Communicative Feedback in Language Acquisition

Mitja Nikolaus

Aix*Marseille université
Socialement engagée

NNT : 2023AIXM0001

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 27 Octobre 2023 par

Mitja NIKOLAUS

Communicative Feedback in Language Acquisition

Discipline
Sciences Cognitives

École doctorale
ED 356 COGNITION, LANGAGE, EDUCATION

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

Composition du jury

Raquel FERNÁNDEZ Professor, University of Amsterdam	Rapporteure
Gabriella VIGLIOCCO Professor, University College London (UCL)	Rapporteure
Okko RÄSÄNEN Associate Professor, Tampere University	Examinateur
Philippe BLACHE Directeur de recherche CNRS, Aix-Marseille University	Président du jury
Laurent PRÉVOT Professeur, Aix-Marseille University	Directeur de thèse
Abdellah FOURTASSI Maître de conférences, Aix-Marseille University	Directeur de thèse

ILCB Institute of Language, Communication and the Brain
Institut Archimède Mathématiques-Informatique Aix*Marseille Université
Laboratoire Parole et Langage 1972 50 2022
LIS Laboratoire d'Informatique & Systèmes UMR 7020

Affidavit

I, undersigned, Mitja Nikolaus, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Laurent Prévot and Abdellah Fourtassi, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the french national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

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

Marseille, 01 June 2023

List of publications and conference participations

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

1. Nikolaus, M.*, Maes, J.*, Auguste, J., Prévot, L., & Fourtassi, A. (2021). **Large-scale study of speech acts' development using automatic labelling.** *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (*Joint First Authors)
2. Nikolaus, M., Maes, J., & Fourtassi, A. (2021). **Modeling speech act development in early childhood : The role of frequency and linguistic cues.** *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*
3. Nikolaus, M., & Fourtassi, A. (2021). **Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in Artificial Neural Networks.** *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 200–210
4. Bodur, K., Nikolaus, M., Kassim, F., Prévot, L., & Fourtassi, A. (2021). **ChiCo : A Multimodal Corpus for the Study of Child Conversation.** *Proceedings of the 23rd International Workshop on Corpora and Tools for Social Skills Annotation (ICMI)*, 158–163.
5. Nikolaus, M., & Fourtassi, A. (2021). **Modeling the Interaction Between Perception-Based and Production-Based Learning in Children's Early Acquisition of Semantic Knowledge.** *Proceedings of the 25th Conference on Computational Natural Language Learning*, 391–407.
6. Bodur, K., Nikolaus, M., Prevot, L., & Fourtassi, A. (2022). **Backchannel Behavior in Child-Caregiver Video Calls.** *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.*
7. Nikolaus, M., Prévot, L., & Fourtassi, A. (2022). **Communicative Feedback as a Mechanism Supporting the Production of Intelligible Speech in Early Childhood.** *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.*
8. Nikolaus, M., Alishahi, A., & Chrupała, G. (2022). **Learning English with Peppa Pig.** *Transactions of the Association for Computational Linguistics*
9. Nikolaus, M., Salin, E., Ayache, S., Fourtassi, A., Favre, B. (2022) **Do Vision-and-Language Transformers Learn Grounded Predicate-Noun Dependencies?** *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. (EMNLP)*

10. Liu, J. Nikolaus, M., Bodur, K. Fourtassi, A. (2022) **Predicting Backchannel Signaling in Child-Caregiver Multimodal Conversations.** *Proceedings of the International Conference on Multimodal Interaction (ICMI'22 Companion)*
11. Nikolaus, M., Maes, J., Auguste, J., Prévot, L., & Fourtassi, A. (2022). **Large-scale study of speech acts' development in early childhood.** *Language Development Research*
12. Bodur, K., Nikolaus, M., Prevot, L., & Fourtassi, A. (2023). **Using video calls to study children's conversational development : The case of backchannel signaling** *Frontiers in Computer Science*
13. Nikolaus, M. Fourtassi, A. (2023). **Communicative Feedback in Language Acquisition.** *New Ideas in Psychology*
14. Cabiddu, F., Nikolaus, M., & Fourtassi, A. (2023). **Comparing Children and Large Language Models in Word Sense Disambiguation : Insights and Challenges** *Proceedings of the 45th Annual Meeting of the Cognitive Science Society.*
15. Nikolaus, M., Prévot, L., & Fourtassi, A. (2023). **Communicative Feedback in Response to Children's Grammatical Errors** *Proceedings of the 45th Annual Meeting of the Cognitive Science Society.*

Participation at conferences and summer schools :

1. ILCB Summer School 2020
2. ILCB Summer School 2021
3. ILCB Summer School 2022
4. CogSci 2021
5. CogSci 2022
6. CogSci 2023
7. NAACL 2021
8. EMNLP 2021
9. ALPS Winter School 2022
10. HSP 2022
11. LREC 2022
12. Lisbon Machine Learning Summer School (LxMLS) 2023
13. Amlap 2023

Résumé

Bien avant que leurs compétences linguistiques ne soient pleinement développées, les enfants participent à des échanges conversationnels. Cela leur permet d'expérimenter avec leurs connaissances linguistiques émergentes et de recevoir un feedback communicatif de la part de leurs interlocuteurs. En s'inspirant des théories de la coordination communicative, nous pouvons formaliser un nouveau mécanisme d'acquisition des langages : Les enfants peuvent améliorer leurs connaissances linguistiques au cours d'une conversation en exploitant les signaux explicites ou implicites de réussite ou de rupture de la communication.

À partir de cette hypothèse, nous menons deux études de corpus qui soulignent le rôle du feedback communicatif en tant que mécanisme soutenant la production d'un langage intelligible, ainsi que l'acquisition de la grammaire de la langue maternelle. Enfin, nous concevons et évaluons des modèles computationnels qui instancient un mécanisme d'apprentissage basé sur le feedback en plus de l'apprentissage statistique et nous démontrons que ce feedback peut améliorer l'acquisition de la sémantique.

Le feedback communicatif fournit un cadre commun à plusieurs lignes de recherche sur le développement de l'enfant et nous permettra d'obtenir une compréhension plus complète de l'acquisition du langage au sein et à travers l'interaction sociale.

Mots clés : acquisition du langage, communication, conversation, feedback communicatif, étude de corpus, modélisation computationnelle

Abstract

Children start to communicate and use language in social interactions from a very young age. This allows them to experiment with their developing linguistic knowledge and receive valuable feedback from their interlocutors. While research in language acquisition has focused a great deal on children's ability to learn from the linguistic input or social cues, little work, in comparison, has investigated the nature and role of Communicative Feedback, a process that results from children and caregivers trying to coordinate mutual understanding. By drawing on insights from theories of communicative coordination we can formalize a new mechanism for language acquisition: We argue that children can improve their linguistic knowledge in conversation by leveraging explicit or implicit signals of communication success or failure.

Based on this hypothesis, we conducted two corpus studies that highlight the role of Communicative Feedback as a mechanism supporting the production of intelligible speech, as well as the acquisition of the grammar of one's native language. Finally, we design and evaluate computational models that instantiate a feedback-based learning mechanism in addition to statistical learning and demonstrate that such feedback can improve the acquisition of semantics.

Communicative Feedback provides a common framework for several lines of research in child development and will enable us to obtain a more complete understanding of language acquisition within and through social interaction.

Keywords: language acquisition, communication, conversation, communicative feedback, corpus study, computational modeling

Contents

List of Figures

List of Tables

Part I.

Communicative Feedback

Summary

1. Communicative Feedback in Language Acquisition

This chapter is based on the article "Communicative Feedback in Language Acquisition" (Nikolaus and Fourtassi 2023), published in *New Ideas In Psychology*.

This chapter provides the introduction as well as the theoretical framework for this thesis. We define Communicative Feedback and describe a framework for categorizing feedback signals based on their explicitness and valence. Further, we describe how, in principle, it can support child language acquisition on multiple levels: The learning of linguistic *form*, *meaning*, as well as language *use*.

In the following, we reviewed articles that fit into this framework by categorizing them according to their focus on the different explicit and implicit feedback signals. Based on this review, we identified several directions for future research, some of which have been addressed in the following parts of this thesis (Parts II and III). Several other topics still remain open for further investigation in future work, as discussed in the last part of this thesis (Part IV).

1.1. Abstract

Children start to communicate and use language in social interactions from a very young age. This allows them to experiment with their developing linguistic knowledge and receive valuable *feedback* from their – often more knowledgeable – interlocutors. While research in language acquisition has focused a great deal on children's ability to learn from the linguistic input or social cues, little work, in comparison, has investigated the nature and role of communicative feedback, a process that results from children and caregivers trying to coordinate mutual understanding.

In this work, we draw on insights from theories of communicative coordination to formalize a mechanism for language acquisition: We argue that children can improve their linguistic knowledge in conversation by leveraging explicit or implicit signals of communication success or failure. This new formalization provides a common framework for several lines of research in child development that have been pursued separately. Further, it points towards several gaps in the literature that, we believe, should be addressed in future research in order to achieve a more complete understanding of language acquisition within and through social interaction.

1.2. Introduction

Research in language acquisition has extensively documented the impressive skills children use to learn from the properties of the language they hear around them (Saffran, Aslin, and Newport 1996) together with the properties of their visual environment (L. Smith and Yu 2008). Such multimodal input is, however, not the only source of information available to children. In particular, children start to actively interact with people very early in development. This early social interaction has long been considered to play an important role in the acquisition of language (e.g., Bruner 1985; Ninio and C. Snow 1988; Tomasello 2003; Kuhl 2007; Eve V Clark 2016; Eve V. Clark 2018; Matthews 2014; Vygotsky 1962; Yurovsky 2018).

The current dominant line of research studying the role of social interaction focuses on children's ability to make inferences about people's communicative intents. For example, when a – more knowledgeable – adult introduces a novel word in an ambiguous context where there are many objects, children have to infer which precise object the adult meant. To make a successful pragmatic inference, children can take into account the context of language use, common ground with the interlocutor, as well as social cues provided by the latter such as gaze and pointing (Tomasello, Carpenter, Call, et al. 2005; Senju and Csibra 2008; Yurovsky and Michael C. Frank 2017; Bohn and Michael C. Frank 2019; Tsuji, Jincho, Mazuka, et al. 2020).

In the current work, we examine the role of another aspect of social interaction in language learning, involving not only pragmatic inference over what the speaker has said or done, but also the explicit negotiation of shared understanding with the interlocutor. Indeed, children start communicating long before their linguistic skills are mature (Bates, Camaioni, and Volterra 1975; Ninio and C. Snow 1988; Halliday

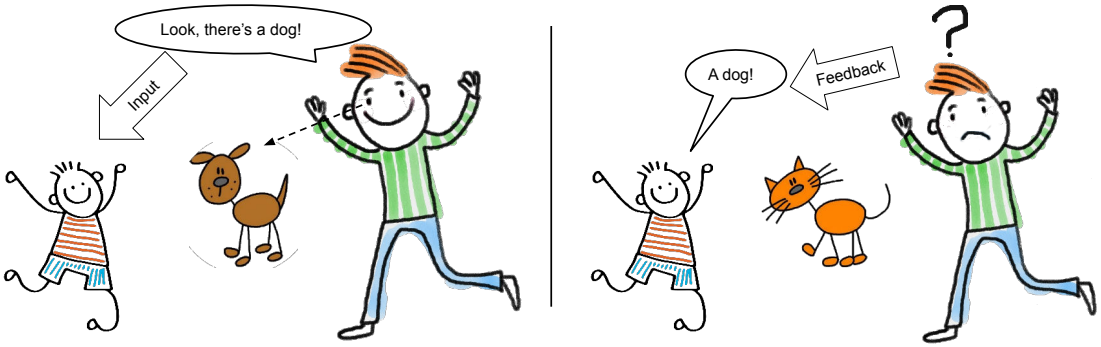

Figure 1.1. – Learning from input and learning from feedback. The child may learn from the linguistic input by listening to what is said and making pragmatic inference about what is meant (Left side: The child learns from the parent's utterances as well as the parent's eye gaze about the meaning of the word "dog"). The child can also learn from positive or negative feedback provided by interlocutors on their own communicative attempts (Right side: The child receives negative feedback (signals of non-understanding, in this case a puzzled face) for using the word "dog" when trying to talk about a cat).

1975; Eve V Clark 2016). Such early attempts at communication succeed at times, but they can also fail because children make phonological, syntactic, semantic, and pragmatic mistakes that impede the transmission of their true intents. In the context of these early conversations, children receive *feedback* from their interlocutors, signaling successful or unsuccessful communication which they can use to fine-tune their linguistic knowledge (see an illustration in Figure 1.1).

We call this mechanism **Communicative Feedback** (hereafter CF) for two reasons. First, to emphasize its link to general communicative principals in conversations – more studied in the adult literature – whereby interlocutors coordinate to understand each other (H. H. Clark 1996; Pickering and Garrod 2021). Second, to differentiate it from another form of feedback – more studied in the developmental literature – often under the name of *corrective* feedback, describing responses from caregivers that provide a correction for potential mistakes in children's utterances.

Corrective feedback has long been debated in the language acquisition literature, especially regarding the question of the learnability of grammar from negative evidence in addition to positive evidence (e.g., Gold 1967). Some researchers questioned its availability or usefulness (e.g., Braine 1971; R. Brown and Hanlon 1970; Marcus 1993) while others have provided evidence to the contrary, especially when corrective feedback takes the *indirect* form of recast/reformulation of the child's erroneous utterance in a more conventional fashion (e.g., Farrar 1992; Saxton 2000; Chouinard and Eve V. Clark 2003; Hiller and Fernandez 2016; Strapp 1999; Nelson, Carskaddon, and Bonvillian 1973).

Communicative Feedback, however, provides signals about *communicative* success (positive signals) or failure (negative signals). Therefore, unlike corrective feedback

and more like adult communicative coordination, CF focuses on understanding the child's communicative *intent* rather than correcting the form, meaning, or use of the child's language. Correction/reformulation may occur, but only after the interlocutor has successfully understood the child's intended meaning. CF only signals whether or not the listener (here, the more knowledgeable interlocutor) understood the communicative intent of the speaker (i.e., child).

Our main proposition is that many aspects of language can be acquired as a *side product* of the child trying to achieve shared understanding in conversation with a more knowledgeable interlocutor (e.g., a caregiver or an older sibling): Positive CF confirms their language use whereas negative CF urges them to revise the way they express their intent in future exchange.

1.2.1. Contributions

The general idea of communication/conversation as a matrix for language acquisition is not new. We can find it proposed in the work of many developmental scientists (e.g., Eve V Clark 2016; Eve V. Clark 2018; Halliday 1975; Bates 1979; Roberta Michnick Golinkoff 1986; Ochs and Schieffelin 1984; Yurovsky 2018). The novelty of the current work is twofold. First, we focus specifically on the role of CF and formalize it by making a systematic link with theories of communicative coordination that have been developed largely with adults (Pickering and Garrod 2021; H. H. Clark 1996). Second, we briefly review lines of experimental research in language acquisition that have been using measures that closely relate to the concept of CF in child-adult conversation and argue that this research can benefit from being unified under the theoretical framework that we propose.

The broad impact of this work is to bridge across two fields that have evolved largely separately (i.e., communicative coordination and language acquisition), providing a unifying framework for different lines of experimental research in the language acquisition literature. This theoretical effort is crucial not only to make sense of what appears to be disparate research goals, methods, and findings, but also to help locate gaps in the scientific literature and open up new promising areas for future research.

1.3. Communicative Feedback

For communication to succeed in a conversation, interlocutors coordinate to achieve and maintain *common ground*, a process also known as communicative grounding (Stalnaker 1978; Lewis 1969; H. H. Clark 1996). Intuitively speaking, this process characterizes conversation as a collaboration between (at least) two interlocutors trying to understand each other. To reach and maintain the state of mutual understanding, listeners send signals of understanding (e.g., acknowledgements), non-understanding (e.g. clarification requests), and mis-understanding (e.g., responding in a non-contingent fashion). The speakers use these signals either to move forward or to revise the expression of their intended meaning (H. H. Clark and Schaefer 1989;

Pickering and Garrod 2021).

Using this framework, we define Communicative Feedback as the signals sent by the listener to indicate communicative success or failure depending on whether or not the listener thinks they understood the intended meaning behind the speaker's linguistic utterance. Such signals have also been referred to as "closures" (H. H. Clark and Schaefer 1989; H. H. Clark 1996) or "commentaries" (Pickering and Garrod 2021).

In both cases (i.e., success and failure), CF can be either implicit or explicit: A listener can either "*say* that he[/she] understands [...], or *demonstrate* that he[/she] understands" (H. H. Clark and Schaefer 1989, p. 267).

Explicit positive signals of understanding are **acknowledgements**, also called "positive commentaries" in Pickering and Garrod (2021). These signals include short non-intrusive backchannel responses ("assertions of understanding" in H. H. Clark (1996); e.g., "uh-huh", "yeah", head nod, smile), as well as paraphrases or verbatim repetitions ("exemplifications of understanding" in H. H. Clark (1996)). [1] With these responses the listener asserts that they have understood the utterance of the speaker.

On the other hand, in the case of communicative failure, the listener can respond with a **clarification request** ("negative commentaries" in Pickering and Garrod (2021)) such as "Huh?", "Which one?", or a confused face. These are explicit signals of non-understanding.

Implicit signals of understanding are sent when the listener provides a response that is **contingent** on the speaker's utterance, as judged from the perspective of the speaker (e.g., responding "I'm at home." to the question "Where are you?"). If the listener responds in a **non-contingent** manner (e.g., responding "I'm fine." to the question "Where are you?"), they provide an implicit signal of communication failure to the speaker. The speaker can detect this misunderstanding if the response is non-contingent from their perspective. A similar concept has been described by H. H. Clark (1996, p. 228) under the name of *displays of understanding*, which can be exemplified, as we did above, by the fact that an answer displays (in part) whether a question was understood correctly or incorrectly.

The proposed classification of CF signals is summarized in Figure 1.2 and will help us sort/unify various experimental studies reviewed in the following sections.

1.4. CF for language learning

While language acquisition can be understood in broader terms, here we focus specifically on the process of learning to *understand* and *use* language in communication. Acquiring language requires the child both to learn how to infer a speaker's intended meaning from an utterance (when listening) and to learn how to produce a linguistic utterance that best conveys their intended meaning (when speaking).

CF-based mechanisms take as a starting point children's productions. Nevertheless, the learning that results form this mechanism is general to both comprehension and

1. See also Tannen (1989) and Norrick (1987) for the coordinative function of repetitions in conversation.

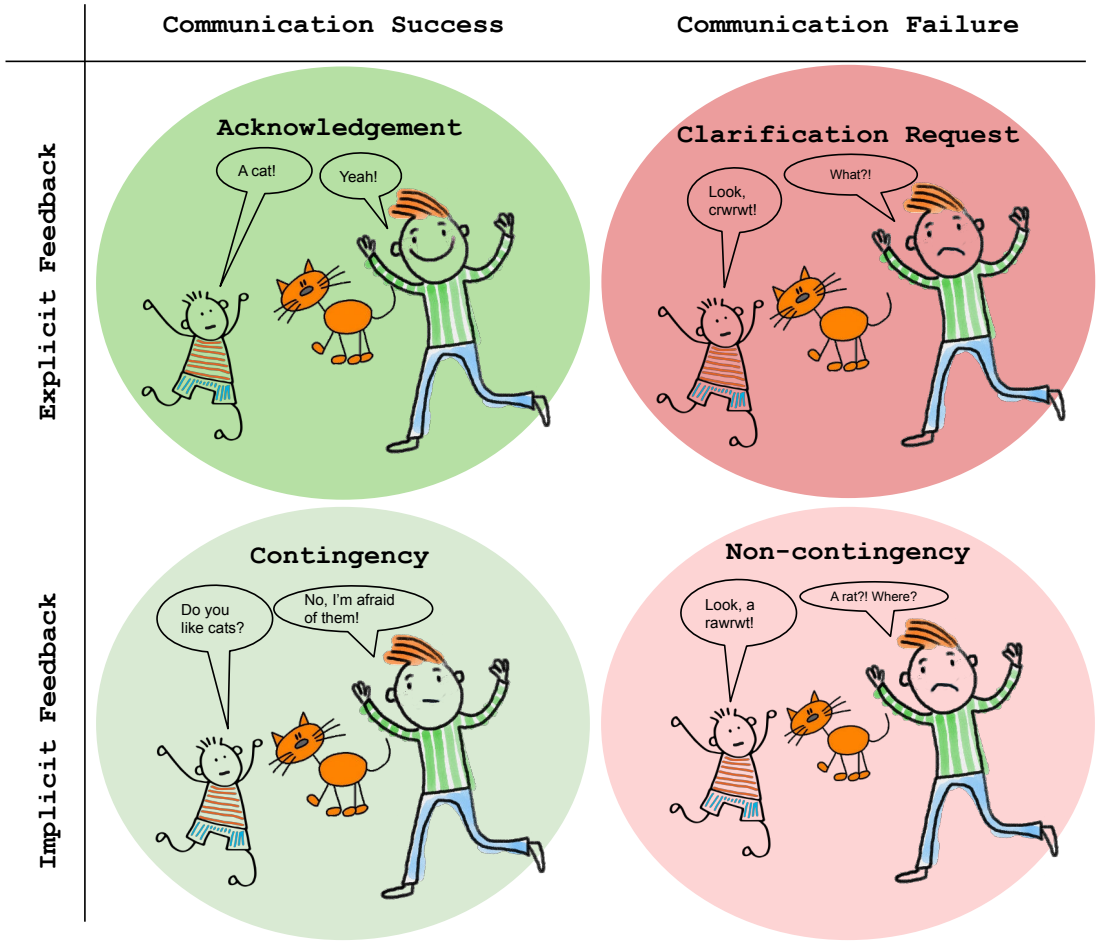

Figure 1.2. – Function and nature of Communicative Feedback signals. We illustrate each signal type with an example. Acknowledgement: The interlocutor acknowledges their understanding by smiling and uttering "Yeah". Clarification request: The interlocutor verbalizes their problem in understanding the child by responding with an open clarification request "What?". Contingency: The interlocutor responds with a relevant answer to the question, thereby providing an implicit signal to the child that they have understood the utterance. Non-contingency: The interlocutor misunderstands the child, responds non-contingently (from the perspective of the child), thereby providing implicit feedback signaling communication failure.

production. In fact, by producing linguistic utterances in the context of a conversation, children can be understood as putting their general linguistic knowledge to test, allowing them to receive feedback on it – whether implicitly or explicitly – from their more knowledgeable interlocutors (e.g., caregivers or older peers).

This view of language use as a driving force for language learning contrasts with traditional theories on language acquisition where major linguistic components, i.e.,

form, meaning, and use (L. Bloom and Lahey 1978) are experimentally compartmentalized and studied as if children learn them in a sequential and independent fashion. That is, children are sometimes understood as learning the *form* (e.g., the phonology of the word "water") based largely on the analysis of the linguistic input. Then, they would learn the *meaning* (i.e., that the form "water" maps on to the concept WATER) based mostly on multimodal association and categorization but also on pragmatic inference in social interaction. Finally, they learn how to *use* language in context to communicate their intent (e.g., the child uttering "Water!" to *request* WATER).

