the science of social interaction”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 378.1875 (2023), p. 20210470 (cit. on p. 53).
- [NF23] M. Nikolaus and A. Fourtassi. “Communicative feedback in language acquisition”. In: *New Ideas in Psychology* 68 (2023), p. 100985 (cit. on p. 32).
- [Rad+23] A. Radford, J. W. Kim, T. Xu, et al. “Robust speech recognition via large-scale weak supervision”. In: *International conference on machine learning*. PMLR. 2023, pp. 28492–28518 (cit. on pp. 42, 43).

- [SY23] S. E. Schroer and C. Yu. “Looking is not enough: Multimodal attention supports the real-time learning of new words”. In: *Developmental Science* 26.2 (2023), e13290 (cit. on pp. 35, 66).
- [DE24] M. Dingemanse and N. J. Enfield. “Interactive repair and the foundations of language”. In: *Trends in Cognitive Sciences* 28.1 (2024), pp. 30–42 (cit. on p. 32).
- [Von+24] W. K. Vong, W. Wang, A. E. Orhan, et al. “Grounded language acquisition through the eyes and ears of a single child”. In: *Science* 383.6682 (2024), pp. 504–511 (cit. on p. 36).
- [Gu+25] Y. Gu, E. Donnellan, B. Grzyb, et al. “The ECOLANG Multimodal Corpus of adult-child and adult-adult Language”. In: *Scientific data* 12.1 (2025), p. 89 (cit. on p. 36).
- [MPL] MPLab. *The MPLab GENKI Database, GENKI-4K Subset*. <http://mplab.ucsd.edu> (cit. on p. 53).

ANNEXES

A. Complementary Results: Classification Reports for Chapter 5

This appendix provides detailed class-wise classification reports for all models evaluated in Chapter 5. These tables complement the summary results reported in Table 5.1.

A.1. Baselines

Table A.1. – Classification report for the majority baseline.

Class	Precision	Recall	F1-Score	Support
0 (Hold)	0.00	0.00	0.00	192
1 (Shift)	0.74	1.00	0.85	544
Accuracy			0.74	736
Macro Avg	0.37	0.50	0.42	736
Weighted Avg	0.55	0.74	0.63	736

Table A.2. – Classification report for the random baseline.

Class	Precision	Recall	F1-Score	Support
0 (Hold)	0.25	0.50	0.33	192
1 (Shift)	0.72	0.47	0.57	544
Accuracy			0.47	736
Macro Avg	0.49	0.48	0.45	736
Weighted Avg	0.60	0.47	0.51	736

A.2. Feature-Based Models

Table A.3. – Classification report for the Feature-Based Logistic Regression model.

Class	Precision	Recall	F1-Score	Support
0 (Hold)	0.32	0.71	0.44	192
1 (Shift)	0.82	0.47	0.60	544
Accuracy			0.53	736
Macro Avg	0.57	0.59	0.52	736
Weighted Avg	0.69	0.53	0.56	736

Table A.4. – Classification report for the gaze MLP model.

Class	Precision	Recall	F1-Score	Support
0 (Hold)	0.37	0.61	0.46	192
1 (Shift)	0.82	0.64	0.72	544
Accuracy			0.63	736
Macro Avg	0.60	0.62	0.59	736
Weighted Avg	0.70	0.63	0.65	736

Table A.5. – Classification report for the combined features MLP model.

Class	Precision	Recall	F1-Score	Support
0 (Hold)	0.49	0.16	0.24	192
1 (Shift)	0.76	0.94	0.84	544
Accuracy			0.74	736
Macro Avg	0.63	0.55	0.54	736
Weighted Avg	0.69	0.74	0.69	736

A.3. Sequence-Based Models

Table A.6. – Classification reports for LSTM models

Model	Class	Precision	Recall	F1-Score	Support
LSTM (F0)	0 (Hold)	0.35	0.40	0.37	192
	1 (Shift)	0.76	0.59	0.66	544
	Accuracy			0.54	736
LSTM (Gaze)	0 (Hold)	0.34	0.42	0.38	192
	1 (Shift)	0.77	0.58	0.66	544
	Accuracy			0.54	736
LSTM (Combined)	0 (Hold)	0.36	0.44	0.40	192
	1 (Shift)	0.78	0.59	0.67	544
	Accuracy			0.55	736

A.4. Representation-Learning Models (Wav2Vec2)

Table A.7. – Classification report for Wav2Vec2 (LR).

Class	Precision	Recall	F1-Score	Support
0 (Hold)	0.37	0.17	0.23	148
1 (Shift)	0.77	0.91	0.83	463
Accuracy			0.73	611
Macro Avg	0.57	0.54	0.53	611
Weighted Avg	0.68	0.73	0.69	611

Table A.8. – Classification report for Wav2Vec2 (MLP).

Class	Precision	Recall	F1-Score	Support
0 (Hold)	0.43	0.34	0.38	148
1 (Shift)	0.80	0.85	0.83	463
Accuracy			0.73	611
Macro Avg	0.62	0.60	0.60	611
Weighted Avg	0.71	0.73	0.72	611

Table A.9. – Classification report for the ensemble of models.

Class	Precision	Recall	F1-Score	Support
0 (Hold)	0.44	0.41	0.42	148
1 (Shift)	0.81	0.83	0.82	463
Accuracy			0.73	611
Macro Avg	0.62	0.62	0.62	611
Weighted Avg	0.72	0.73	0.72	611