However, more recent theories on language acquisition do highlight synergies when learning form and meaning (e.g., Landau and L. R. Gleitman 1985; Babineau, Havron, Dautriche, et al. 2022; Feldman, Griffiths, Goldwater, et al. 2013; Abend, Kwiatkowski, N. J. Smith, et al. 2017; Räsänen and Rasilo 2015; Fourtassi, Regan, and Michael C. Frank 2020; Dupoux 2018; Christophe, Millotte, Bernal, et al. 2008), as well as when leveraging information about how language is used in context to learn various linguistic structures (Eve V Clark 2016; Eve V. Clark 2018; Bohn and Michael C. Frank 2019; Tomasello 2003). Most relevant to our proposal are the studies showing that children do not wait to have mastered the form and meaning before they start using language to communicate with people around them (Bates, Camaioni, and Volterra 1975; Halliday 1975; C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. 1996). In fact, the CF-based mechanisms assume that the feedback children receive on their early – correct or incorrect – language use allows them to refine their linguistic knowledge, a priori, at every level.

We illustrate the general idea of CF-based mechanisms in Figure 1.3, using word learning as an example. In brief, the CF-based mechanisms can be characterized as instances of social reinforcement. Signals of communicative success lead to positive reinforcement, thus comforting the child in their word choice. In contrast, signals of communicative failures lead to negative reinforcement, thus prompting the child to revise their knowledge and, in future exchange, use different words to try and better convey the intended meaning.

A crucial property of the CF-based mechanisms is that they do not require the interlocutor to explicitly teach or correct linguistic knowledge. Learning takes place as a *side product* of the child and interlocutor trying to understand each other. In fact, the mechanisms do not even require the interlocutor to interact with the child differently than they would do with any mature speaker of the language: Upon hearing the child's utterance, the interlocutor – as in a typical conversation between adults – produces positive CF or negative CF, depending on whether or not they have understood the message as intended by the child.

If the interlocutor thinks they understood the child's intended meaning, they can acknowledge the receipt and/or move forward with the interaction in a contingent fashion. This is, as described above, a positive signal to the child, confirming – and thereby strengthening – the child's linguistic use in such a context. If the interlocutor did not understand or misunderstood the message, they may ask for clarification or respond in a non-contingent fashion (from the child's point of view): Both are negative signals to the child, inviting knowledge revision.

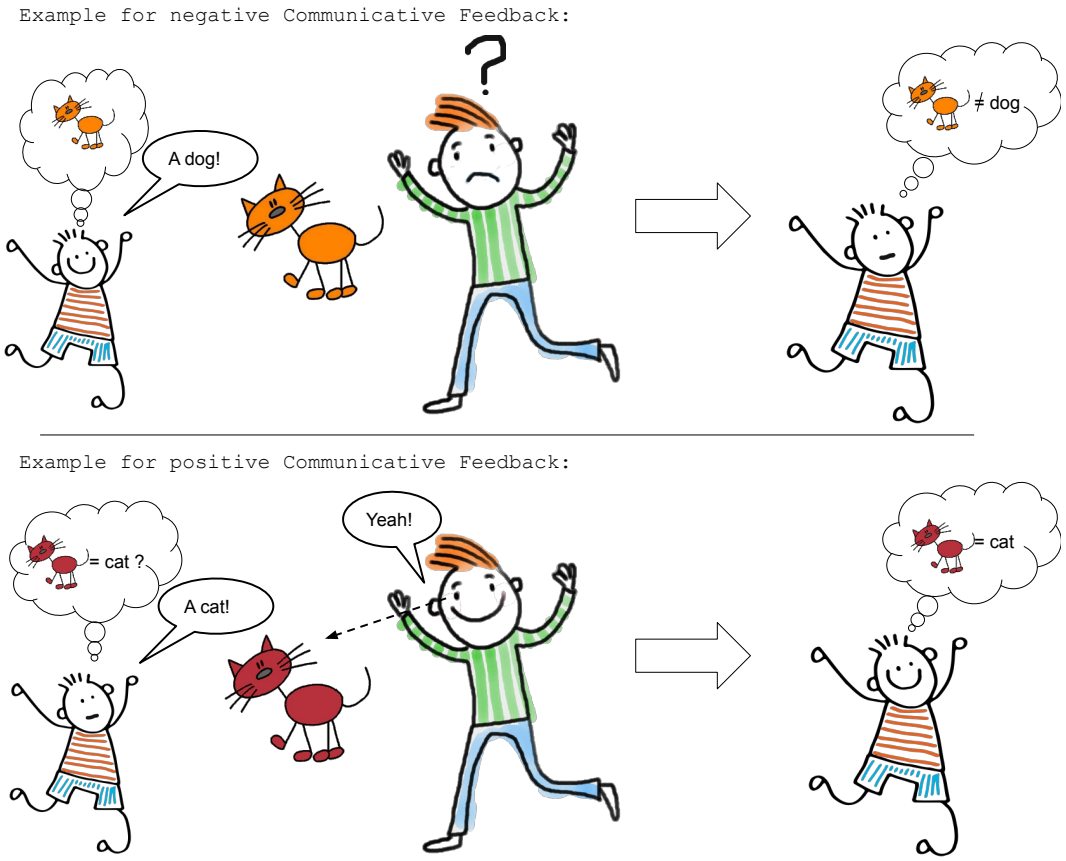

Figure 1.3. – We illustrate the CF-based mechanism with an example of word learning. The first illustration (top) shows a child that overgeneralizes the word "dog" to both cat and dog. Upon encountering a cat, they might say "A dog!" to draw the caregiver's attention to the cat. The caregiver would most probably react with a puzzled face, or ask for clarification, thereby providing rather negative CF to the child. Through this short interaction, the child can revise their knowledge about the meaning of "dog".
Later on (illustration at the bottom), the child might have learned about the word "cat" but might not be totally sure about its meaning. When encountering a new animal that looks like a cat, they might say "A cat!" The caregiver would most likely attend to the cat and respond contingently, thereby sending positive CF to the child, and strengthening the child's knowledge about the word cat.

It follows that the CF-based mechanism is an indirect way to learn language. The learner *uses* language in context and continuously updates their knowledge based on the feedback received. Such error-driven learning mechanisms have been proposed to play a major role in human learning more generally (Friston 2009; A. Clark 2015) and are increasingly being applied to language acquisition (Cox, Fusaroli, Keren-Portnoy, et al. 2020; Babineau, Havron, Dautriche, et al. 2022).

A question one could raise is the following: Why propose to study an indirect mechanism of language learning when previous research has focused on more direct mechanisms such as corrective feedback?

There are three main reasons. First, corrective feedback can only operate when the child produces relatively minor mistakes which do not impede the understanding of their intended meaning. Indeed, only if the interlocutor first understands the child's communicative intent can they then correct the mistake with a more conventional language use (e.g., via reformulation). In contrast, CF-based mechanisms are more general and can be useful even in early stages of development when children are barely intelligible (this will become clearer in the following section).

Second, while instances of corrective feedback have been observed in many naturalistic studies of child-caregiver interactions (e.g., Chouinard and Eve V. Clark 2003; Hiller and Fernandez 2016; Saxton 2000; Strapp 1999), it is unclear the extent to which this parenting style is constant across cultures. In fact, there is evidence that caregivers in some cultures talk only rarely directly to their young children or do not specifically adapt their language when talking to children (Shneidman and Goldin-Meadow 2012; Cristia, Dupoux, Gurven, et al. 2019; Casillas, P. Brown, and Levinson 2020; Ochs and Schieffelin 1984). In contrast, CF-based mechanisms are not specific to child-caregiver interactions. They rely on fundamental properties of human communication (H. H. Clark 1996; Pickering and Garrod 2021), making them much more likely to be universal across cultures. There is indeed accumulating evidence that feedback signals of the sort described above are present in a diversity of languages and cultures (although not always studied in the context interaction with children), such as for acknowledgements (Cutrone 2005; Maynard 1990; Liesenfeld and Dingemanse 2022), communicative repair (Dingemanse, Roberts, Baranova, et al. 2015; Ochs and Schieffelin 1984; Schegloff 2006), and time-contingent responses (Richman, Miller, and LeVine 1992; Marc H. Bornstein, Catherine S. Tamis-LeMonda, Tal, et al. 1992).

Finally, the feedback-based learning mechanism has the advantage that the learner can play an active role in shaping the learning process by engaging in curiosity-driven learning. That is, a child can choose to selectively initiate, shape and/or put more attention on topics with high amount of uncertainty in order to receive optimally informative responses depending on their current state in the learning process (Moulin-Frier, Nguyen, and Oudeyer 2014; Kidd, Piantadosi, and Aslin 2012; Gelderloos, Kamelabad, and Alishahi 2020; Twomey and Westermann 2018; Foushee, Srinivasan, and F. Xu 2022).

1.5. Empirical evidence for CF-based mechanisms

We looked in the development literature for experimental evidence supporting CF-based mechanisms in language learning. In the following, we provide an overview of major studies in each type of CF, in light of the classification made in Figure 1.2.

1.5.1. Acknowledgements

Within this category, we consider all responses that explicitly confirm understanding from the listener's side. These form an explicit positive CF signal to the speaker (in our case: the child).

Acknowledgements include backchannels, as well as certain kinds of repetitions. Backchannels are short non-intrusive vocalizations that signal attention, understanding, or agreement from the listener (Yngve 1970; Schegloff 1982; Bangerter and H. H. Clark 2003). [2] They can be verbal (e.g., "yeah", "right", "uh-huh") or non-verbal (e.g., smiling, nodding). Regarding repetitions, certain exact repetitions as well as paraphrases can function as acknowledgement, i.e. to communicate the receipt of information (Demetras, Post, and C. E. Snow 1986; C.-c. Huang 2011; H. H. Clark 1996).

While there is research on children's ability to produce and interpret backchannel signals in the context of child-caregiver interaction (Hess and Johnston 1988; Dittmann 1972; Bodur, Nikolaus, Prévot, et al. 2022), we found very few studies investigating the potential effect of received backchannels on children's language learning. We can mention the work by Peterson, Jesso, and McCabe (1999) who conducted an intervention study with preschool children, investigating the effect of caregivers' backchannel responses on children's narration (among other narrative-eliciting behaviors such as asking more open-ended and context-eliciting questions). They found that children in the intervention group showed more improvement in vocabulary and narrative skills both immediately after the intervention and in a follow-up testing one year after the intervention. In another work, Newport, H. Gleitman, and L. R. Gleitman (1977) included a "Note on Reinforcement" (p. 172), suggesting that backchannels "may constitute confirmatory evidence for a child trying to build some hypotheses about how to speak English effectively", because they indicate understanding of what the child said. They based this claim on their finding that the rate of caregivers' use of interjections (which include backchannels) was positively correlated with growth of children's productive vocabulary as well as their use of verb inflections and auxiliaries.

Regarding the role of repetitions, Demetras, Post, and C. E. Snow (1986) found that in naturalistic child-caregiver conversations, exact repetitions are much more frequently used in response to well-formed (semantically, syntactically and phonologically appropriate) than to ill-formed child utterances, thereby providing a useful positive feedback signal.

1.5.2. Clarification requests

Clarification requests (also referred to as other-initiated repairs) are used by listeners to signal difficulty or lack of understanding (M. R. J. Purver 2004; Schegloff, Jefferson, and Sacks 1977). They form an explicit negative CF, signaling to the speaker that

2. Schegloff (1982) argues that backchannels such as "uh-huh" ("continuers") are not strictly signaling understanding in all cases, but sometimes just an invitation for the speaker to continue (as the listener is passing on the opportunity to initiate a repair). We consider this still as a positive (but probably weaker) feedback signal to the speaker.

their intended meaning has not been communicated successfully. Importantly, this negative CF signal can be used by language learners not only to revise their message in the upcoming conversational turn, but also to take into account the communicative failure to improve their linguistic knowledge for future interactions. Clarification requests can also be verbal (e.g., "what?", "which one?") or non-verbal (e.g., frowning). We consider both open and restricted clarification requests as part of the CF-based mechanism, as they both signal a lack of understanding. [3]

In a naturalistic study of four mother-child dyads, Demetras, Post, and C. E. Snow (1986) found that mothers use clarification requests more often in response to ill-formed child utterances (semantically, syntactically or phonologically inappropriate) than to well-formed ones. They conclude that clarification requests therefore form a useful negative feedback signal. [4]

Regarding children's sensitivity to clarification requests, it has been shown that even preverbal infants attempt to repair conversations if interlocutors show signs of non-understanding (Roberta Michnick Golinkoff 1986). Studying children in their early stages of language use, Gallagher (1977) found that caregivers' clarification requests are understood by 2-to-3 year-olds and they follow up on these requests by revising or repeating their utterances. In the case of revision, which was more frequent, children either expanded on the original utterance or adapted the pronunciation. Saxton, Houston–Price, and Dawson (2005) studied the effect of clarification requests on grammatical errors with 2- and 4-years old children using an intervention paradigm. They found that children were more likely to correct grammatical errors (than to introduce an error) when prompted with clarification requests.

Several other studies explored the ways that children perceive and react to clarification requests at different stages of development (e.g., Corrin 2010; Wilcox and Webster 1980; Forrester and Cherington 2009; Bosco, Bucciarelli, and Bara 2006; Gallagher 1981; Brinton, Fujiki, Loeb, et al. 1986; Anselmi, Tomasello, and Acunzo 1986; Lustigman and Eve V. Clark 2019; Eve V. Clark and de Marneffe 2012; Carmiol, Matthews, and Rodríguez-Villagra 2018). We refer readers to Eve V. Clark (2020) for a more comprehensive overview on the role of clarification requests for language acquisition.

1.5.3. Contingency (or lack thereof)

We use the term contingency in a broad sense as any felicitous response (verbal or non-verbal) from the listener that is coherent/compatible with the speaker's utterance

3. The specificity of the feedback signal varies with the kind of clarification request. For open clarification requests ("What?"), the speaker only gets a binary feedback: The message has not been understood. Restricted clarification requests (e.g., Child: "I went to xxx." Adult: "You went where?") offer more specific feedback (arguably more valuable) on the part of the utterance that has not been understood. Restricted offers (Child: "I falled", Adult: "You fell?") offer the most specific feedback, as they additionally provide a possible repair. In that way, they are very close to corrective feedback, however we still include them within the framework of CF because they are part of general conversational management, and not specific to correcting children's mistakes.

4. However, see Marcus (1993) for an important critique of these results.

(e.g., responding on-topic to a statement or answering with "yes!" to a yes-no question). A contingent response is an *implicit* CF that shows the listener has understood the speaker's intended meaning.

Non-contingency, by opposition, is defined as any response that is incoherent with the speaker's utterance (e.g., an off-topic response or answering with "yes!" to a greeting), implicitly indicating to the speaker that the listener did not understand their communicative intent. It has been shown that from an early age, children are aware of breakdowns in social coordination more generally (Tronick, Als, Adamson, et al. 1978; Markova and Legerstee 2006; Bourvis, Singer, Saint Georges, et al. 2018), and try to re-establish communication, e.g., by using self-initiated repairs when the caregiver's response does not seem to match their expectations (Forrester 2008; Morgenstern, Leroy-Collombel, and Caët 2013).

Contingency is a notoriously challenging concept to operationalize (from the researcher's third point of view) because it requires inferring the child's communicative intent and judging whether the interlocutor's response is compatible with this intent. Both are non-trivial tasks. That being said, researchers have used various measures to approximate contingency in the context of children's early conversations.

1.5.3.1. Temporal contingency

Temporal contingency has been mainly used in studies with pre-verbal infants, especially regarding the development of their vocalizations into speech-like sounds. It describes responses that *follow* a child's communicative attempt within a short temporal delay, usually one to two seconds (Warlaumont, Richards, Gilkerson, et al. 2014; Goldstein, King, and West 2003; K. Bloom, Russell, and Wassenberg 1987).[5] The idea is that if a speaker receives a response (as opposed to silence or a delayed response), this provides positive reinforcing feedback.

Using controlled experimental paradigms, researchers have found that infants' proportion of speech-like (syllabic) sounds over vocalic sounds increased if caregivers responded time-contingently, as compared to when they responded at random timepoints (K. Bloom 1988; K. Bloom, Russell, and Wassenberg 1987; Goldstein, King, and West 2003).

Similar effects have been reproduced in more naturalistic settings. For example, Warlaumont, Richards, Gilkerson, et al. (2014) analyzed home recordings from child-caregiver conversations and found that (1) caregivers are more time-contingent on child speech-related vocalization (e.g., babbling) than on non-speech-related vo-

5. This contrasts with a closely related line of research on caregiver *responsiveness* which has studied responses that match the child's focus of interest rather than responses that provide feedback on the child's production (McGillion, Herbert, Pine, et al. 2013; Donnellan, Bannard, McGillion, et al. 2020; Gros-Louis, West, and King 2014; Z. Wu and Gros-Louis 2014; Akhtar, F. Dunham, and P. J. Dunham 1991; Carpenter, Nagell, Tomasello, et al. 1998; C. S. Tamis-LeMonda, M. H. Bornstein, and Baumwell 2001; Masek, McMillan, Paterson, et al. 2021).

We consider that these measures of contingency are therefore dealing with contingent *input* rather than contingent *feedback*. (Distinctions between feedback and "input at the right time" have been discussed in previous work (Poulson 1983; K. Bloom 1984; Goldstein and Schwade 2008)

calization (e.g., laugh or cry) and (2) children were more likely to continue with a speech-related utterances if they received a time-contingent response than if the caregiver was unresponsive.

Finally, Lopez, Walle, Pretzer, et al. (2020) found that sequences made of child canonical babbling, followed by caregiver time-contingent response, followed by repeated child canonical babbling were predictive of productive vocabulary later in the child's development.

Note that for studies that have focused on the role of social feedback in helping children transition from early vocalization (e.g., crying) to speech-related sounds (i.e., babbling), it is not straightforward to equate communicative feedback, as we defined it above, with the caregiver's temporal contingency because the child's production may lack communicative intent and the caregiver's reaction is unlikely to be driven by an effort to "understand." Indeed, babbling is still unintelligible speech; it does not make the communicative intent, if there is any, clearer than mere vocalic sounds. It is, therefore, likely that this early form of social reinforcement is driven by a desire for emotional connection/ attachment (Bowlby 1969; Ainsworth and Bowlby 1991) – without necessarily being about mutual understanding.

That said, we still consider this line of research to be related to our proposal. We believe this early form of "emotional connection"-based reinforcement represents a precursor, if not a basis, for later communication-based reinforcement when children start being able to talk about their intents in an (at least partly) intelligible fashion.

1.5.3.2. Content contingency

As soon as children's vocalizations start to to be intelligible, we can go beyond time-contingency and use measures of contingency that also take into account the the *content* of utterances.

Hoff-Ginsberg (1987) put forward the notion of topic-continuing replies to describe responses that refer to an entity or event that was referred to in the child's prior utterance. Caregiver's topic-continuing response behavior was found to elicit higher child responsiveness and to be predictive of children's vocabulary knowledge at later stages (Hoff-Ginsberg 1987; Hoff 2003).

The effect of negative feedback in the form of non-contingent responses to young infant's communicative attempts has been studied using controlled conversational paradigms (Shwe and E. Markman 1997; Grosse, Behne, Carpenter, et al. 2010). In these studies, the researchers showed that infants revise and repair their requests for objects in the case of misunderstanding, i.e. if their interlocutor did *not* understand their request correctly (e.g., if they responded "Oh, you want the paper?! Here you are!" to a child's request for a ball).

The research that aims at measuring lexical and semantic alignment in child-caregiver conversations can also be seen as capturing some aspects of contingency. In particular, many have investigated the extent to which caregivers re-use some of children words (or semantically related words) in their follow-up utterances (Fernandez and Grimm 2014; Yurovsky, Doyle, and Michael C Frank 2016; Misiek, Favre, and

Fourtassi 2020) and some have found this behavior to predict later development in linguistic skills (Fusaroli, Weed, Fein, et al. 2021; Denby and Yurovsky 2019).

1.5.3.3. Action contingency

Linguistic utterances do not only elicit verbal responses (e.g., a yes-no question eliciting a verbal answer), it can also elicit action (e.g., a request to hand over the ball). In the example of a request, the listener might just provide the speaker with the requested object as a response. As this is a form of successful communication, it provides positive CF. If a request is not met with the right action, this constitutes negative CF.

Whitehurst and Valdez-Menchaca (1988) studied the acquisition of foreign-language words for toys in 2 to 3 years old children. They found that children performed better in production and comprehension tests if they were (selectively) reinforced when making a correct production of the word by handing the corresponding toy to the child (and allowing them to play with it).

1.6. Conclusion

While the idea that children learn language (partly) in and through conversation is not new, here we made this link more systematic by drawing on insights from theories on conversational coordination. We focus specifically on the role of Communicative Feedback that a – more knowledgeable – listener (an adult caregiver or an older sibling) provides to the speaker (here the child), signaling communicative success or breakdown. The main argument is that such signals, though they may lack a teaching agenda, can be picked up on by children and used to refine their language skills, leading to more successful communication in future exchange.

Using this framework, we bridged across several lines on research in language acquisition that have been pursued largely independently but which, according to our framework, all investigate how children's learning can be improved by leveraging the explicit or implicit Communicative Feedback in a dialog. Further, our review of this literature – in the light of the big picture – has revealed several gaps that suggest themselves as priorities for future research in order to paint a more complete picture of children's language learning in an interactive context.

1.7. Directions for Future Work

In the light of our theoretical framework where we propose an explicit link between conversational coordination and language acquisition, the above literature review reveals several research gaps and points towards many directions for promising future work. Some of these have been addressed within the context of this thesis: Chapter 3 explores the role of temporal contingency and clarification requests on

the development of intelligible speech in early childhood. Further, it includes a reproduction from earlier findings regarding the role of temporal contingency of the development of speech (Warlaumont, Richards, Gilkerson, et al. 2014). Chapter 4 uses similar methodology to investigate the role of Communicative Feedback signals in response to children's grammatical errors. Both of these studies rely on a model for automatic annotation of speech acts in child-caregiver interactions, which is introduced in Chapter 2.

In the third part of this thesis (III), we propose to use computational models as a means for studying language acquisition at scale. In the first chapter (5) we describe a paradigm for the acquisition evaluation of word-level and sentence-level semantics in multimodal neural networks which was then used in chapter 6 to evaluate various models that integrate feedback-based learning algorithms with more traditional statistical learning algorithms. More specifically, the models provided proof-of-concept implementations using reinforcement learning as instantiation of Communicative Feedback and to investigate whether utterance-level feedback can in principle be leveraged by learners to improve their semantic knowledge. Further, it explored possible interactions between the different learning mechanisms.

Further open research directions that have not (or only partly) been addressed within the scope of this thesis are discussed in Chapter IV.

Part II.

Evidence for CF from Corpus Studies

Summary

2. Automatic Annotation of Speech Acts in Child-Caregiver Conversations

This chapter is based on the article "Large-Scale Study of Speech Acts' Development in Early Childhood" (Nikolaus, Maes, Auguste, et al. 2022), published in *Language Development Research*.

For the study of language acquisition in conversation, as proposed in the framework of Communicative Feedback, speech acts form an essential part. They allow us to classify the communicative intent of an interlocutor, which can in some cases directly be mapped to explicit positive or negative feedback signals: A backchannel response provides positive feedback to the interlocutor, while a clarification request provides negative feedback. Future work could leverage speech acts additionally to estimate the *contingency* of responses, thereby offering a way to quantify *implicit* Communicative Feedback signals (see also Section 7.3).

The following chapter introduces a tool for the automatic annotation of speech acts in child-caregiver conversations. Previous studies on speech acts in child-caregiver conversations have been investigating rather small samples of children, raising the question of how their findings generalize both to larger and more representative populations and to a richer set of interaction contexts. Here we propose a simple automatic model for speech act labeling in early childhood based on the INCA-A coding scheme (Ninio, C. E. Snow, Barbara A. Pan, et al. 1994). After validating the model against ground truth labels, we automatically annotated the entire English-language data from the CHILDES corpus. The major theoretical result was that earlier findings generalize quite well at a large scale. Further, we introduced two complementary measures for the age of acquisition of speech acts which allows us to rank different speech acts according to their order of emergence in production and comprehension.

The developed models are shared with the community so that researchers can use it with their data to investigate various question related to language use both in typical and atypical populations of children.

In the following chapters (Chapter 3 and 4), the proposed model is used to identify clarification requests in naturalistic child-caregiver conversations. Based on these annotations, we executed two corpus studies on the role of communicative feedback for learning to produce intelligible and grammatical speech when learning English.

2.1. Introduction

Research on language learning has largely focused on investigating how children acquire language form (e.g., phonology, lexicon, and syntax) and content (e.g., word and sentence meanings). Yet, an important aspect of language learning, which has received less attention, is the mastery of how to use language adequately in natural social interactions (L. Bloom and Lahey 1978). This mastery involves, in particular, using linguistic utterances to encoding and decode communicative intents (Grice 1975) or speech acts that characterize the illocutionary force of an utterance (e.g question, assertion, and request) (Searle 1976). Children's learning of speech acts is crucial for their ability to engage in coherent conversations. For example, it is important to recognize that an utterance is a "question" requiring an "answer", or that it is a "request" requiring "acceptance" or "refusal", instead.

Several taxonomies have been proposed that purport to capture children's emergent repertoire of speech act categories in the context of early child-caregiver social interactions (for reviews, see Cameron-Faulkner 2014; Casillas and Hilbrink 2020), the most comprehensive to date is the Inventory of Communicative Acts and its abridged version INCA-A (Ninio, C. E. Snow, Barbara A. Pan, et al. 1994).

C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) used INCA-A to study the emergence of speech act major classes in a longitudinal corpus of children aged 14 to 32 months old. [1] They documented several important findings that not only informed our understanding of language use development, but also shed light on how children's emerging linguistic skills interface with the development of their social-cognitive competences. By analyzing the development of the number of distinct speech acts as well as the distribution of speech acts used by children, they showed that when children utter their first words, they already express a range of simple communicative intents such as requests and questions. The repertoire of speech acts was observed in this study to increase rapidly within the first years of life, in tandem with development in social-cognitive and linguistic skills: Children become able to express more sophisticated speech acts such as "promise", "prohibit", and "persuade". Using the same coding scheme, Rollins (2017) and Rollins (1999) has shown that investigating speech act development can also help us study atypical cognitive development such as autism.

While this previous effort has been influential in the study of language use development, it has relied on hand annotation to code the data, which has limited the researchers' ability to explore how their findings generalize to larger population of children and across different interactive contexts. In fact, INCA-A is a rather complex scheme with a large number of categories (e.g., 67 different types of illocutionary acts) and its hand-annotation — including the effort of train annotators — is prohibitively expensive to deploy at a large scale.

1. While the terms "speech act" and "communicative intent" have sometimes been used by different researchers to mean slightly different things or to refer to different taxonomies, here — and for simplicity — we use them interchangeably to refer to the categories of communicative intents at the utterance level, as defined in the INCA-A coding scheme.

2.1.1. Current study

The current study aims at addressing this gap using recent advances in automatic speech act labeling. Using Snow et al.'s child-caregiver corpus and its INCA-A annotation, we tested various models on their ability to map utterances to corresponding speech acts and we selected the one that provided the best performance on a testing set made of unseen utterances from the same corpus.

Using this model, we examined how previous findings in speech act development generalized at scale. To this end we proceeded in two steps: First, we validated the chosen model by testing its ability to replicate key findings from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996). More specifically, we reproduce developmental patterns regarding the number of distinct speech acts as well as the distribution of speech acts used by children from 14 to 32 months of age. Second, and after successful validation, we used the model to automatically label the entire North American English-language section of CHILDES (MacWhinney 2014) and compared the results of this large-scale analysis to the original findings.

Additionally, we proposed methods for quantifying the age of acquisition of a speech act both in terms of production and comprehension. These measures have allowed us to rank different speech acts according to their order of emergence. We first examined this order of emergence with data in C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996), and second, thanks to our automatic labeling tool, we tested how this developmental trajectory generalized across all English language corpora in CHILDES.

This chapter is organized as follows. First, we introduce the dataset and provide an overview of models for automatic annotation of speech acts that we evaluated in our study. Further, we define the measures for speech act emergence in production and comprehension. In the results sections we compare the performance of the selected models and present replications the findings of C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) using automatically generated labels. Additionally, the results contain predicted ages of acquisition for each speech act using both manually-annotated and automatically-annotated data. Finally, we discuss the results in the context of language development in general and point out limitations of the current approach which offer possibilities for future research.

2.2. Datasets and Methods

2.2.1. Datasets

New England Corpus For model training and validation, we use ground-truth labels from the dataset collected by C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) which is the largest child-caregiver interaction dataset annotated for speech acts. This dataset was collected for a longitudinal study of 52 children aged 14, 20 and 32 months old. Child-caregiver dyads were invited for three sessions that consisted of semi-structured free play. All conversations were recorded, transcribed,

and annotated with INCA-A coding scheme. There were 55,941 labeled utterances in total.

English-Language CHILDES In order to test how findings from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) generalize to a larger dataset of children and across different contexts, we use the entire North American English-language subset of CHILDES made of children in the same age range (i.e., between 14 and 32 month old), resulting in 2078 different transcripts totaling 354 children. [2]

2.2.2. INCA-A Coding Scheme

INCA-A is the most comprehensive coding scheme to date that was designed to capture children's emerging speech acts it the context of spontaneous social interaction with a caregiver (Ninio, C. E. Snow, Barbara A. Pan, et al. 1994). The coding scheme has two coding tiers: 1) the interchange level that annotates the topic of the conversation (e.g., "discussing a recent event"), and may span multiple utterances, and 2) the illocutionary force level (e.g., "Ask a yes/no question") which is determined at the utterance level. Here, we focus on the illocutionary force. INCA-A has 67 different speech act types, which are grouped into several high-level categories such as directives, declarations, commitments, markings, statements, questions, evaluations, and other vocalizations. [3]

2.2.3. Automatic Classification of Speech Acts

Speech act classification (also referred to as dialogue act tagging in the field of Natural Language Processing) describes the task of annotating utterances in dialogue with their respective speech act category. Given a transcript of a conversation and a speech act coding scheme, each utterance in the transcript is assigned one of the speech acts in the coding scheme (Stolcke, Ries, Coccaro, et al. 2000).

Early work used Hidden Markov Models to map utterances to speech acts using a set of lexical, collocational, and prosodic cues (Stolcke, Ries, Coccaro, et al. 2000). Subsequent work has used Recurrent Neural Networks (RNNs) such as Long short-term memory networks (LSTMs) for encoding transcribed utterances in order to leverage the sequential structure of the data (Khanpour, Guntakandla, and Nielsen 2016). More recent approaches combine hierarchical deep neural network encoders with Conditional Random Field (CRF) decoders (Kumar, Agarwal, Dasgupta, et al. 2018). While the encoder is aware of relationships between the different utterances of a transcript and thus models dependencies in the *feature space*, the CRF can model transition probabilities in the *label space*. In this way, it can for example learn common

2. For fair comparison, we excluded very short transcripts where the number of children's utterances was less than the minimum number of children's utterances in transcripts of the New England corpus at the same age.

3. Refer to the appendix for the full list of speech acts.

adjacency pairs (Schegloff and Sacks 1973) in conversation, e.g. that questions are usually followed by answers.

Following this brief review, we considered and compared the following models.

2.2.3.1. Baselines

As this work is the first to propose automatic speech act annotation using the INCA-A coding scheme on child-caregiver conversations, we run several baselines in order to obtain reference performances on this specific task.

Majority Classifier As a first simple baseline, we consider the majority classifier, which always predicts the most frequent speech act.

Random Forests We use the reference implementation of a random forests algorithm from scikit-learn (Pedregosa, Varoquaux, Gramfort, et al. 2011). As features, we provide the model with the speaker (caregiver or child), bag-of-words, part-of-speech tags (that are present in the corpus [4]), and the number of words in the utterance.

Support Vector Machine Using the same features as for the random forests model, we train and evaluate a linear support vector machine from scikit-learn.

2.2.3.2. Conditional Random Field

Next, we consider a CRF as annotation model. We hypothesized this model would outperform the baselines thanks to its ability to track transition probabilities in the label space. We use *pycrfsuite*[5] (Okazaki 2007) to implement the CRF. We extend the set of features used by the baseline models and add bigrams and repetitions (words that are repeated from the previous utterance, as well as the number of repeated words normalized by the utterances length) to provide the model with some context of the previous utterances. [6] The model uses the whole conversation in a transcript to find the most probable sequences of labels using the Viterbi algorithm.

2.2.3.3. Hierarchical LSTM + CRF

We further consider a model that is inspired by state-of-the-art speech act annotation models in other domains. More specifically, we implement a hierarchical LSTM encoder combined with a CRF decoder similar to the implementation of Kumar, Agarwal, Dasgupta, et al. (2018). The encoder processes the utterances within a transcript

4. The POS tags in CHILDES were automatically generated using the Morphological Analysis algorithm (MOR; MacWhinney 2000) which yields a high accuracy rate on CHILDES adult data (above 99%).

5. https://github.com/scrapinghub/python-crfsuite

6. In preliminary experiments we tested adding all the exact words of previous utterances as features to the model but observed, if anything, a small degradation in performance.

on two levels. We add a special token representing the speaker identity to the beginning of each utterance. Afterwards, for each utterance, one-hot encodings of the words are passed through word embeddings, and are then encoded using the word-level LSTM. The last hidden representation of this LSTM forms the latent utterance representation, which is then passed into the utterance-level LSTM. This higher-level LSTM processes the utterances sequentially and generates conversation-context-aware representations. The output of each timestep of the utterances LSTM is then passed as features to a CRF, which predicts the corresponding speech act. The model has access to contextualized utterance representations as well as the history of speech acts for the classification task. A high-level overview of the architecture of this model can be found in the appendix (Figure 1).

2.2.3.4. BERT

Given recent developments in NLP regarding the success of pre-trained contextualized embeddings (Devlin, Chang, K. Lee, et al. 2019), we additionally test the performance of a model where utterances are encoded using BERT. The success of these models relies on self-attention mechanisms that allow the model to create contextualized representations with long-range dependencies as well as setups in which the encoder is pre-trained on large-scale data before being fine-tuned on the actual task. Here we replace the word-level LSTM of the Hierarchical LSTM + CRF model with a pre-trained publicly available implementation of DistilBERT (Wolf, Debut, Sanh, et al. 2020). The weights of BERT are fine-tuned on the task. Details on the hyperparameters of the neural network models can be found in the Appendix.

2.2.4. Measures of Speech Act Emergence

Here we introduce measures of speech acts' age of emergence, both at the level of children's production and comprehension.

2.2.4.1. Production

By analogy to work in word learning (J. C. Goodman, Dale, and P. Li 2008; Braginsky, Yurovsky, V. A. Marchman, et al. 2016), we define the age of acquisition of a speech act in production as the month by which at least 50% of the observed children produce it.[7] More precisely, for each speech act S, we proceed as follows:

1. For each age in the dataset (i.e., 14, 20 and 32 months), calculate the proportion of children who are producing S at least twice.
2. Perform a logistic regression over these proportions.
3. Measure the age of first production as the age where the logistic regression curve surpasses the value 0.5.

7. In line with C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996), we consider that a child acquired a speech act if it is produced at least twice at a certain age.

2.2.4.2. Comprehension

Studying speech act emergence only from a production point of view may underestimate children's pragmatic competence. Thus, we additionally introduce a measure for children's comprehension, which we define as the ability of children to respond to a target speech act in a contingent fashion (e.g., responding to a "yes/no question" with "yes" or "no"). More precisely, for each speech act S, we proceed as follows:

1. Find all utterances produced by the caregivers labeled as S.
2. Find all cases where these utterances are followed by an utterance of the child.
3. For each occurring follow-up utterance, annotate whether its speech act is contingent as a response to S. [8] We manually annotated the contingency of all combinations of speech act categories that appear in the data. Using this annotation, we could label each child utterance that follows a caregiver utterance as either possibly contingent or non-contingent based on the corresponding speech act category. The contingency annotation can be found in the GitHub repository: https://github.com/mitjanikolaus/childes-speech-acts.
4. For each age (14, 20 and 32 months), calculate the proportion of contingent follow-up utterances.
5. Perform a logistic regression over the proportion. [9]
6. Measure the age of comprehension as the age where the logistic regression curve surpasses the value 0.5.

2.3. Results and Analyses

First, we compare performance across all models presented above on the New England corpus. Second, we choose the best performing model and test the extent to which its predicted labels replicate major findings obtained using gold labels from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996). Finally, we use the model to automatically label the North American section from CHILDES and explore how original findings from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) on the emergence of speech acts generalize to this larger dataset.

2.3.1. Comparing Models of Speech Act Labeling

We evaluate our models on the speech act annotations of utterances in the New England corpus (C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. 1996). We employ 5-fold cross validation so that we evaluate (and later utilize in all analyses)

8. Annotating contingency was done using a binary scale, indicating whether the speech act was *possibly* contingent (1) or clearly non contingent (0). A speech act was considered contingent (1) if it can form a coherent response with respect to the previous speech act, and non contingent (0) otherwise.

9. We only regard data points where the proportion was calculated over at least 2 examples, i.e. where there were at least two utterances with follow-ups.

Table 2.1. – Accuracy for all models.

Model	**Accuracy**
Majority Classifier	13.44% (±2.81%)
Random Forests	62.81% (±6.29%)
Support Vector Machine	62.42% (±6.97%)
Conditional Random Field	**72.33%** (±4.23%)
Hierarchical LSTM + CRF	69.77% (±3.70%)
+ BERT	68.50% (±4.29%)
Inter-Annotator Agreement	81% to 89%

only the predicted labels on the parts of the corpus that were not seen by the model in the training phase. To this end, and to obtain labels for the whole New England corpus, we train models on 5 different training sets, always holding out 20% of the data. Then we use each of the trained models to label their respective test sets which together form a set of predicted speech act labels for the whole New England corpus.

We report the mean and standard deviation (based on the five cross-validation runs) of each model's accuracy in Table 2.1.

The majority classifier had a high score given the relatively large label space. This could be explained by the fact the label distribution is heavily skewed (Figure 2.1). A small set of speech acts are used very frequently while several others are rarely used. As for other baseline models, i.e., random forests and support vector machine, the scores are relatively high despite the fact that they do not have access to the conversation history or dependencies in the label space. Our more sophisticated models (Hierarchical LSTM with and without BERT) did not improve performance much, which could be explained by the lack of large-scale training data. Further, in the case of the BERT-based model, we hypothesize that we do not see any performance gains because this model is pre-trained on large text corpora (based on e.g. Wikipedia) that do not have much in common with the dynamics of child-caregiver conversations.

Finally, we find that the CRF model shows the highest accuracy scores, outperforming the baselines as well as the more complex neural network models. Its large performance gains over the baseline are most likely explained by its ability to track transition probabilities in the label space. This property is crucial for the task of speech act annotation; given a speech act sequence, certain speech acts are very likely to follow and others are not. The CRF is the best-performing model, and thus, it is the one we employ for the rest of analyses in this chapter.

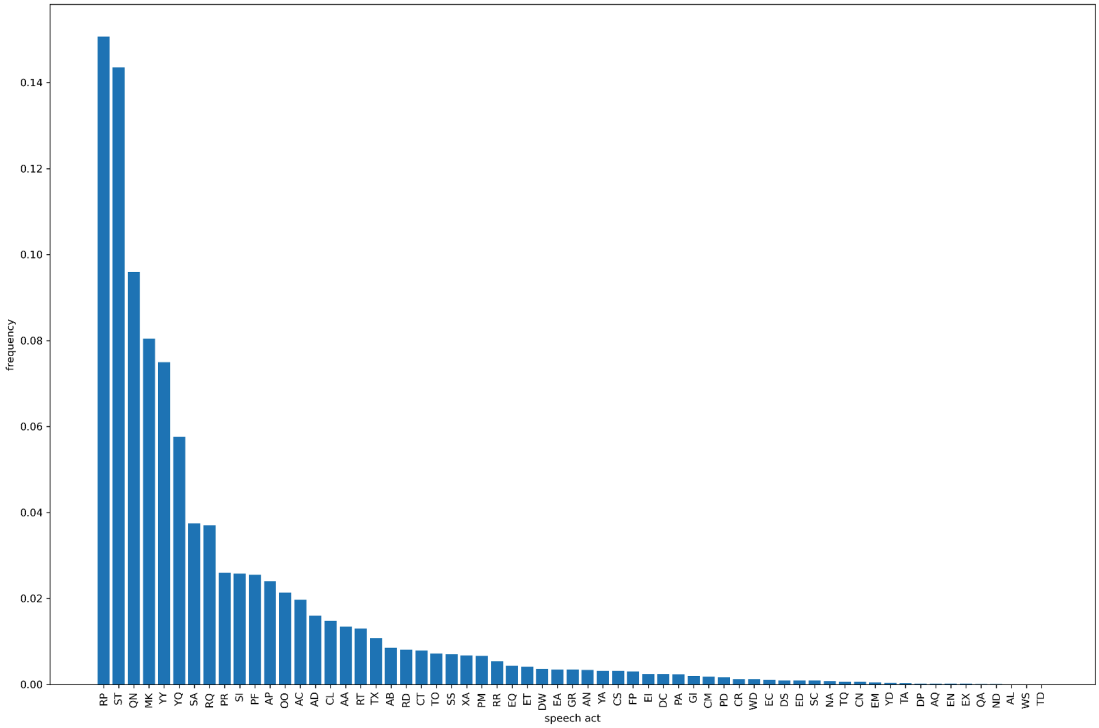

Figure 2.1. – Distribution of frequencies of all speech acts in the New England corpus. Labels from the INCA-A tagset are listed in the Appendix.

2.3.1.1. Amount of Training Data

We further investigate the effects of the amount of training data on the performance of the CRF model. Figure 2.2 presents the test accuracy as a function of training set size for this model. The performance indicated in Table 2.1 was obtained when the model was trained on 80% of the dataset (around 44,000 utterances). However, from the learning curve in Figure 2.2 we can see that the model actually achieves decent scores (around 65% accuracy) when trained on only 5,000 annotated utterances, and almost converged when trained on about 20,000 annotated utterances.

2.3.1.2. Error Analysis

To gain a better understanding of our best performing model (the CRF), we perform an error analysis. For each speech act category, we calculate precision, recall and f1-score. Results can be found in the Appendix. The variance of the f1-scores for different categories is remarkably high, with values ranging from 0 to 95%. Performance is best for speech acts `QN` ("Ask a product-question") and `EA` ("Elicit onomatopoeic or animal sounds.") and worst for speech acts such as `CR` ("Criticize or point out error in nonverbal act") and `AL` ("Agree to do something for the last time.").

One important factor affecting the per-label performance is the availability of training examples and the distribution of speech acts in the dataset is heavily skewed with

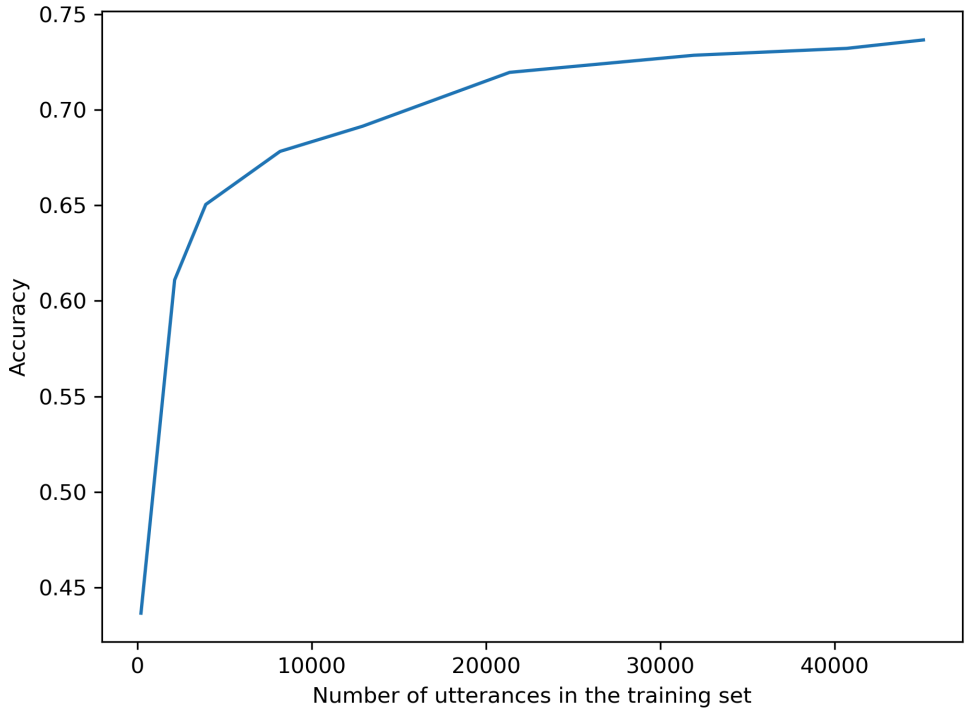

Figure 2.2. – CRF: Accuracy as a function of training set size.

a long tail (see Figure 2.1). For labels with only very few training examples the model struggles to pick up important features. Indeed we find a high correlation between the frequency of labels and their respective f1-score (Spearman correlation coefficient: 0.59, $p < 1 \cdot 10^{-5}$). The example in Table 2.2 illustrates this finding. In the conversation, all speech acts have been predicted correctly by our model except for the last utterance ("You're a nut"), which is labeled as ST ("Make a declarative statement") while the ground-truth label is DS ("Disapprove scold protest disruptive behavior"). Indeed, the speech act DS occurs very few times in the training data (only 40 examples, i.e., less than 0.1% of the training data).

Another factor that affects the model's performance is what appears to be ambiguities in the definition of some categories in the INCA-A coding scheme. In particular, many pairs of speech acts are either very similar or hierarchically related (see Cameron-Faulkner and Hickey (2011) for a similar observation). More concretely, there are pairs of speech act categories that describe overlapping communicative intents (e.g., "Criticize or point out error in nonverbal act" (CR) can overlap with "Disapprove scold protest disruptive behavior" (DS) and pairs of speech acts where the meaning of one act appears to be covered by the other broader act (e.g., the speech act "Praise for motor acts i.e for nonverbal behavior." (PM) is part of "Approve of appropriate behavior." (AB)). Such overlaps in the definition of some categories do not help the model make clear distinctions between the affected categories and, thus, tend to conflate them.

We provide an example for this phenomenon in Table 2.3. In this conversation,

Table 2.2. – Excerpt of a conversation from the New England Corpus (Child: Liam, Age: 14 months, Transcript: 99) with manually-annotated speech acts ("Manual") and predicted speech acts ("CRF"). Labels from the INCA-A tagset are listed in the Appendix.

	Speech act	
	Manual	CRF
Mother: We're having a little problem here in the corner.	ST	ST
(Mother stands up)		
(Child unplugs cord from wall again)		
Mother: Liam !	CL	CL
(Mother takes hold of Child's hand)		
Mother: No!	PF	PF
(Mother takes hold of cord and tries to pull it out of Child's hand, Child holds onto cord)		
Mother: Let go.	RP	RP
(Child lets go of cord, Mother plugs cord back into wall, Child watches what Mother does with cord)		
Mother: No.	PF	PF
(Mother picks up Child)		
Mother: You're a nut.	DS	ST

the mother's utterance "Good girl" is labeled by the CRF as "Approve of appropriate behavior." (AB), which is not incorrect, but differs from the human annotation, which categorizes it as "Praise for motor acts i.e for nonverbal behavior." (PM). We hypothesize that collapsing overlapping categories would improve the model performance. Indeed, we experimented with an alternative coding scheme where we collapsed certain categories and the model achieves a higher average performance of 75.35% (±4.17%) accuracy. However, for the remainder of this work, we continue using the original coding scheme to ensure comparability to the work of C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996).

2.3.2. Replicating Findings from Snow et al. (1996)

Here we validate the CRF model by testing its ability to lead to conclusions similar to the ones obtained in C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996). To this end, and as we mentioned earlier, we proceed in two steps: First, we replicate major findings in C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) using their hand-annotated labels. Second, we compared them to the corresponding findings obtained using the labels that were predicted using our CRF model. In addition to replicating main analyses from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) (i.e., development of the size and distribution of speech acts), we also tested the models with a new, more specific task that consists of

Table 2.3. – Excerpt of a conversation from the New England Corpus (Child: Joanna, Age: 20 months, Transcript: 32) with manually-annotated speech acts ("Manual") and predicted speech acts ("CRF"). Labels from the INCA-A tagset are listed in the Appendix.

	Speech act	
	Manual	CRF
Mother: Take it [= book] out of the box. (*The child struggles with both hands on the open book. Afterwards, the child pulls the book up and out of the box*)	RP	RP
Mother: Good girl.	PM	AB

predicting the precise normative age of acquisition of speech acts in both production and comprehension.

2.3.2.1. Development of the Number of Distinct Speech Acts

Figure 2.3 shows the proportion of children producing a given number of different speech act types for the three age groups studied in C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) (This is a direct replication of Figure 2 in the original paper). Next to each bar obtained from the hand-annotation (in blue) we plot the corresponding bar from the automatic labeling by CRF on the same dataset (in orange).

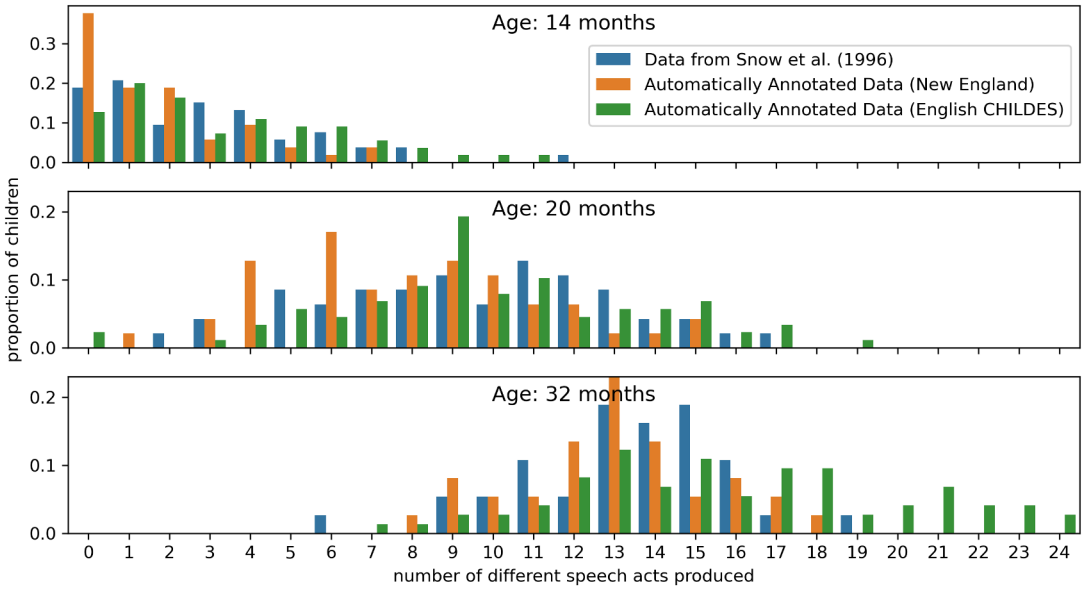

Figure 2.3. – Proportion of children producing a given number of distinct speech act types at 14, 20, and 32 months old. Note that the y-axis for the bottom two figures has been shortened for better visibility.

We can see that the patterns observed in C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) are well captured by automatic labeling data: At 14 months, most children produce only a handful of speech act types, such as statements (ST), repetitions (RT) and markings (MK). This number increases on average for children aged 20 months where now a substantial proportion of children become able to produce around 10 different speech act types (now starting to use for example requests (RP), stating intent (ST) and product questions (QN)). Finally, at 32 months, children typically produce between 10 and 20 different speech act types (starting to use for example polar questions (YQ)). When compared to hand annotated data in the New England corpus, the model was able to capture not only the rough number of speech act types produced at each age range, it was also able to capture quite well the variability between children at each age.

We can quantify the similarity between the hand- and automatic-annotation-based distributions by computing their Jensen-Shannon distances. This measure quantifies the dissimilarity between two probability distributions with values ranging from 0 (maximally similar) to 1 (minimally similar). The similarities of distributions from manually and automatically annotated data were as follows: 0.262 (at 14 months), 0.367 (at 20 months), and 0.186 (at 32 months).

2.3.2.2. Development of the Distribution of Speech Acts

Figure 2.4 shows the replication of the analysis on the development of the distribution of speech acts (cf. Table 9 in C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996)). This analysis compares the proportions of utterances that fall within each speech act category for the three age groups. Similar to the previous graph, next to each bar obtained from the hand-annotation (in blue) we plot the corresponding bar from the automatic labeling by CRF (in orange). We can see that the frequency distributions look remarkably similar in each age group (see Appendix for the legend of what each speech act label refers to). Jensen-Shannon distances of automatically annotated data (New England) compared to data from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) were: 0.089 (14 months), 0.103 (20 months), 0.080 (32 months).

2.3.3. Generalizing Findings to Data in CHILDES

In the previous subsection, we validated the model by comparing findings from predicted and hand-annotated labels of the same data. Here, we use the trained model to automatically annotate data from English corpora in CHILDES. The goal is to investigate the extent to which findings obtained in C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) generalize to a larger number of children and to the variety of communicative contexts represented in these new corpora.

More precisely, we trained the CRF on the whole New England corpus (no held-out test set) and used it to annotate speech acts on transcripts of children aged between 14 to 32 months old in the North American English corpora of CHILDES (excluding

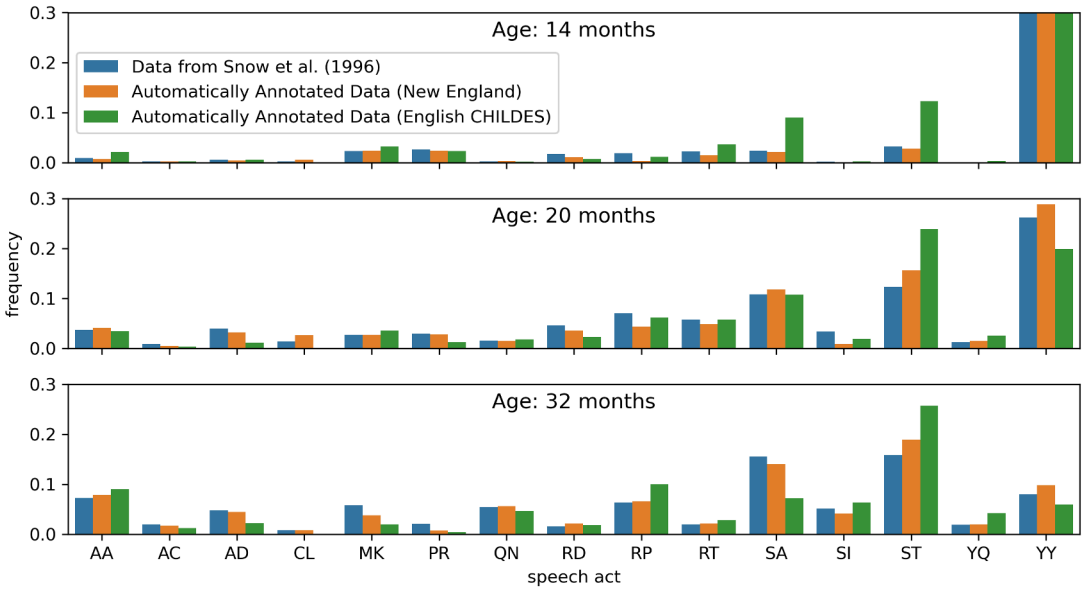

Figure 2.4. – Frequency distribution of speech acts for different ages. Note that the y-axes have been trimmed for better visibility (The frequencies for YY at 14 months are around 0.6).

transcripts from the New England corpus). Next, we perform the same analyses as in the previous section using the large-scale annotated data.

2.3.3.1. Development of the Number of Distinct Speech Acts

The green bars in Figure 2.3 show the number of different speech act types produced by children from CHILDES. Developmental patterns are very similar to the original graphs (in orange), with the exception of the oldest age group (i.e., 32 months) where we found that more children produced a relatively larger number of different speech acts (more than 20). Jensen-Shannon distances of automatically annotated data (English CHILDES) compared to data from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) were: 0.209 (at 14 months), 0.222 (at 20 months), and 0.418 (at 32 months).

2.3.3.2. Development of the Distribution of Speech Acts

We present the frequency distribution of speech acts for children from CHILDES in the green bars of Figure 2.4. Again, patterns obtained by C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) generalize very well. Jensen-Shannon distances of automatically annotated data (English CHILDES) compared to data from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996): 0.204 (14 months), 0.173 (20 months), 0.197 (32 months).

2.3.4. Age of Acquisition of Speech Acts

In this section, we present results for the age of acquisition of speech acts in terms of production and comprehension using the measures defined in the Section "Measures of Speech Act Emergence".

2.3.4.1. Production

We calculated the age of acquisition for a subset of 25 speech acts [10] using both the manually-annotated labels from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) and the automatically generated labels from the CRF on the same dataset. Examples for regression plots and predicted ages of acquisition for all speech acts can be found in the appendix. Then, we calculated the Spearman rank-order correlation [11] to examine whether the *order* of emergence of speech acts is correctly captured by the automatically annotated data.

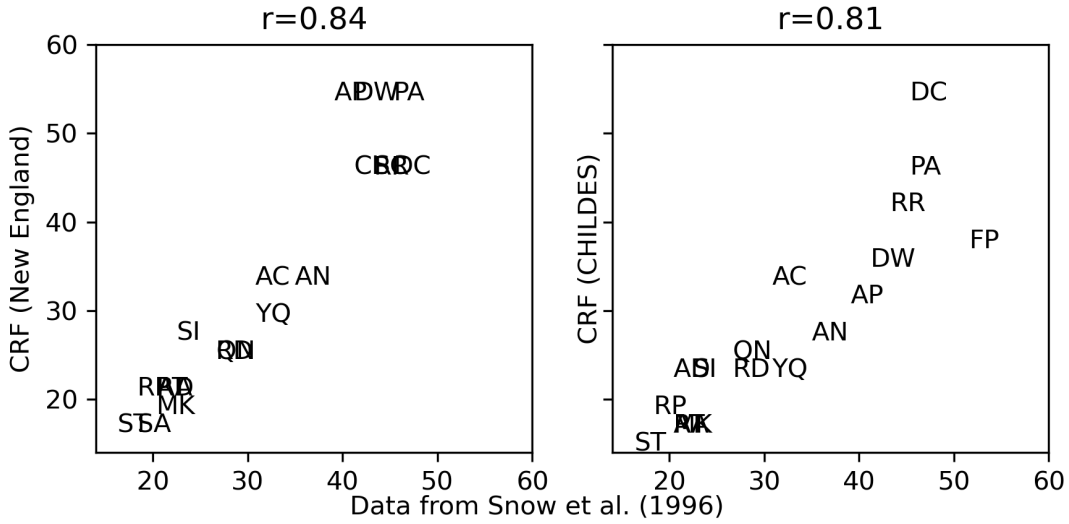

Figure 2.5. – Correlation of age of acquisition in terms of production as calculated using data from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) and automatically annotated data for the New England corpus and CHILDES. Note that some speech acts are not displayed because the axes limits were set to 60 months for better visibility of early development. However, the correlation was calculated for all values.

10. These were the ones for which we could fit a logistic regression using at least two data points. While the number of acts we keep may seem small compared to the original size (65 possible speech acts excluding categories for unintelligible speech acts, YY and OO), it is due to the fact that the frequency distribution is highly skewed: Most categories occurred rarely in the corpus (Figure 2.1) and therefore did not provide enough data to be used in the calculation of age of acquisition.

11. The rank-order correlation was computed over the subset of 25 speech acts for which an age of acquisition could be calculated, details in the Appendix.

The resulting high correlation (see Figure 2.5 (left); $r \approx 0.84$, $p < 1 \cdot 10^{-6}$) indicates that the automatically generated labels can provide reasonable estimates for the developmental trajectory of speech acts.

We also calculated ages of acquisition using the predicted labels on CHILDES data. Figure 2.5 (right) shows the correlation with the ages calculated using New England data. Spearman rank-order correlation was $r \approx 0.81$ ($p < 1 \cdot 10^{-6}$).

2.3.4.2. Comprehension

To illustrate the emergence of speech acts in terms of comprehension, we first show observed adjacency pairs for adult-child turns for different ages in Figure 2.6. The youngest children respond with unintelligible utterances or utterances without clear function (YY, OO) in most of the cases displayed. Children at 20 months show some consistent patterns in their response behavior: Polar and product questions (YQ, QN) are answered with adequate responses (AA, SA). Polite requests (RQ) are either accepted (AD) or refused (RD). Requests or suggestions (RP) are also usually accepted or refused, although in some cases children answer with a statement (ST), which is not contingent. Additionally, there is still a large amount of utterances without clear function (YY). Only by the age of 32 months, most of the parents' utterances are addressed with contingent responses (at least as captured at the broad level of speech act categories).

Examples for predicted ages of acquisition for all speech acts can be found in the appendix. We observe that while there are similar trajectories in production and comprehension for some speech acts (e.g. RR), we also observed some striking differences in other cases. For example, "demands for permission" (FP) is produced very late (around 52 months), but they are already understood a lot earlier (around 14 months).

As done for the production measure, we calculated the age of acquisition using both the ground-truth labels from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) and the automatically generated labels from the CRF on the same dataset, as well as using generated labels on the English CHILDES data. As in production, the Spearman rank-order correlation coefficient [12] (see Figure 2.7, left; $r \approx 0.46$, $p < 0.01$) indicates a statistically significant positive correlation (however lower than for the production measure). For the correlation with predicted labels on CHILDES data, the Spearman rank-order correlation was $r \approx 0.63$ ($p < 1 \cdot 10^{-5}$; see Figure 2.7, right). [13]

12. The rank-order correlation was computed over the subset of 47 speech acts for which an age of acquisition in terms of comprehension could be calculated, i.e. cases in which we could fit a logistic regression using at least two data points, details in the Appendix.

13. As we said above, we chose to fit the age of acquisition using logistic regressions following the method used for the AoA of words Michael C. Frank, Braginsky, Yurovsky, et al. (2021). The main limitation here was the sparsity of available annotated data: The study by Snow et al. (1996) only considers 3 different age groups: Children at 14, 20, and 32 months. While the fitted curves were good for production, this was less obvious for comprehension data based on contingency (see the graphs in the appendix). Note, however, that for our analysis, i.e., correlating AoA from predicted vs. hand-annotated speech acts (Figures 6 and 7), we only needed the ranking of AoA, not necessarily absolute values of ages. So, one simple way to test the robustness of these correlations is the following:

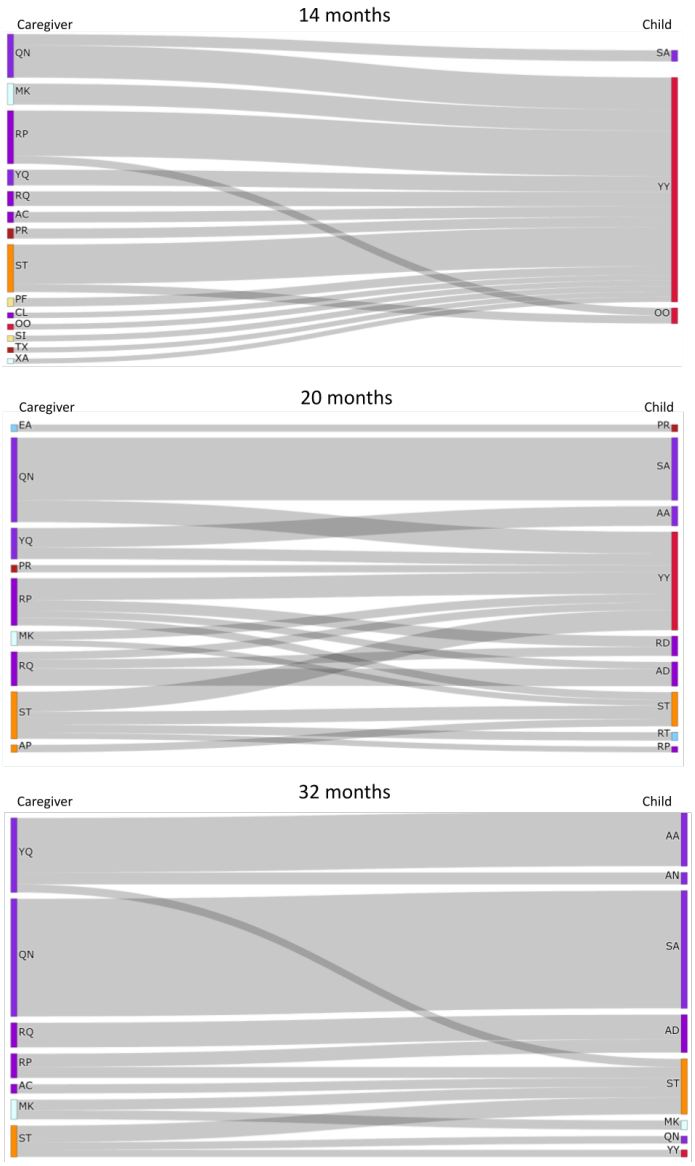

Figure 2.6. – Adjacency pairs of speech acts for children of 14, 20, and 32 months. Utterances by the caregiver are on the left, responses by the children on the right. Filtered to display speech acts that occur in at least 0.01% of the data for better visibility. The colors indicate the higher-level interchange type for each speech act (see C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. 1996).

Instead of estimating the AoA using logistic regressions, we can estimate the ranking without fitting any model and directly from the data. More specifically, we computed the proportion of children that produced (or understood) a given speech act (averaged over the three-time points) and ranked the speech acts according to these proportions as a proxy for their order of acquisition. The resulting rank-order correlations obtained using this model-free method were very close to the correlations

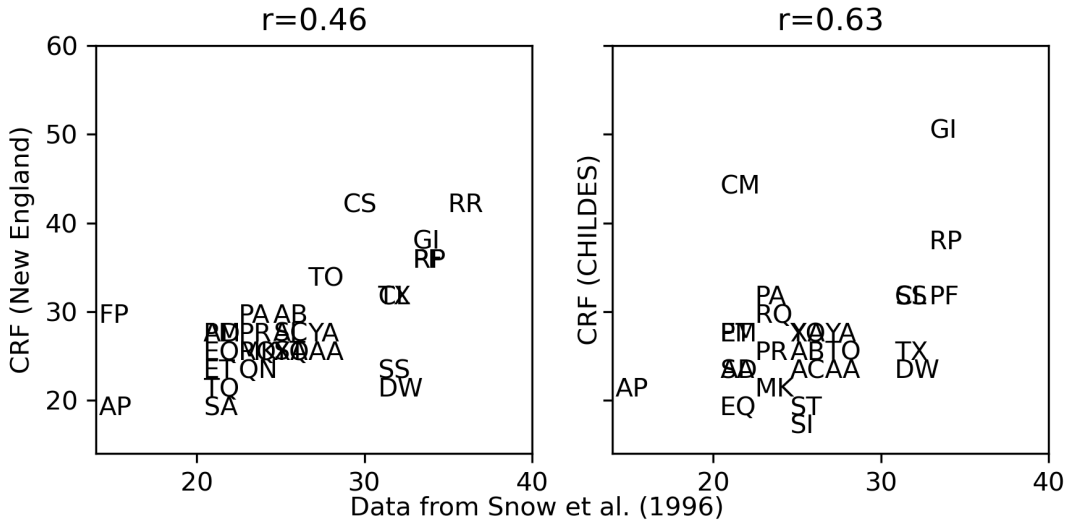

Figure 2.7. – Correlation of age of acquisition in terms of comprehension as calculated using data from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996) and automatically annotated data for the New England corpus and CHILDES. Note that some speech acts are not displayed because the axes limits were set to 60/40 months for better visibility of early development. However, the correlation was calculated for all values.

Figure 2.8 shows the full distribution of age of emergence in both production and comprehension. It shows that, overall, comprehension of speech acts precedes their production. Indeed, a paired t-test (using only speech acts for which we could calculate an age of acquisition both in production and in comprehension) shows a mean difference of 2.51 months ($p < 0.05$). [14]

Finally, we ask how the trajectory of emergence in comprehension compares to that of production. For instance, does production follow the same pattern/order of comprehension, only delayed? Pearson's correlation between the two developmental trajectories is $r \approx -0.07$ ($p \approx 0.76$), indicating that speech acts emerge differently in production and comprehension, and suggesting that these two dimensions of development may be explained by different factors.

2.3.4.3. Development of Speech Acts Beyond 32 Months

Since CHILDES contains data for children beyond the age range studied in C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996), we could also make predictions about the age of acquisition of some speech acts that could not be calculated using the New England corpus because they were not yet acquired by children by

found using the regression method, thus corroborating these findings.

14. When using the alternative coding scheme with collapsed speech act categories (see Section "Error analysis"), this difference increases to 9.61 months ($p < 0.01$).

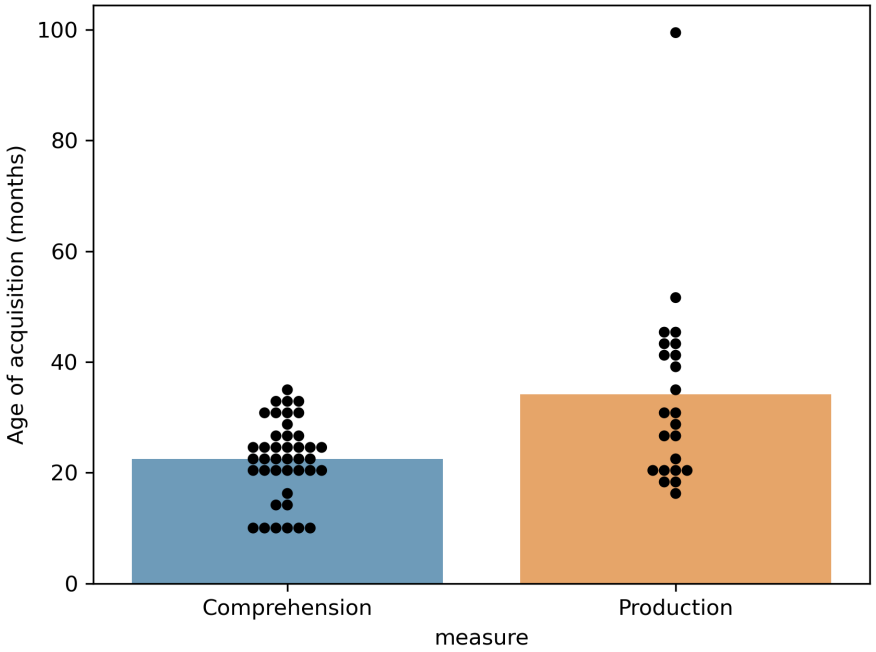

Figure 2.8. – The distribution of the speech acts' age of emergence in comprehension and production.

32 months. To this end, we use all transcripts up to 54 months (data become sparse beyond that age). Using this larger set of annotations, we can for example estimate the age at which children produce speech acts such as prohibitions (PF, at 84.9 months), give reason (GR, at 87.0 months), polite requests (RQ, at 66.2 months), and make promises (PD, at 130.7 months)). These predictions are consistent with the developmental literature showing a late acquisition of some of these speech acts (Matthews 2014). A table of all results can be found in the Appendix.

2.4. Discussion

The way children master language use in social interaction is an important frontier in the study of language development (L. Bloom and Lahey 1978; C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. 1996; Matthews 2014; Eve V. Clark 2018; Casillas and Hilbrink 2020). Answering this question has also the potential for impact in clinical applications (e.g., early and automatic detection of communicative difficulties). However, the investigation of this phenomenon in ecological valid settings requires complex, large-scale data annotation which is prohibitively expensive to do by hand only.

In the current work, we introduced a simple model that allows for reliable *automatic* labeling of major speech act categories in the context of child-caregiver social

interactions. We trained the model on a dataset that was previously hand-annotated using INCA-A, a comprehensive coding scheme for speech acts in early childhood (Ninio, C. E. Snow, Barbara A. Pan, et al. 1994; C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. 1996). When tested on parts of the data it had not seen in the training, the model predicted speech acts that captured quite well the major findings reported in this earlier work such as the average trajectory of speech act development and the patterns of variations between children.

Besides providing a valuable tool that we make available to the community, a major theoretical contribution of this work was testing how earlier findings — obtained using hand annotation of a small number of children — generalize to a larger and different sample. We tested this generality by automatically labeling the entire American English section of CHILDES for speech acts. We found that, across all major analyses, children show, overall, patterns that were very similar to the ones reported by (C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. 1996). The main difference was that older children in the larger dataset produced noticeably more speech act types than children of similar age in the original study (Figure 2.3, bottom). This difference could be due to the fact that the larger dataset contains a richer set of conversational contexts, giving children the opportunity to perform more distinct speech act types. [15]

Anther contribution of this work is the introduction of two measures to quantify the age of emergence of speech acts in children's production and comprehension. We found that these two measures (i.e., comprehension and production) did not correlate, indicating that they provide non-redundant information about development and suggesting that speech acts may develop differently in production and comprehension. In particular, factors that would be relevant for learning in production may not necessarily be the same in comprehension, especially in the rather *asymmetrical* context of child-caregiver interactions.

To illustrate, take the case of "Yes/no requests" (RQ) vs. "yes/no questions for information." (YQ). In production, we replicated C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996)'s finding that children produce yes/no questions as requests later than yes/no questions for information (very few children produced the first act and only at 32 months). This fact is also in line with the literature on politeness which suggests that children produce polite requests quite late (Axia and M. R. Baroni 1985). Interestingly however, in comprehension we found that on average children responded contingently to the yes/no requests at about the same age as they do to yes/no questions for information.

When using automatically annotated data from our model, we found that their predicted measures of age of acquisition correlated to a high degree with the ages of acquisition predicted from manually labeled data, especially in production. In a

15. Another observation was that the proportion of children producing no speech acts (i.e., 0 in Figure 3) at 14 months is noticeably higher in the automatically annotated data than in the original data. This means that our model classified more utterances as unintelligible or utterance without function than the human annotators. We hypothesize that the highly skewed distribution of speech acts in the dataset for children at this age, with many (but not all) utterances actually being without clear function, leads the model to overfit to this case and miss some actually meaningful utterances.

direct application, the model allowed us to estimate the age of acquisition of some late emerging speech acts (e.g., "promise" and "give reason") thanks to automatic labeling of new data children that were older in CHILDES than in the original New England corpus.

While the automatic labeling model provides a high average accuracy score, the per-label scores showed high variability. While, as we argued above, some of this variability can be explained by the frequency of occurrence in the training data and by ambiguities in the definition of some categories in the coding scheme, we speculate that other factors could be in play as well, especially the *linguistic variability* with which a speech act can be expressed. [16]

For example, there is a variety of ways one can express the act of "giving reasons" (GR) in linguistic terms, which makes it relatively hard to recognize based only on the linguistic features of its instances (F-score = 0.3). In comparison, the set of linguistic terms typically used to express, say, the act of "requesting repetition" (RR) or "eliciting question" (EQ) is much more constrained, making their recognition easier (F-scores are 0.53 and 0.81, respectively), although all three categories have roughly similar (low) frequency of occurrence in the data. Take also the case of "stating intent" (SI) and "prohibiting" (PF). Both of these speech acts are similarly frequent (around 300 occurrences), but the F-score for PF is much higher than the one for SI (0.76 and 0.43, respectively). This difference could also be due to the fact that "prohibiting" is much more constrained linguistically than "stating intent."

Researchers have made a similar argument about the role that linguistic variability can have on their learnability by children (e.g. L. Bloom and Lahey 1978). This analogy is to be taken with a grain of salt though. More generally, it is not warranted to make a direct link between the learnability of speech act categories by our model and their learnability by children: In the first case, the model was aimed at optimizing prediction accuracy and had been trained on labeled data. In the second case, children learn without having access to the true labels of the utterances. Models that aim at "discovering" categories in an unsupervised fashion are more likely to be insightful about the learnability of speech act categories by children (e.g. Bergey, Marshall, DeDeo, et al. 2022).

2.4.1. Limitations and Future Work

Our model learns how to recognize speech acts from their linguistic instances only. While the scores were quite good and allowed us to replicate major findings that were obtained using human annotations, future work should seek to build more comprehensive models that integrate multimodal cues — besides verbal language — that likely play a role in signaling communicative intents including vocal and visual cues (e.g. Fernald 1989; Tomasello, Call, and Gluckman 1997; Senju and Csibra 2008; Trujillo, Simanova, Bekkering, et al. 2018). This effort will involve collecting multimodal

16. Indeed, the higher the variability within a given category, the more examples the model needs to learn it.

data of spontaneous child-caregiver conversations (e.g. Bodur, Nikolaus, Kassim, et al. 2021; Shi, Gu, and Vigliocco 2022; Sullivan, Mei, Perfors, et al. 2022) as well as the development of machine learning methods for the automatic annotation of speech acts using linguistic, acoustic, and visual features.

Another limitation concerns the measures we used to quantify the age of acquisition. While it is easier to quantify acquisition through production, it is trickier to have a perfect measure of comprehension in a natural, uncontrolled context. Here, we provided a contingency-based measure. Such an operationalization has allowed us to uncover new interesting phenomena (namely that children understand some speech act before they produce them).

However, measuring contingency is a notoriously difficult task, especially in a naturalistic setting and with verbal data only. First, responses can be contingent in various ways: For example, asking a yes-no question like "Do you want a banana?" can be followed by many speech acts that can all be contingent such as "Yes!", "I just ate one", or "now?". Other speech acts such as declarative statements do not necessarily require a response, so the listener might understand the communicative intent without necessarily giving a response. In this work, we partly avoided these difficulties by using a broad binary annotation that judged whether a response was possibly contingent or totally inappropriate (e.g., a "greeting" after a "yes-no question").

In addition to these theoretical difficulties, there are practical difficulties related to the fact that children (especially the younger ones) may respond contingently but in a non-verbal fashion (a case that is not captured by the current model). Besides, they sometimes respond in an unintelligible fashion (a case which we had to classify as non-contingent). Another case is when they do not respond at all (leading to more data exclusion). However, when children do not respond (e.g., after being asked a question), it does not necessarily mean that they did not understand the speech act. For example, children may lack the appropriate vocabulary to formulate an adequate response or they may just not be interested in following up.

Finally, we did not take into account the timing of responses (as several CHILDES corpora lack timestamps in the transcripts). This is important, because if a child's response only follows a caregiver's utterance after a long temporal delay, it may not be an actually response, but a new initiation. Thus, it would not be appropriate to judge the contingency of this "response" with respect to the caregiver's utterance that preceded it.

All these reasons may contribute to making our contingency measure *under-estimate* children's early age of comprehension. That is, it is very likely that children understand many speech acts at a much earlier age than what we report in this work. That said, some results using this measure, especially the fact that comprehension precedes production in some categories, would still hold. In fact, if anything, a more accurate measure of comprehension would just make such conclusions stronger.

Finally, we found several limitations the INCA-A coding scheme when automatically labeling utterances, including overlapping as well as hierarchically related categories (cf. the error analyses section as well as Cameron-Faulkner (2014) for similar observations). In the future, the coding scheme should be updated in order to make it less

ambiguous for automatic annotation.

To conclude, this work has introduced both novel research tools and measures that we hope will pave the way to a more quantitative approach to the study of children's speech act development in the wild.

3. CF for Learning to Produce Intelligible Speech

This chapter is based on the article "Communicative Feedback as a Mechanism Supporting the Production of Intelligible Speech in Early Childhood (Nikolaus, Prévot, and Fourtassi 2022), published in the *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*.

The review of the literature in Section 1.5 showed that there has been only a limited number of studies that explore the role of Communicative Feedback under naturalistic conditions. Further, many studies deal with very early stages of linguistic development, such as the transition from non-speech to speech-related vocalizations (Warlaumont, Richards, Gilkerson, et al. 2014; Goldstein, King, and West 2003; K. Bloom 1988; K. Bloom 1984). In this chapter, we addressed this gap by executing a large-scale corpus study on naturalistic child-caregiver conversations in English (MacWhinney 2014). We used the speech act annotation model presented in the previous Chapter (2) in order to automatically annotate clarification requests in our large-scale data. [1]

As a first step, we reproduced results from earlier work regarding the development of speech-related sounds (Warlaumont, Richards, Gilkerson, et al. 2014) on a larger and more diverse dataset. Then, we turned to examine the quality of positive and negative Communicative Feedback signals that caregivers provide in terms of time-contingent responses and clarification requests in supporting children's production of intelligible speech.

The results of our analyses suggest that caregivers use clarification requests more often in response to unintelligible utterances than to intelligible ones and children improve their intelligibility when prompted with a clarification request. Regarding the role of implicit feedback, we found that caregivers provide more time-contingent responses to intelligible utterances and children produce more intelligible utterances if their caregivers are responsive.

This study provided evidence that general social feedback mechanisms that govern human communication can support child language acquisition not only in very early stages of speech development, but they can also aid the child to produce their first intelligible words. Further work is required to investigate the universality as well as the limitations of such learning mechanisms, and look into possible effects on the

1. Unfortunately there was no speech acts within the INCA-A coding scheme (Ninio, C. E. Snow, Barbara A. Pan, et al. 1994) that could be directly mapped to the function of acknowledgements. Therefore, we left the analyses of such explicit positive feedback for future work.

learning of even later stages, i.e. the production of grammatical sentences (as covered in Chapter 4).

3.1. Introduction

Much of computational research in language acquisition has traditionally focused on investigating learnability from the linguistic input. While such an approach has been insightful about the role of the input in language development, it tends to consider – whether implicitly or explicitly – that children only passively absorb the information they are exposed to. However, children start to actively interact with people around them very early in development and use their growing linguistic knowledge to establish some form of rudimentary communication. This early social interaction also plays a role in the acquisition of language (Bruner 1985; Tomasello 2003; Kuhl 2007; Eve V. Clark 2018).

Currently, the dominant line of research on the role of social interaction focuses on children's ability to make pragmatic inferences about caregiver's communicative intents, taking into account the context of language use, common ground, as well as social cues such as gaze and pointing (Tomasello, Carpenter, Call, et al. 2005; Senju and Csibra 2008; Yurovsky and Michael C. Frank 2017; Bohn and Michael C. Frank 2019; Tsuji, Jincho, Mazuka, et al. 2020).

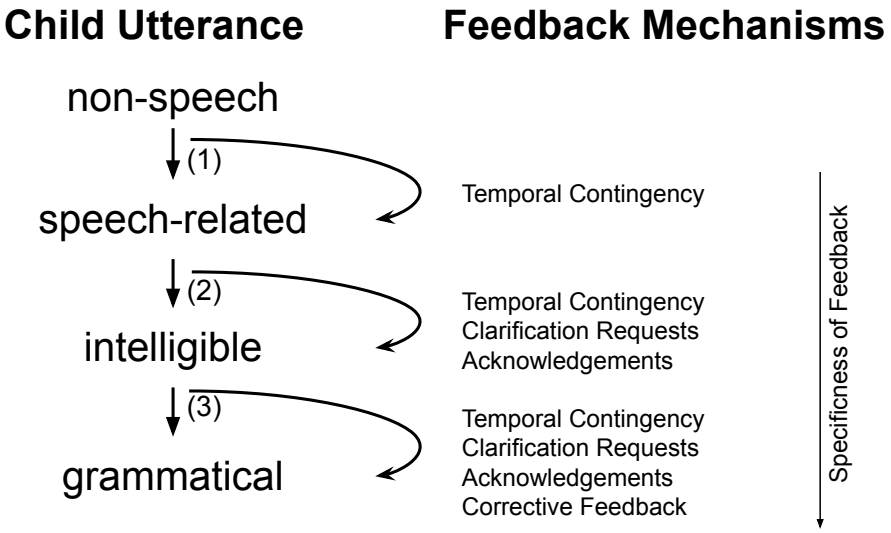

Figure 3.1. – Developmental steps in children's linguistic productions before they reach the final grammatical/well-formed stage. As children move from one stage to the next, the range of available feedback mechanisms and their specificness increases (as the communicative intent of the child becomes increasingly easier to decode). Communicative Feedback (the subject of the current work) includes contingency, clarification requests, and acknowledgements, but excludes corrective feedback.

Another dimension of social interaction is that it offers opportunities for caregivers to provide *feedback* on children's linguistic productions. Children start communicating long before their linguistic skills are mature (i.e., intelligible and grammati-

cally sound). Their vocalizations start from being non-speech (e.g., crying, laughing) and become increasingly speech-related (e.g., babbling), but still largely unintelligible. Then, their linguistic productions become increasingly intelligible[2] although not always grammatical (e.g., "Want play!"). Finally, children's productions become grammatical/well-formed (e.g., "I want to play!"). As children move from one stage to the next, the range of available social feedback mechanisms and their specificness increases (i.e. their ambiguity decreases, see Figure 3.1).

To support the transition from non-speech to speech-related vocalization (Transition (1) in Figure 3.1), the caregiver can provide feedback in the form of *temporal contingency* (e.g., by responding more often and faster to speech-related than to speech-unrelated vocalization, thus "reinforcing" the former), a mechanism that has been studied for example by Warlaumont, Richards, Gilkerson, et al. (2014).[3] Once children become able to produce speech-related utterances (that can be either intelligible or unintelligible), additional feedback mechanisms become possible, e.g., *clarification requests* (M. R. J. Purver 2004) can be given following unintelligible vocalizations, encouraging the child to produce more intelligible speech (Transition (2) in Figure 3.1; see e.g., Eve V. Clark (2020)).[4] Finally, *corrective feedback* (e.g., a correct reformulation of an incorrect utterance by the child) has been studied more extensively in the development literature (e.g., R. Brown 1973; Saxton 2000; Chouinard and Eve V. Clark 2003; Hiller and Fernandez 2016), but this form of feedback can only be provided at the last stage (Transition (3) in Figure 3.1) where children are capable of producing intelligible utterances. Indeed, if an utterance is unintelligible, the caregiver cannot infer the child's communicative intent and, thus, will not be able to *correct* or reformulate its linguistic expression.

3.1.0.1. Communicative Feedback

Here, we focus on a subset of the social feedback mechanisms reviewed above which we call *Communicative Feedback* (CF). We define CF as the form of social feedback whose goal is to signal *communicative success or failure*, rather than to correct the linguistic *content* of a child's production. Therefore, CF includes rather non-specific mechanisms such as time-contingent responses and clarification requests (e.g., "What?", "Which one?").

We are interested in this communication-focused feedback because it has been understudied compared to content-focused feedback (especially corrective feedback), although it is arguably a better candidate for a universal mechanism. Indeed, contrary to corrective feedback, Communicative Feedback is not specific to language

2. We define intelligibility by contrast to unintelligible utterances whose *communicative intent* is difficult to infer (e.g., babbling). An utterance is intelligible if a communicate intent can be decoded from it even though it is not necessarily grammatically correct.

3. This mechanism has also been referred to as "responsiveness". Here we call it temporal contingency in order to distinguish it from other kinds of contingency, namely input-contingency or content-contingency (see also McGillion, Herbert, Pine, et al. 2013).

4. Further, the caregivers can provide explicit positive feedback in the form of acknowledgements. We left the analyses of such feedback for future work.

acquisition, as it is a fundamental mechanism for communication in general (see "communicative grounding"; Stalnaker (1978) and H. H. Clark (1996)). Further, both communicative repair mechanisms and caregiver's temporal contingency have been observed in a diversity of cultures (Dingemanse, Roberts, Baranova, et al. 2015; Richman, Miller, and LeVine 1992; Marc H. Bornstein, Catherine S. Tamis-LeMonda, Tal, et al. 1992). Finally, CF is the only mechanism that can a priori be used across all stages of child utterance development, as explained in Figure 3.1.

3.1.0.2. Communicative Feedback and language acquisition

Though CF is more about communication management than about correcting or refining the content of the child's utterance, it can still help with language learning, indirectly, via a reinforcement-like mechanism. CF can be positive or negative, signaling communicative success or failure, respectively. The hypothesis is that the child would seek positive signals and avoid negative signals, motivated by the desire to be understood. Utterances that receive positive feedback are more likely to be correct and will be repeated, whereas utterances that receive negative signals are more likely to be incorrect and will be avoided or revised in future interactions.

Negative CF signals include a) lack of contingency (e.g., the caregiver providing a non-contingent response, or no response at all) and b) clarification requests, which could be verbal or non-verbal (e.g., "What?" or a puzzled face). Positive CF signals include a) high contingency (e.g., fast and on-topic verbal responses, successful shared attention) and b) explicit verbal and non-verbal signs of understanding (backchannel responses, repeating the child's utterance, a cheering face and a head nod, etc.).

We find in the development literature evidence that supports CF as a mechanism for language learning, although most research has investigated only the early stages of utterance development described in Figure 3.1. For example, it has been shown that caregivers are more responsive to child speech-related utterances than to non-speech utterances and that, critically, children also react differently to positive and negative CF, producing more speech-related utterances following high temporal contingency (K. Bloom 1988; Goldstein, King, and West 2003; Warlaumont, Richards, Gilkerson, et al. 2014).

Very few studies examined CF for later stages of development. Eve V. Clark (2020) documented caregiver's use of clarification requests and Gallagher (1977) and Saxton, Houston–Price, and Dawson (2005) tested how children revise their utterances when they receive such requests. However, though very insightful to our understanding of the phenomenon, this previous work remains incomplete as it has either relied on qualitative/anecdotal reporting or on experimentally controlled settings (as opposed to systematic and quantitative study of natural/spontaneous child-caregiver conversations).

3.1.0.3. The current study and novelty of our work

We conduct a quantitative large-scale corpus study on the role of CF in children's language development. This work makes two main contributions. First, we test the extent to which previous work by Warlaumont, Richards, Gilkerson, et al. (2014) — on how temporal contingency can help children transition towards speech-related utterances (i.e., Transition (1) in Figure 3.1) — can be replicated with different datasets of child-caregiver interactions. Second, we investigate how CF (both temporal contingency and clarification requests) can help children transition towards more intelligible utterances (i.e., Transition (2) in Figure 3.1).

To ensure reproducibility, we make the source code of all analyses publicly available: https://github.com/mitjanikolaus/childes-communicative-feedback.

3.2. Methods

3.2.1. Unit of analysis: U-R-F

To study CF in child-caregiver conversations, we use as unit of analysis a 3-part micro-structure sequence consisting of 1) child's utterance, 2) caregiver response (or lack thereof), and 3) the child follow-up (following previous work like Warlaumont, Richards, Gilkerson, et al. (2014)). [5] Hereafter, we will call this sequence U-R-F (Utterance, Response, Follow-up). Table 3.1 shows some examples of U-R-F from the dataset we used.

Table 3.1. – Examples of U-R-F sequences taken from the Thomas corpus (Lieven, Salomo, and Tomasello 2009).

Child utt.	**Caregiver resp.**	**Child follow-up**
Moon	A big moon.	And a firework.
Uh no big smoke .	<no response>	xxx .
[=! babble]	what?	put in there please.

3.2.2. Data

We used transcribed conversations from an English subset of the CHILDES corpus (MacWhinney 2014). The subset involves children aged 10 to 48 months [6] for which timing information (start and end time of each utterance) is available. We converted the children's ages into equidistant bins of 6 months for plotting and analyses. In total,

5. We disregard case where the follow-up occurs more than 60s after the response.

6. We chose 10 months as a minimum age because at this age children typically start to produce their first intelligible utterances. As a maximum age, we chose 48 months because data in CHILDES becomes sparse after this age.

our data consists of 21 corpora [7] with 1787 transcripts from 326 children. We extracted and analyzed a total of 367,774 U-R-F sequences.

3.2.3. Annotations

For each U-R-F, we annotate the **speech-relatedness** and **intelligibility** of all child utterances and follow-ups, as well as whether there was a caregiver **response** and whether the response was a **clarification request.**

3.2.3.1. Speech-relatedness

All corpora in CHILDES follow the CHAT transcription format (MacWhinney 2017) which includes so called "paralinguistic events". [8] All utterances that contain at least one transcribed word or one speech-related event were annotated as speech-related, others as non-speech.

3.2.3.2. Intelligibility

We labeled all utterances as either intelligible or unintelligible using a rule-based approach on the transcriptions. Not all corpora transcribed unintelligible speech exactly the same way, so we manually verified which conventions were used in each corpus.

In CHAT, phrases are either explicitly labeled as unintelligible ("xxx") or labeled as phonological fragments (e.g., "&baba", "baba@p"). The latter case is used for "vocalizations that cannot be mapped to words" (and are therefore coded phonetically instead, see MacWhinney 2017). As they cannot be mapped to words, we assume they are also unintelligible to the interlocutor. [9] Further, there are some event codes that refer to unintelligible utterances and babbling (e.g., "&=vocalize", "&=babble", "baba@b"). Utterances that contain at least one unintelligible word were labeled as unintelligible, all others as intelligible.

3.2.3.3. Temporal contingency

While the contingency of caregiver response behavior can by measured in many ways (McGillion, Herbert, Pine, et al. 2013), here we focus on one instantiation of contingency known as temporal contingency. Using timing information available in the transcripts, we annotate for each child's utterance whether the caregiver response is given or not. Following Warlaumont, Richards, Gilkerson, et al. (2014), we considered

7. Namely: Bernstein, Bloom, Braunwald, Brent, Edinburgh, Gleason, MPI-EVA-Manchester, MacWhinney, McCune, McMillan, Nelson, NewmanRatner, Peters, Rollins, Sachs, Snow, Soderstrom, Thomas, Tommerdahl, VanHouten, Weist

8. Events in CHILDES can be transcribed either as "paralinguistic events" ("[=! crying]"), or "simple events" ("&=crying").

9. There may however be exceptions in which the utterance remains intelligible, we will return to this case in the discussion.

all cases in which a caregiver's utterance follows the child's utterance within a response latency of one second as **response**. All other cases (a caregiver's utterance that follows with a greater delay, or no utterance at all) are considered as **no response**. [10]

3.2.3.4. Clarification requests

To detect clarification requests, we used a model that was recently developed for automatic annotation of speech acts in child-caregiver conversations (Nikolaus, Maes, Auguste, et al. 2022). This model uses the INCA-A coding scheme, which was specifically designed for the study of child-caregiver conversations (Ninio, C. E. Snow, Barbara A. Pan, et al. 1994).

All utterances that were labeled as "Eliciting questions (e.g., hmm?)" (EQ) or "Requests to repeat utterance" (RR) were treated as clarification requests. The most common utterances falling into these categories are open clarification requests, e.g., "what?", "hm?", "what, darling?", "huh?". Less frequently, there are also restricted ones such as "what about backside?" or "some what?".

3.3. Analyses

Our analyses are organized into three main parts: 1) replicating work by Warlaumont, Richards, Gilkerson, et al. (2014) on the development of speech-relatedness via temporal contingency, 2) investigating the development of speech intelligibility via temporal contingency, and 3) investigating the development of speech intelligibility via clarification requests.

3.3.0.1. General developmental trajectories

Since we are both replicating work on the development of speech-related speech and investigating the development of intelligible speech, we start our analysis by providing an overview of the developmental trajectories of both phenomena in our dataset (Figure 3.2). As expected, children's utterances become increasingly speech-related as well as increasingly intelligible. The proportion of speech-related utterances converges clearly before the proportion of intelligible utterances.

3.3.1. Development of Speech-related Vocalizations (Replication of Warlaumont, Richards, Gilkerson, et al. (2014))

Following Warlaumont, Richards, Gilkerson, et al. (2014), we first calculated a measure for caregiver's temporal contingency on child speech-relatedness. This measure

10. We also ran all experiments with a more conservative response latency threshold (2 seconds) and this higher threshold did not change the conclusions of this work.

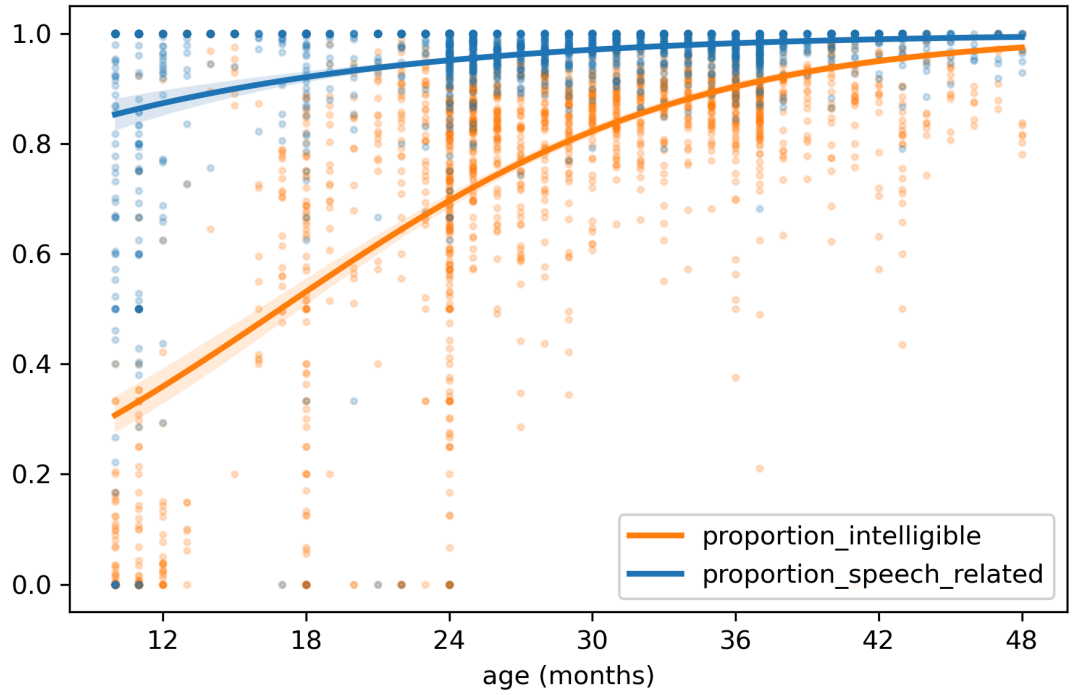

Figure 3.2. – Proportion of speech-related utterances and intelligible utterances. Each data point represents a transcript. The plot shows fitted logistic regression curves and their 95% confidence intervals.

was defined as the ratio of caregiver responses to speech-related child utterances subtracted by the ratio of caregiver responses to non-speech utterances. When applied to our dataset, we obtain:

$$\frac{\#(U_{speech} \wedge R_{response})}{\#U_{speech}} - \frac{\#(U_{non-speech} \wedge R_{response})}{\#U_{non-speech}} \approx 0.13 \tag{3.1}$$

A one-sided t-test indicated that the value was significantly greater than 0 ($SE = 0.008$, $p < 0.001$), replicating the original results, and confirming that temporal contingency contains useful information for learning speech-related vocalizations.

Second, we calculated a measure for the child's follow-up depending on whether there was a caregiver's response to the child's utterance. This measure was defined (in the previous study) as the ratio of speech-related child follow-ups to speech-related utterances that received a response subtracted by the ratio of speech-related child follow-ups to speech-related utterances that did not receive a response:

$$\frac{\#(U_{speech} \wedge R_{resp} \wedge F_{speech})}{\#(U_{speech} \wedge R_{resp})} - \frac{\#(U_{speech} \wedge R_{no\_resp} \wedge F_{speech})}{\#(U_{speech} \wedge R_{no\_resp})} \approx 0.01 \tag{3.2}$$

This value was also significantly positive ($SE = 0.003$, $p < 0.05$), again replicating the results of the original study and confirming that children are sensitive to temporal

contingency when learning speech-related vocalizations. [11]

Note that, for comparison with Warlaumont, Richards, Gilkerson, et al. (2014), our replication study used the same measures. However, for our next analyses, we will use mixed-effect models instead, because they allow more rigorous statistical testing as well as the ability to control for other variables. We also ran the equivalent mixed-effects models for this replication, and the results confirmed the significance of both effects, even when controlling for age.

3.3.2. Development of Intelligibility via Temporal Contingency

Following the general reasoning in Warlaumont, Richards, Gilkerson, et al. (2014), we first study the extent to which caregiver's time-contingent response behavior depends on the intelligibility of the child's utterance. Second, we study the extent to which the child's follow-up show improved intelligibility when following responsive behavior from the caregiver.

3.3.2.1. Caregiver's temporal contingency

Figure 3.3 shows the results of how caregivers' response behavior depends on the intelligibility of the children's utterances.

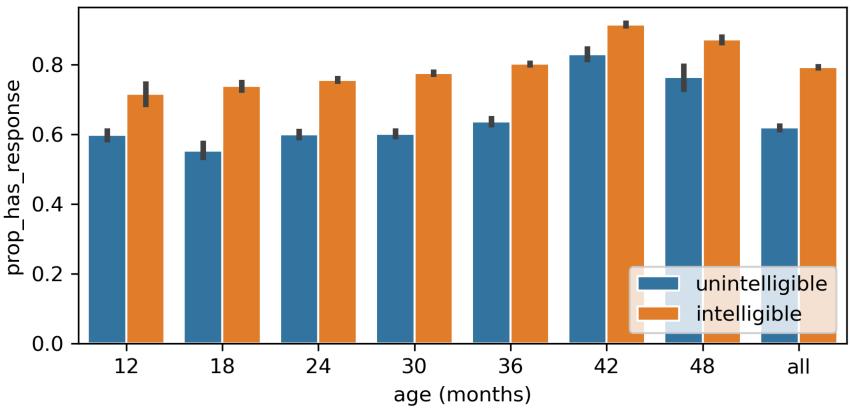

Figure 3.3. – Comparison of proportion of caregiver responses for intelligible and unintelligible child utterances.

To quantify this effect, we used the following mixed-effects GLM that predicts whether a caregiver response was given as a function of whether the child utterance was intelligible:

11. Warlaumont, Richards, Gilkerson, et al. (2014) obtained the following values: 0.065 for caregiver contingency and 0.036 child contingency. The difference in effect sizes could arise from differences in the properties of the datasets used in the original vs. the replication, e.g., different conversational contexts or because in our data (which relies on transcriptions instead of automatic speech classification) probably not all non-speech utterances were transcribed exhaustively.

$$\texttt{has\_resp} \sim \texttt{utt\_is\_intelligible} * \texttt{age} + (1|\texttt{child}) \quad (3.3)$$

The estimated fixed effects were: `utt_is_intelligible` : $\beta = 0.81, SE = 0.01, p < 0.001$; `age` : $\beta = 1.23, SE = 0.03, p < 0.001$; `utt_is_intelligible:age` : $\beta = -0.29, SE = 0.05, p < 0.001$.

These results confirm the qualitative observations in Figure 3.3, that is, `utt_is_intelligible` is a predictor of caregiver's response contingency (more intelligible utterances leads to more contingency and vice versa). This effect was significant even controlling for age.

3.3.2.2. Child sensitivity to temporal contingency

In Figure 3.4, we show how the intelligibility of the child's follow-up depends on whether there was a caregiver's response to an intelligible child's utterance.

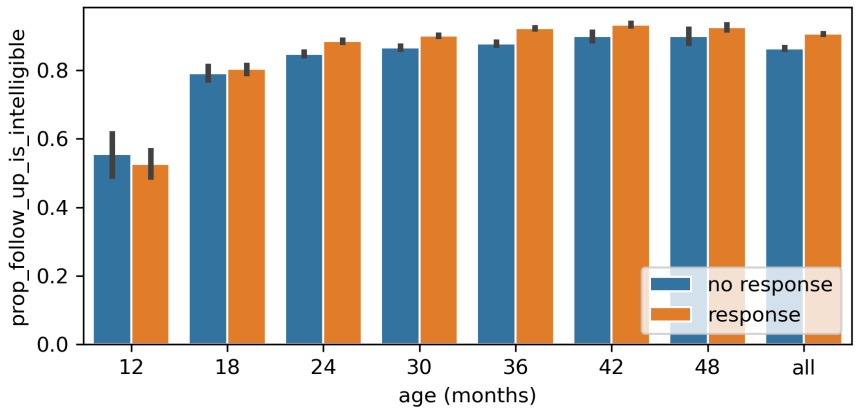

Figure 3.4. – Comparison of the proportion of intelligible follow-ups depending on whether the child's previous intelligible utterance received a response from the caregiver or not.

To quantify this effect, we similarly used a GLM that predicts whether a child follow-up is intelligible as a function of whether there was a caregiver's response or not, taking into account only U-R-Fs for which the initial utterance by the child was intelligible:

$$\texttt{follow\_up\_is\_intelligible} \sim \texttt{has\_resp} * \texttt{age} + (1|\texttt{child}) \quad (3.4)$$

The estimated fixed effects were: `has_resp` : $\beta = 0.38, SE = 0.01, p < 0.001$; `age` : $\beta = 0.77, SE = 0.04, p < 0.001$; `has_resp:age` : $\beta = 0.12, SE = 0.06, p = 0.06$.

Here again, the statistical analysis confirms the qualitative observations in Figure 3.4, that is, there was a positive impact of caregiver's responses (`has_resp`) on children's follow-up being more intelligible. This effect was significant above and beyond the child's age.

3.3.3. Development of Intelligibility via Clarification Requests

We first study the extent to which caregivers' clarification requests are dependent of the intelligibility of the child's utterance. Second, we study the extent to which children's follow-ups increase in intelligibility after a clarification request.

3.3.3.1. Caregiver's clarification requests

Figure 3.5 shows how caregiver's clarification requests depend on the intelligibility of the preceding child utterance across development.

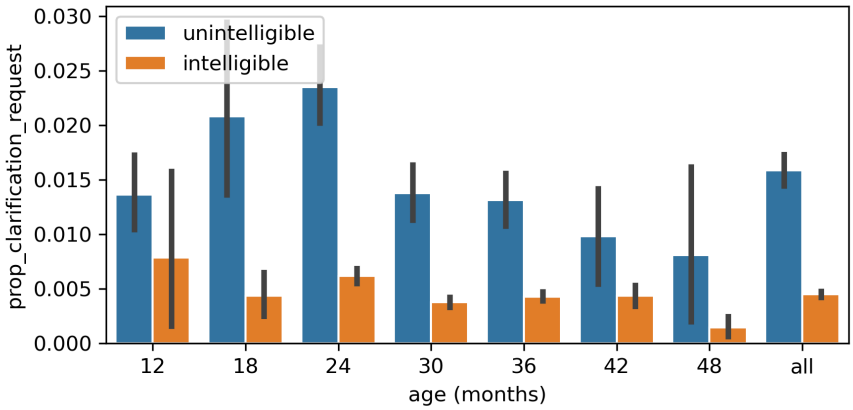

Figure 3.5. – Comparison of proportion of clarification requests for intelligible and unintelligible child utterances.

We used the following GLM predicting whether the caregiver's response is a clarification request, as a function of whether the child's utterance was intelligible:

$$\texttt{resp\_is\_clar\_req} \sim \texttt{utt\_is\_intelligible} * \texttt{age} + (1|\texttt{child}) \quad (3.5)$$

The estimates were: $\texttt{utt\_is\_intelligible}: \beta = -1.14, SE = 0.06, p < 0.001$; $\texttt{age}: \beta = -0.59, SE = 0.15, p < 0.001$; $\texttt{utt\_is\_intelligible}:\texttt{age}: \beta = 0.02, SE = 0.21, p = 0.9$.

As expected, we found a negative effect of `utt_is_intelligible` showing that clarification requests are more likely to be made by the caregiver after unintelligible utterances from children. This effect was also significant above and beyond age.

Also, there are less clarification requests by the caregiver with increasing child age but no significant interaction term, indicating that the relationship between intelligibility and clarification requests does not vary with increasing age.

3.3.3.2. Child sensitivity to clarification requests

Here, we investigate if caregiver's clarification requests lead to more intelligible follow-ups from children. For this analysis, we do not compare the intelligibility of child follow-ups as a function of the presence vs. absence of clarification request (similarly to the analysis on sensitivity to temporal contingency, where we compare the distinction response vs. no response), because the temporal-contingency-based mechanism creates a confound. More precisely, when caregivers do not give a clarification request, this is usually because the child utterance was already intelligible and is thus likely to receive a response from the caregiver, leading to the continuation of intelligibility in child follow-up. In both cases (presence vs. absence of clarification request), we can predict high follow-up intelligibility, hence the confound we need to avoid.

Thus, to be able to test the specific effect of clarification request without interference from the temporal contingency mechanism, we compare the intelligibility of the follow-up to that of the child's previous utterance within the *same* U-R-F.

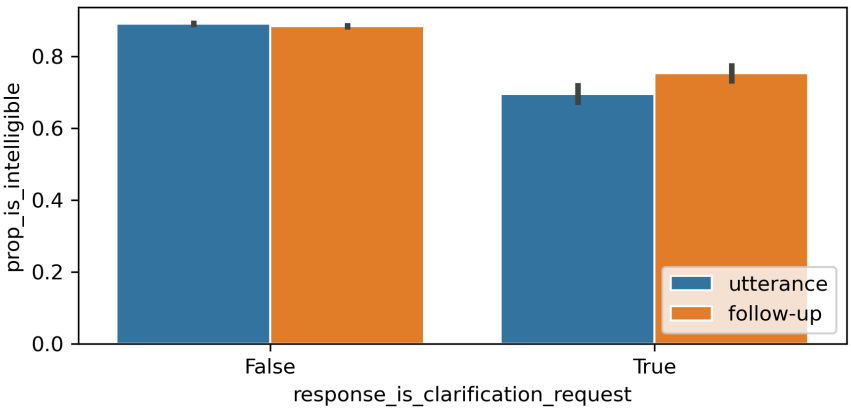

Figure 3.6. – Comparison of proportion of intelligible utterances before (utterance) and after (follow up) clarification requests and other responses.

Figure 3.6 compares the effect of the caregiver's response (presence vs. absence of clarification requests) on the intelligibility of utterances before and after the response. We observe that in the absence of a clarification request, *both* the child's follow-up and her previous utterance are more intelligible. This observation illustrates the confound previously mentioned: Comparing intelligibility of follow-ups alone would be misleading. However, when we compare intelligibility before and after a response, we observe that intelligibility improved more when the response is a clarification request (right side of Figure 3.6). We quantify this effect through testing an interaction in the following model, demonstrating that children are sensitive to this kind of CF and that they can use it to improve their intelligibility:

$$\texttt{is\_intelligible} \sim \texttt{resp\_is\_clar\_req} * \texttt{is\_follow\_up} + (1|\texttt{age}) + (1|\texttt{child}) + (1|\texttt{urf\_id}) \quad (3.6)$$

The estimates were as follows: resp_is_clar_req : $\beta = -1.02, SE = 0.04, p < 0.001$; is_follow_up : $\beta = 0.14, SE = 0.03, p < 0.001$; and more importantly, the interaction term: resp_clarification_req*is_follow_up : $\beta = 0.46, SE = 0.05, p < 0.001$. The positive interaction term demonstrates that the difference between before and after is larger in the case of clarification request responses, than it is in other responses.

Next, we zoom in on the case of clarification requests (bars on the right in Figure 3.6, and we test whether the observed effect holds over development (Figure 3.7).

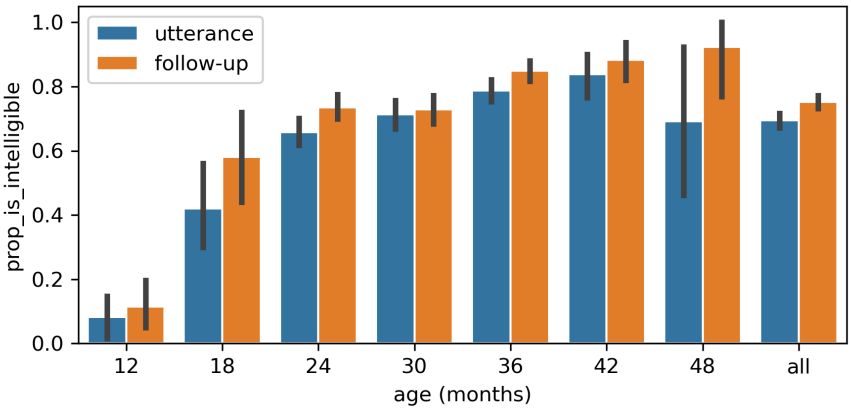

Figure 3.7. – Comparison of proportion of intelligible utterances before (utterance) and after (follow up) a clarification request.

We used the following model, taking into account only the subset of U-R-Fs where the response is a clarification request:

$$\texttt{is\_intelligible} \sim \texttt{is\_follow\_up} * \texttt{age} + (1|\texttt{child}) + (1|\texttt{urf\_id}) \quad (3.7)$$

The estimates were: is_follow_up : $\beta = 0.44, SE = 0.1, p < 0.001$; age : $\beta = 2.74, SE = 0.43, p < 0.001$, is_follow_up:age : $\beta = 0.06, SE = 0.52, p = 0.9$, indicating that the effect remains significant above and beyond age. The intelligibility of children's utterances increased after receiving negative feedback in the form of a clarification request from the caregiver.

3.4. Discussion

The broad goal of this work was to investigate how some general social feedback mechanisms that are part of human communication (H. H. Clark 1996) can help

children learn language. As a case study, we explored how this feedback can support children to produce intelligible speech.

What is special about this feedback (which we call communicative feedback, CF) is that it does not aim to correct or refine the content of the child's utterances (as in the case of corrective feedback). It only seeks to establish/repair communication via positive or negative signals of understanding. We argued that CF can help with language development indirectly: As children seek to be understood, they are sensitive to social signals that indicate whether their communicative intent (e.g., requesting an object or seeking attention) was successfully achieved, and they revise/adjust their expression if necessary, aligning with the correct linguistic conventions.

We provided evidence that CF is useful for learning how to produce intelligible speech. To this end, we investigated not only *positive* CF (temporal contingency) but also one kind of *negative* CF (clarification requests).

Our results indicated that both are contingent on the intelligibility of child utterances across all observed ages, thus providing a useful feedback signal. Critically, we also found evidence that children are sensitive to these signals and produce more intelligible utterances if their caregivers are responsive and improve the intelligibility of utterances if the caregiver asks for clarification.

Limitations and future research directions This work relied on publicly available transcriptions of child-caregiver interactions in naturalistic environments. The recordings could be of varying quality and there might be cases in which utterances are transcribed as "unintelligible" because of poor audio quality, or noise. These cases are a confound to our analyses, since we considered all such cases as unintelligible to the caregiver, while they might just be unintelligible to the transcribing person. That said, manual verification of several examples suggested that in most cases the utterances were most likely also unintelligible to the interlocutor. Further, we observed a continuous increase of the intelligibility of children's utterances in our corpora (see Figure 3.2, which indicates that the intelligibility is not (only) a phenomenon of the transcription but an indicator of the children's linguistic development.

We quantified children's sensitivity to CF by measuring its effect on immediate child follow-ups and observed a significant influence. This operationalization could *overestimate* actual learning effects, which could be forgotten in the long term. However, it could also *underestimate* learning effects which may become visible only at a later point in time. Future research is required to explore the long-term effects of CF.

Regarding negative CF signals, we studied the role of clarification requests. While we were able to demonstrate their usefulness, their presence in the observed conversations was rather scarce (We analyzed a total of 2235 clarification requests, which formed 0.5% of all U-R-F sequences).

In future research, many other positive and negative CF signals could be quantified, including facial expressions (e.g., frowning as negative feedback), actions (e.g., providing requested objects as positive feedback), and content contingency (e.g., responding on-topic as positive feedback). This effort will require collecting more multimodal data of child-caregiver conversations where such cues can be captured, as well as the

development of machine learning methods that can perform annotation at scale.

4. CF for Learning to Produce Grammatical Speech

This chapter is based on the article "Communicative Feedback in Response to Children's Grammatical Errors" (Nikolaus, Prévot, and Fourtassi 2023), published in the *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

In the previous chapter, we provided evidence for the role of Communicative Feedback for a transition from speech-related to intelligible vocalizations. The following chapter deals with even later stages of linguistic development: The transition from intelligible words to grammatical utterances.

Previous work on the development of grammatical speech has either solely focused on the linguistic input, or focused on the role of *corrective* feedback. This chapter provides a first study on grammaticality within the framework of Communicative Feedback. We investigated the role of Communicative Feedback in response to children's grammatical errors in naturalistic child-caregiver dialog. To support our corpus study, we again used the speech act annotation model presented in Chapter 2. However, we found that our model predominantly detected only a subset of clarification requests. We therefore developed additional methods for the detection of repetition-based clarification requests. Additionally, we automatically detected caregiver's acknowledgements in the conversations using both keyword-based and repetition-based methods.

We found evidence for both positive and negative feedback signals that are useful for learning the grammar of one's native language: Caregivers are more likely to provide acknowledgments if an utterance is grammatical, and they are more likely to ask for clarification if an utterance is ungrammatical. Further, we investigate how children react in response to negative Communicative Feedback signals and find evidence that grammaticality is improved in direct follow-ups to negative feedback signals. This study provides the largest and most comprehensive evidence supporting the presence and effectiveness of Communicative Feedback signals in grammar learning, broadening the literature on Communicative Feedback in language acquisition more generally: We find that such mechanisms can support language acquisition not only in the very early stages of development (such as learning to produce speech-related and intelligible words), but can in principle also support the learning of one's native language's grammar.

While the current study provided correlational evidence, further investigations regarding *causal* relationships between Communicative Feedback and grammar learning are required to achieve more definite conclusions. To this end, one approach could

be to design controlled experimental paradigms in which the effects of Communicative Feedback on grammar learning are tested.

However, some effects of language learning could only become visible in more long-term learning setups, which are not feasible to study within the scope of a single experiment. In order to shed light on these effects, corpus studies on longitudinal data as well as computational simulations, in a similar fashion as outlined in Part III, could be employed.

4.1. Introduction

Long before their linguistic skills are fully developed, children engage in conversational exchanges with people around them, which allows them to refine their linguistic knowledge by leveraging various signals of language use in interaction (Bohn and Michael C. Frank 2019; Tomasello 2003; Eve V. Clark 2018; Bruner 1985). One such signal is *corrective* feedback (and its variants such as negative evidence, reformulations, or recasts). It describes situations in which the caregiver provides the child with a corrected form of an erroneous utterance ("I goed to school." - "You *went* to school?"). This phenomenon has been studied extensively in the developmental literature, but the research community has not reached a consensus regarding its availability to children and/or its effectiveness for first language acquisition (e.g., Chouinard and Eve V. Clark 2003; Hiller and Fernandez 2016; Nelson, Carskaddon, and Bonvillian 1973; Marcus 1993; R. Brown and Hanlon 1970; Farrar 1992; Eve V. Clark 2020; Saxton, Backley, and Gallaway 2005; Demetras, Post, and C. E. Snow 1986; Morgan, Bonamo, and Travis 1995).

In the current study, we focus on the role of another kind of social signal that has come to be called *Communicative* Feedback (hereafter, CF). CF represents signals that the listener (here, the caregiver) sends to the speaker (i.e., the child) to indicate communicative success or failure depending on whether the listener thinks they understood the meaning intended by the speaker (for an overview, see Nikolaus and Fourtassi 2023). The main difference with corrective feedback is that CF does not necessarily aim at correcting the child but rather at reaching and maintaining mutual understanding between interlocutors (H. H. Clark 1996). Despite having a communicative rather than a teaching agenda, the child can still use such signals to learn, either by revising their erroneous linguistic assumptions or by confirming/reinforcing their correct knowledge (Nikolaus and Fourtassi 2021b).

Suppose the child produces an erroneous utterance. Even if the child is not corrected (as in the case of recasts), they can receive *negative* signals of communication breakdown (e.g., a clarification request) that they can use to revise the expression of their communicative intent ("Went to school." - "Who went to school?" - "He went to school."). [1] If the child produces a well-formed utterance, the caregiver can provide acknowledgments (e.g., backchannels), offering the child *positive* feedback of communication success and indirectly confirming their current hypotheses about the linguistic structures they used ("He went to school." - "Oh I see!").

Previous work has provided evidence for the role of such feedback in children transitioning from non-speech to speech-like vocalizations (i.e., babbling, Goldstein, King, and West 2003; Warlaumont, Richards, Gilkerson, et al. 2014; Lopez, Walle, Pretzer, et al. 2020) and for transitioning to the first intelligible words (Nikolaus, Prévot, and Fourtassi 2022). Here, we investigate whether similar communicative signals are available and helpful to children in regard to the development of grammatical speech (at both the morphological and syntactic levels).

1. See also Saxton (2000) about the difference between negative *evidence* and negative *feedback*.

We consider negative CF when caregivers provide clarification requests ("Went to school." - "What?") and positive CF when caregivers provide backchannel responses (e.g., "uh-huh") or acknowledge repetitions ("He went to school." - "He went to school."). Some previous work did study the role of clarification requests and of exact repetitions as a (weak) learning signal (Demetras, Post, and C. E. Snow 1986; Bohannon and Stanowicz 1988), but findings from these studies have been difficult to interpret, given several methodological issues and contradicting results (Morgan and Travis 1989; Morgan, Bonamo, and Travis 1995; Marcus 1993). More recently, the effects of negative feedback have been revisited in a corpus study of one child (Saxton 2000) as well as in an intervention paradigm (Saxton, Houston–Price, and Dawson 2005). The findings suggest that children are indeed responsive to negative feedback as shown by an increase of grammatical follow-ups in response to error-contingent clarification requests.

The current study Here, we present the largest (in terms of sample size) and most comprehensive corpus study of CF for grammatical errors in child-caregiver naturalistic conversations. We considered a wide range of positive and negative CF signals, including exact as well as partial repetitions ("I went to school." - "You went to school."), backchannel responses, and clarification requests of various kinds (open and restricted requests, as well as recasts). Thanks to automatic measures, we analyzed these cues in large-scale data of English-learning children conversing with their caregivers (MacWhinney 2014). We tested both 1) the usefulness of CF as reliable signals to children (i.e., more negative CF following ungrammatical utterances and more positive CF following grammatical utterances), and 2) the effect of these signals on children's grammatically as reflected in children's immediate follow-up utterances.

To ensure reproducibility, we make the source code of all analyses publicly available: https://github.com/mitjanikolaus/childes-communicative-feedback.

4.2. Methods

4.2.1. Data

We analyzed 3-part micro-structure sequences consisting of 1) child's utterance, 2) caregiver response, and 3) the child follow-up (following previous work like Warlaumont, Richards, Gilkerson, et al. (2014), Nikolaus, Prévot, and Fourtassi (2022), and Bavelas, Gerwing, and Healing (2017)). Hereafter, we will call this sequence URF (Utterance, Response, Follow-up). An example for such a sequence could look as follows:

> Utterance (Child): *Need some milk*
> Response (Caregiver): *Hm?*
> Follow-up (Child): *I need some milk.*
>
> — EllisWeismer corpus, LT/42pc/22175.cha

Our analyses are based on transcribed conversations from a subset of the English CHILDES corpora (MacWhinney 2014). We follow Nikolaus, Prévot, and Fourtassi (2022) for the extraction of URF sequences and discard all sequences that contain non-speech and non-intelligible utterances.

4.2.2. Annotations and Corpus Selection

4.2.2.1. Grammaticality

Corpora in CHILDES follow the CHAT transcription format (MacWhinney 2017), which supports the annotation of grammatical errors using dedicated coding schemes. In order to obtain a better understanding of the quality and quantity of errors annotated, we grouped all annotated errors into error type classes, using a coding scheme slightly adapted from Hiller and Fernandez (2016) and Saxton, Houston–Price, and Dawson (2005).[2] A list of error types can be found in the legend of Figure 4.1. We excluded all utterances for which no obvious mapping could be made (sometimes error annotations are annotations of slang, such as "I'd [: I would]", or "I was runnin' [: running]"; these cases are not considered grammatical errors).

Figure 4.1 presents the proportion of child errors for different corpora. We find that there are only a few corpora in which a substantial number of errors were annotated. Inspecting more closely the distribution of error types, it becomes apparent that certain corpora only focused on certain error types (e.g., in the Kuczaj corpus, there were almost exclusively `tense_aspect` errors annotated). In the following, we only considered corpora that (1) included at least 1% errors in child utterances and (2) included a range of different error types annotated. This left us we the following set of seven candidate corpora: Thomas, Providence, MPI-EVA-Manchester, Braunwald, Lara, EllisWeismer, and Bates.

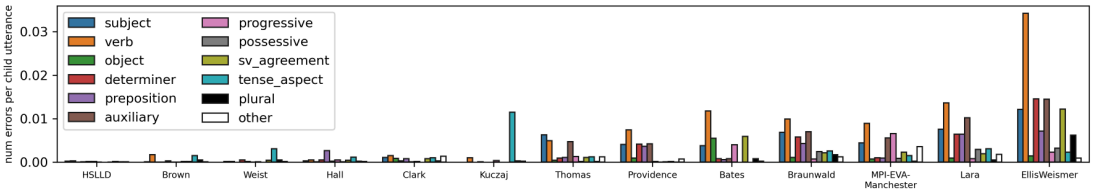

Figure 4.1. – Proportion of child errors normalized by total number of child utterances. We only display English CHILDES corpora that include at least 100 annotated errors.

For these candidate corpora, we performed some manual annotations as a sanity check for our research purposes. For each corpus, we randomly selected 3 transcripts.

2. To simplify their coding scheme, we don't distinguish errors of omission/insertion/substitution and group regular and irregular past tense errors in the group `tense_aspect` (thereby also including errors with e.g. participles). Further, we merge regular and irregular plural errors and include all kinds of subject-verb agreement errors (third-person s, wrong use of is/are) in the group `sv_agreement`).

The first author annotated grammaticality (on a binary scale) of all children's utterances until we reached a threshold of 100 annotated utterances within the transcript.

Table 4.1. – Inter-annotator agreement (Cohen's κ), Precision, and Recall for the grammatical error annotations in 7 different corpora.

	Cohen's κ	Precision	Recall
Bates	0.43	0.89	0.31
Thomas	0.05	1.00	0.03
MPI-EVA-Manchester	0.41	0.80	0.31
Providence	0.65	0.92	0.52
Braunwald	0.49	0.86	0.42
Lara	0.62	0.77	0.57
EllisWeismer	0.68	0.91	0.59

As shown in Table 4.1, inter-annotator agreement between our annotations and the annotations in CHILDES vary to a large degree. Importantly, however, we find that error precision is overall high and recall rather low, indicating that numerous errors were not annotated in CHILDES, but the errors that *were* annotated are generally agreed upon. As the exact annotation guidelines are unfortunately not available for the error annotations in these corpora, we can only speculate that annotators were probably not focusing solely on the error annotations, but performing them as part of the transcription process.

For the remainder of this work, we only consider corpora for which we obtained substantial agreement scores: Providence (Demuth, Culbertson, and Alter 2006), Lara (Rowland and Fletcher 2006), and EllisWeismer (Moyle, Weismer, Evans, et al. 2007). These contained a total of 109,536 micro-structure sequences (URF) from 664 transcripts of 127 children. 5,593 (5.1%) of the children's initial utterances were ungrammatical, and 4,049 (3.7%) of their follow-ups. The children were 12 to 48 months old.

4.2.2.2. Clarification Requests

We annotated clarification requests using two complementary approaches. First, we use a model for automatic annotation of speech acts in child-caregiver conversations (Nikolaus, Maes, Auguste, et al. 2022) and select all utterances that are labelled as "Eliciting questions (e.g., hmm?)" or "Requests to repeat utterance". The model detected mostly open clarification requests, such as "what?", or "huh?".

Secondly, in order to include other kinds of clarification requests such as restricted requests and restricted offers (Dingemanse and Enfield 2015) we also considered questions that are marked by repetition of the previous utterance (e.g., "Went to the house" - "Who went to the house?"). Previous research has found that repair often involves repetition (Jefferson 1972; Fusaroli, Tylén, Garly, et al. 2017; Schegloff,

Jefferson, and Sacks 1977; K. H. Kendrick 2015; Dingemanse and Enfield 2015; M. Purver, Hough, and Howes 2018).

We calculated repetition scores for all child utterances followed by a caregiver utterance. After excluding a set of stopwords and stemming, we calculated:

$$\texttt{rep\_utt} = \texttt{\#words\_overlap}/\texttt{\#words\_utt} \tag{4.1}$$

and

$$\texttt{rep\_response} = \texttt{\#words\_overlap}/\texttt{\#words\_response} \tag{4.2}$$

, where `#words_overlap` is the number of words that are both in the utterance and the response, `#words_utt` the words of the utterance and `#words_response` the words of the response. We only counted unique words. Then, we randomly sampled a set of 200 utterance-response pairs for which the response was a question (marked by a question mark in the transcript) and the repetition ratios were greater than 0. We manually annotated these pairs for whether the response was a clarification request or not.[3] Figure 4.2 shows the relationship between the two repetition ratio measures and whether a response is a clarification request. The annotated data was used to train a logistic regression model that classifies clarification requests based on the repetition ratios.[4] This classifier reached an F-score of 0.82 on the training set. The fitted decision boundary is shown in the graph. We found that for distinguishing clarification requests from other responses, mainly the response repetition ratio (`rep_response`) was important. We annotated another 100 utterances as evaluation set and obtained an F-score of 0.85 (precision: 0.93, recall: 0.78).[5]

4.2.2.3. Acknowledgements

To annotate acknowledgments, we included all responses that start with specific keywords (e.g., “uhhuh”, “mhm”, “okay”, “alright”, “yeah”). We excluded cases in which these keywords are following a question, in that case they are responses and not backchannels. This keyword-based method includes many common backchannel responses, but misses repetition-based acknowledgements (e.g., "It isn't very nice is it?” – “It isn't.”). To identify such acknowledgements, repetition ratios can also be used as a feature (Fernández, Ginzburg, and Lappin 2007).

We manually annotated another 200 utterance-response pairs, but this time including only responses that were not finished with a question mark. All responses that approve the understanding of the child's utterance were marked as acknowledge-

3. We included requests for confirmation (“He went to school.” - “Did he?”), as they also communicate a negative feedback signal to the speaker: The interlocutor is not sure whether they understood the speaker correctly (see also K. H. Kendrick 2015).

4. We evaluated also a non-linear SVM on the data but found that performance did not improve.

5. Manual inspection of misclassified examples showed that cases in which the caregiver asks a follow-up question were sometimes wrongly classified as clarification requests. Other cases with synonyms in restricted offers were not classified as clarification requests (e.g., “I want spoon.” - “You'd like a spoon?”). Additionally, the stemmer did not stem certain colloquial word forms (“wanna”) correctly, which led to incorrect repetition ratios.

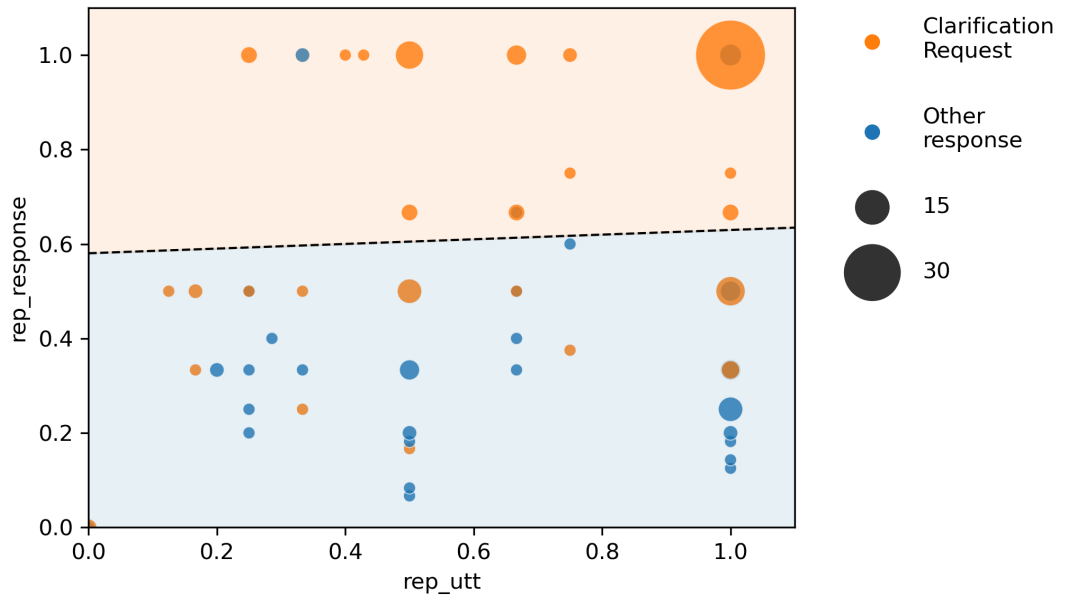

Figure 4.2. – Clarification requests and other responses as a function of repetition ratios. As many points have the same repetition ratios, the number of points is indicated by the size of the dots. The decision boundary is shown as a striped line.

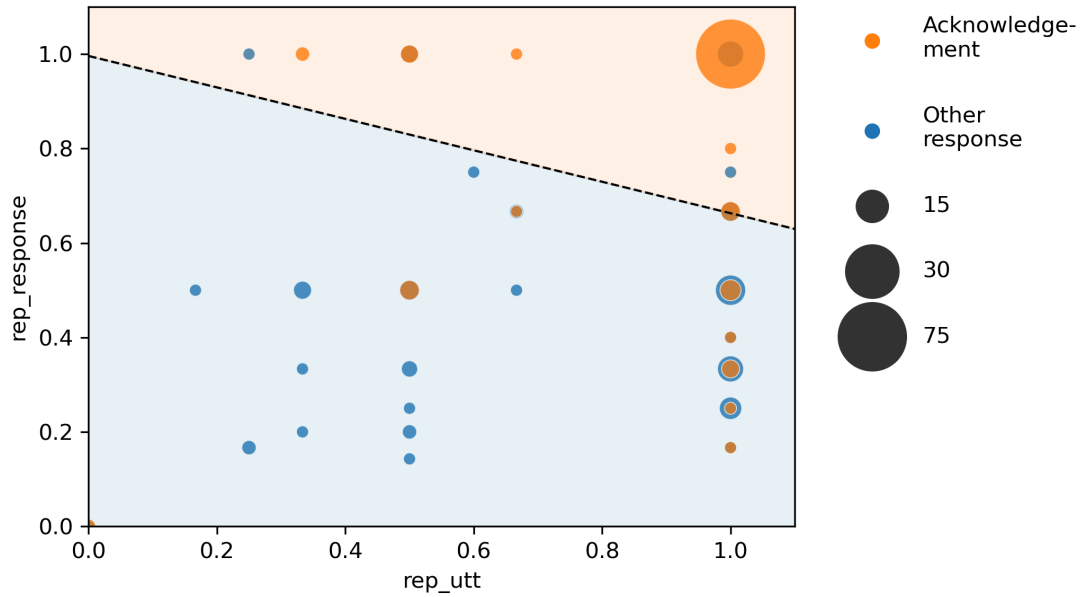

Figure 4.3. – Acknowledgements and other responses as a function of repetition ratios. Number of points is indicated by dot size. The decision boundary is shown as a striped line.

ments. Figure 4.3 shows the relationship between the two repetition ratio measures and whether a response is an acknowledgement. [6] We also fit a logistic regression to classify acknowledgements, which reached an F-score of 0.82 for the training utterances and 0.84 (precision: 0.82, recall: 0.82) on a separate set of 100 evaluation utterances. [7]

Table 4.2 provides an overview on the automatically annotated clarification requests and acknowledgements.

Table 4.2. – Number of clarification requests and acknowledgements that were annotated using speech acts, keywords and repetition features.

Clarification Requests		
Speech Act	Repetition	Total
519	7,540	8,028
Acknowledgements		
Keyword	Repetition	Total
20,398	14,103	32,214

4.3. Analyses

4.3.1. Caregiver's Clarification Requests

Figure 4.4 compares the difference in proportion of clarification requests to grammatical and ungrammatical child utterances. The graph suggests that clarification requests are used more often in response to ungrammatical sentences. We supported this hypothesis using a mixed-effects GLM that predicts whether a clarification request was given as a function of whether the child utterance was grammatical and child age (in months), including a random intercept for the child identifier:

```
resp_is_clar_req ~ utt_is_grammatical * age + (1|child)   (4.3)
```

The estimated fixed effect for the grammaticality of the child utterances was negative (utt_is_grammatical : $\beta = -0.516, SE = 0.055, p < 0.001$), validating our observation.

The other fixed effect estimates indicated that caregivers use a decreasing number of clarification requests with increasing age of the child (age : $\beta = -0.912, SE =$

6. We did not stem the words for calculating the repetition ratios for this case, as this would hide small morphological modification which could in fact be corrections. In this case the utterance is not an acknowledgement, but rather the contrary.

7. Manual inspection showed that many misclassified examples are cases in which the response is a repetition with minimal changes, which however changes the overall semantics and therefore pragmatics of the utterance.

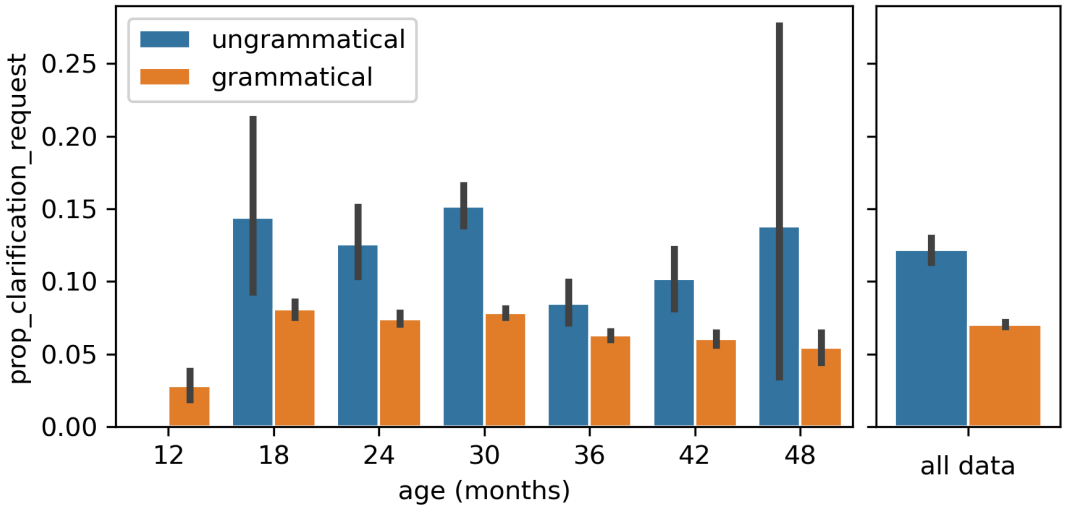

Figure 4.4. – Proportion of caregiver's clarification requests to children's grammatical and ungrammatical utterances. The error bars indicate 95% confidence intervals.

0.117, $p < 0.001$), and a slight decrease of the main effect with increasing child age (utt_is_grammatical*age : $\beta = 0.566, SE = 0.221, p < 0.05$).

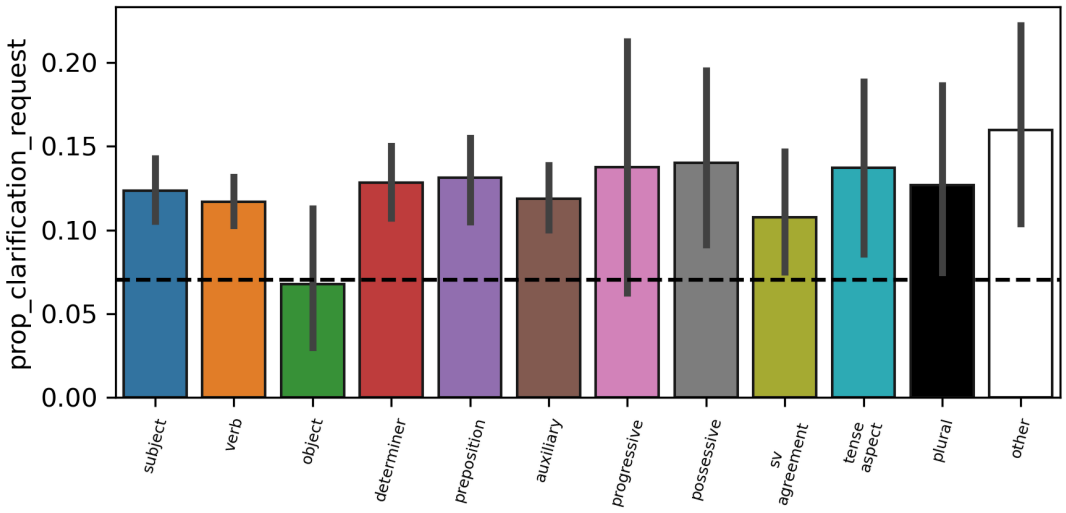

Figure 4.5. – Proportion of caregiver's clarification requests to children's utterances with different error types. The baseline ratio (proportion of clarification requests after grammatical utterances) is indicated as a striped line.

Subsequently, as illustrated in Figure 4.5, we were interested in whether caregivers respond with clarification requests primarily to certain kinds of grammatical errors, or whether they are used for all kinds of errors. We found that the pattern of increased

use of clarification requests is not specific to certain kinds of errors, but holds for almost all error types (with the exception of syntactic `object` errors).

4.3.2. Caregiver's Acknowledgements

Next, we turn to positive feedback which is provided in the form of acknowledgements. Figure 4.6 compares the difference in proportion of acknowledgements to grammatical and ungrammatical utterances. We evaluated the following GLM:

```
resp_is_acknowledgement ~ utt_is_grammatical * age + (1|child)   (4.4)
```

We found that acknowledgements are used more often in response to grammatical utterances (utt_is_grammatical : $\beta = 0.2, SE = 0.033, p < 0.001$), but this pattern is decreasing over the child's age as indicated by a negative interaction term (utt_is_grammatical*age : $\beta = -0.891, SE = 0.167, p < 0.001$).

With increasing age, the overall probability of acknowledgements decreased slightly (age : $\beta = -0.198, SE = 0.086, p < 0.05$).

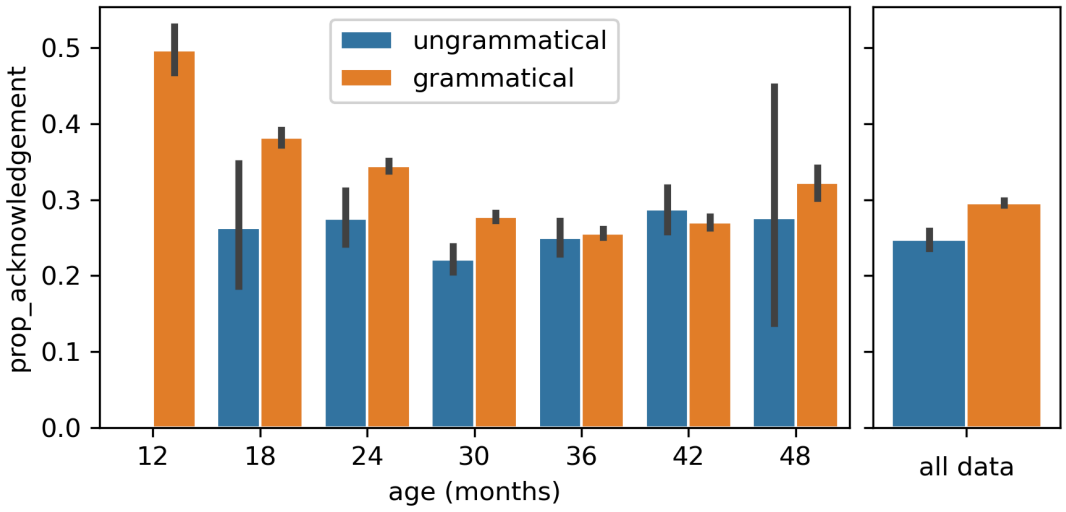

Figure 4.6. – Proportion of caregiver's acknowledgements to children's grammatical and ungrammatical utterances.

4.3.3. Children's Follow-ups

Finally, we explored whether children directly respond to the negative feedback by increasing the grammaticality in their follow-up utterances. Figure 4.7 compares the proportion of grammatical child utterances before (utterance) and after (follow-up) the caregiver's response. In case the caregiver's response is a clarification request (right

side), we observe a slight increase in grammaticality in the children's follow-ups. We fit a mixed-effects model to predict whether an utterance is grammatical depending on whether it is a follow-up (is_follow_up) and whether the response is a clarification request (resp_is_clar_req), including random intercepts for the child identifier, child age, and conversation identifier:

$$\texttt{is\_grammatical} \sim \texttt{resp\_is\_clar\_req} * \texttt{is\_follow\_up} + (1|\texttt{age}) + (1|\texttt{child}) + (1|\texttt{urf\_id}) \quad (4.5)$$

We found a significant positive interaction term resp_is_clar_req*is_follow_up : $\beta = 0.603, SE = 0.072, p < 0.001$, demonstrating that the difference in grammaticality before and after a response is larger in the case of clarification request responses, than it is for other responses.

The other fixed effects indicated that grammaticality of follow-ups is generally lower after clarification requests (resp_is_clar_req : $\beta = -0.231, SE = 0.039, p < 0.001$), and that it is higher in follow-ups (is_follow_up : $\beta = 0.446, SE = 0.036, p < 0.001$)).

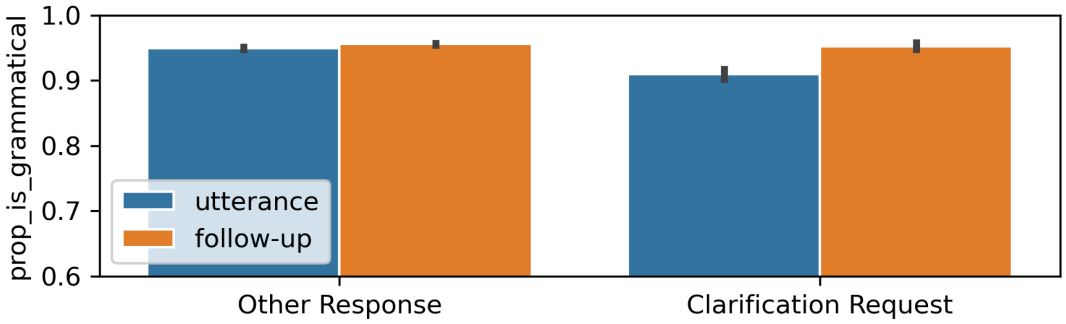

Figure 4.7. – Proportion of grammatical utterances before (utterance) and after (follow-up) a caregiver response, which could be a clarification request (right side) or any other kind of response (left side).

In a follow-up analyses we looked more closely at the kind of clarification request provided by the caregiver. We found that the grammaticality in follow-ups increased only after repetition-based clarification requests, and there was no effect after clarification requests that were annotated by speech acts (which are mostly open requests).

4.4. Discussion

In the present corpus study, we investigated communicative signals provided by caregivers that could support children's acquisition of syntactic and morphological aspects of their native language. We found that caregivers provide both positive and negative communicative feedback in a reliable fashion: They use acknowledgments

more often in response to children's grammatical utterances, and clarification requests more often in response to ungrammatical sentences. When analyzing children's ability to capitalize on these signals, we found that, indeed, the grammaticality of children's utterances increased in direct follow-ups to clarification requests from caregivers. The current studies provide quantitative evidence supporting the presence and effectiveness of communicative feedback signals in grammar learning, enriching the growing research on CF in language acquisition (Eve V. Clark 2020; Nikolaus and Fourtassi 2023).

The role of negative feedback Previous research on the presence of negative feedback led to mixed results (Morgan and Travis 1989; Bohannon and Stanowicz 1988; Demetras, Post, and C. E. Snow 1986; Marcus 1993). In particular, these results have been put into question for several reasons. Some lack inferential statistics, some findings could possibly be explained by averaging artifacts, and some studies conflated multiple error categories (phonological, lexical, and grammatical) (Marcus 1993). Another major limitation of these previous studies was the relatively small sample size (due to the labor-intensive nature of manual annotation), which has led to generalizability issues. Here, we were able to perform a much larger-scale analysis thanks to automatic annotation techniques and aggregation of previously made and publicly available manual annotations. Further, we ensured that our analyses were not subject to previous criticisms. We focused exclusively on grammatical errors, and we employed mixed-effect GLMs which control more efficiently for possible averaging artifacts.

One surprising finding of the current study was that caregivers provided negative feedback to virtually all kinds of children's grammatical errors covered in this study (Figure 4.5). This seems to go against the predictions made by theories of CF for language learning (Nikolaus and Fourtassi 2023). Indeed, one would expect that, if caregivers cared more about communication (and not as much about correction), they would provide more feedback in response to grammatical mistakes that impede more significantly the understanding of children's intended meaning, which is arguably the case for only a subset of the errors and would not include, e.g., errors of past tense inflection (i.e., whether the child says "goes" or "went" does not really impact the transmission of the child's intended meaning, and therefore, should *a priori* receive no negative CF signals). We investigated this observation a bit deeper by focusing on the specific case of `tense_aspect` errors. We found that the negative CF signals in response to such errors are exclusively repetition-based (no open clarification requests). By examining the corresponding URF sequences, we found that they are often recasts ("I waked your mommy right up." - "You woke my mommy right up?"), and sometimes verbatim repetitions of the error without correction: "They breaked it open." - "They breaked it open?"). This suggests, indeed, that caregivers do not glean over grammatical mistakes even when these mistakes do not impede understanding. That said, caregivers' recasts can be understood as providing the child with *both* corrective feedback (a candidate understanding) and negative communicative feedback. Thus, the results of the current work do not only speak to theories of communicative

feedback but also to some aspects of corrective feedback. Further work is needed to determine how communicative and corrective feedback precisely interact in caregiver language use across development.

The role of positive feedback Several studies found that caregivers respond with exact repetitions more often in response to grammatical than to ungrammatical utterances, and could therefore function as positive feedback (Penner 1987; Bohannon and Stanowicz 1988; Demetras, Post, and C. E. Snow 1986). However, as mentioned earlier, these studies have been criticized for several methodological issues. Regarding the specific case of exact repetitions, Marcus (1993) argued that the findings could be merely driven by the fact that almost all adults' utterances are grammatical, and therefore exact repetition of children's grammatical (as opposed to ungrammatical) utterances does not constitute evidence of parental sensitivity to children's grammatical errors. This argument does not apply to our case: The current study does not study *caregiver's sensitivity* to errors, but, rather, the availability of communicative cues that children can leverage for learning even if they do not have a corrective intent on the part of the caregiver.

Our results suggest that positive feedback in the form of acknowledgement is provided predominantly in response to children's grammatical sentences. However, this effect decreased significantly with age (cf. Figure 4.6). It appears that caregivers provide such *explicit* positive feedback mostly in early stages of development, until around the third year of age (this result should be taken with a grain of salt, as the data in the current study is not equally distributed across ages; it is concentrated for children aged 2 to 3 years and is very sparse for the rest). One interpretation of this finding is that, as soon as children start to produce longer and more sophisticated utterances, other (more *implicit*) forms of positive feedback become possible such as the contingency of a caregiver's response to a child utterance (see also Hoff-Ginsberg 1987; Nikolaus and Fourtassi 2023). In other words, children require less explicit encouragement after each correct utterance; as they grow older, they can feel understood merely by having a coherent exchange with their interlocutor. In order to obtain a complete picture of CF signals in language acquisition, such implicit signals should be the focus of future work.

Limitations and future research directions One possible confound of our analysis on grammaticality of children's follow-ups after clarification requests (cf. Figure 4.7) is that children are more likely to produce shorter utterances (e.g., one-word replies) as follow-up to clarification requests and these are more likely to be grammatical. We investigated this possibility by restricting the analysis to utterances that have a minimum length of 2 (or 3) words. In both cases, the effect on the grammaticality of the follow-ups decreased but was still significant for the case of minimum length of 2 words.

More generally, this points to limitations of evaluating children's follow-ups as a means to study the children's sensitivity to feedback (Figure 4.7, as well as Saxton 2000;

Saxton, Houston–Price, and Dawson 2005): This approach is only taking into account immediate and verbalized evidence of children's learning. In many cases, the child might actually understand and take the feedback into account, but not demonstrate it overtly/immediately. Such more long-term effects on learning can be studied using longitudinal data collection coupled with in-lab testing (Bergelson and Aslin 2017).

Another limitation of the present study is that it only considered verbal instantiations of communicative feedback signals that were possible to extract from the transcripts. Future work should include signals that are communicated non-verbally (e.g., head nods, frowns) or using prosodic cues (e.g., rising pitch) using multimodal corpora (e.g., Shi, Gu, and Vigliocco 2022; Bodur, Nikolaus, Prévot, et al. 2023).

Finally, the current analysis was only based on children learning English. Evidence suggests that communicative feedback signals such as clarification requests (Dingemanse, Roberts, Baranova, et al. 2015; Lustigman and Eve V. Clark 2019; Ochs and Schieffelin 1984) and acknowledgements (Liesenfeld and Dingemanse 2022; Cutrone 2005; Maynard 1990) are universally used in human conversations, and can therefore be leveraged by children from different languages and cultures. Future work is required to investigate this hypothesis.

Part III.

Computational Models of CF in Language Acquisition

Summary

5. Evaluating the Acquisition of Semantic Knowledge in Multimodal NNs

This chapter is based on the article "Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in Artificial Neural Networks" (Nikolaus and Fourtassi 2021a), published in the *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

The review of studies in the first part of this thesis (Section 1.5) showed that studies about Communicative Feedback rely almost exclusively on experimental or corpus-analysis methods. This points to a lack of studies leveraging computational modeling, which have historically been an essential research approach for the study of language development (Roy and Pentland 2002; Fazly, Alishahi, and Stevenson 2010; Abend, Kwiatkowski, N. J. Smith, et al. 2017; Kachergis, V. A. Marchman, and Michael C. Frank 2021; Khorrami and Räsänen 2021; Yu and Ballard 2007; Michael C. Frank, N. D. Goodman, and J. B. Tenenbaum 2009). Computational models allow us to study aspects of learning that are difficult to address with experimental and/or corpus studies alone. More specifically, they help us to precisely instantiate the learning mechanism of interest, control its effect by studying it separately from other mechanisms, but also investigate how it interacts with other mechanisms. Further, more recent deep-learning based models allow us to test whether the mechanism of interest scales up to learning from more naturalistic input and simulate its developmental properties over long time scales.

Here, we investigated to what extend cross-situational learning, which has been successfully tested in laboratory experiments with children, scales up to more natural language and visual scenes using a large dataset of crowd-sourced images with corresponding descriptions. We implemented Artificial Neural Networks to model the statistical learners and evaluated their learning using a series of tasks which showed that the model acquires rich semantic knowledge on multiple levels, regarding both word-level and sentence-level semantics. Further, the results suggest that the model is mirroring some common patterns of language learning in early childhood.

This study forms the basis for the study in the following Chapter (6) on the interaction of input-based and feedback-based learning mechanisms. In the first place, it serves as a baseline comparison of a learner that does not leverage feedback-based learning mechanisms. Secondly, the evaluation paradigms developed within the scope of this work were directly applied to compare different learning setups: combining

both input-based and feedback-based learning mechanisms as well as alternations of the two.

The evaluation method is directly inspired by 2-alternative forced choice paradigms as commonly used in laboratory studies of language acquisition (Bergelson and Swingley 2012; Noble, Rowland, and Pine 2011; Gertner and Fisher 2012). Keeping such as close link between model and human evaluation will facilitate more direct comparisons of the behavior of models and humans in future work. To this end, researchers should perform experiments with increasingly realistic input data: The input should be provided in the form of raw video and audio instead of images and text (e.g., Nikolaus, Alishahi, and Chrupała 2022) and should optimally be from a first-person point of view (e.g., Sullivan, Mei, Perfors, et al. 2022).

5.1. Introduction

In order to acquire their native language, children learn both how to associate individual words with their meanings (e.g., the word "ball" refers to the object ball and the word "kick" refers to that act of kicking) and how to map the relationship between words in a sentence onto specific event configurations in the world, e.g., that the sequence of words "Jenny kicks the ball" maps on to the event where the referent of the first noun (i.e., Jenny) is performing the act of kicking on the second (i.e., the ball). This is a difficult task because it requires that children learn these associations and rules in a largely unsupervised fashion from an input that can be highly ambiguous (Quine 1960). It is still unclear how children overcome this challenge.

Previous experimental studies on child language acquisition have focused on *evaluating* children's learning using controlled tasks that typically take the form of a two-alternative forced-choice paradigm. For example, in order to test the learning of an individual word meaning, we can utter this word to the child (e.g., "ball") and present her with two pictures representing correct (i.e., a ball) and incorrect referents (e.g. a cup), and we test if the child reliably prefers the correct one (Bergelson and Swingley 2012). Similarly, in order to evaluate children's understanding of sentence-level semantics such as a the agent-patient relationship, we can utter a sentence such as "Jenny is tickling Mike" and present the child with two pictures where either Jenny or Mike are doing the tickling, and we test if the child reliably prefers the correct picture (e.g. Noble, Rowland, and Pine 2011; Gertner and Fisher 2012).

While we have been able to evaluate children's knowledge using such controlled tests, research has been less compelling regarding the *mechanism of learning* from the natural, ambiguous input. One promising proposal is that of cross-situational learning (hereafter, XSL). This proposal suggests that, even if one naming situation is highly ambiguous, being exposed to many situations allows the learner to narrow down, over time, the set of possible word-world associations (e.g. Pinker 1989).

While in-lab work has shown that XSL is cognitively plausible using toy situations (L. Smith and Yu 2008), effort is still ongoing to test if this mechanism scales up to more natural learning contexts using machine learning tools (e.g. Chrupała, À. Kádár, and Alishahi 2015; Vong and Lake 2020). This previous work, however, has focused mainly on testing the learning of individual words' meanings, while here we are interested in testing and comparing both word-level and sentence-level semantics.

5.1.1. The Current Study

The current study uses tools from Natural Language Processing (NLP) and computer vision as research methods to advance our understanding of how unsupervised XSL could give rise to semantic knowledge. We aim at going beyond the limitations of in-lab XSL experiments with children (which have relied on too simplified learning input) while at the same time integrating the strength and precision of in-lab learning evaluation methods.

More precisely, we first design a model that learns in an XSL fashion from images

and text based on a large-scale dataset of clipart images representing some real-life activities with corresponding – crowdsourced – descriptions. Second, we evaluate the model's learning on a subset of the data that we used to carefully design a series of controlled tasks inspired from methods used in laboratory testing with children. Crucially, we test the extent to which the model acquires various aspects of semantics both at the word level (e.g., the meanings of nouns, adjectives, and verbs) and at the sentence level (e.g. the semantic roles of the nouns).

Further, in order for an XSL-based model to provide a plausible language learning mechanism in early childhood, it should not only be able to succeed in the evaluation tasks, but also mirror children's learning trajectory (e.g., a bias to learn nouns before predicates). Thus, we record and analyze the model's learning trajectory by evaluating the learned semantics at multiple timesteps during the training phase.

5.1.2. Related Work and Novelty

While supervised learning from images and text has received much attention in the NLP and computer vision communities, for example in the form of classification problems (e.g. Yatskar, L. Zettlemoyer, and Farhadi 2016) or question-answering (e.g. Antol, Agrawal, J. Lu, et al. 2015; Hudson and Manning 2019), here we focus on *cross-situational learning* of visually grounded semantics, which corresponds more to our understanding of how children learn language

There is a large body of work on cross-situational word learning (N. Goodman, J. Tenenbaum, and Black 2007; Yu and Ballard 2007; Fazly, Alishahi, and Stevenson 2010), some of them with more plausible, naturalistic input in the form of images as we consider in our work (Á. Kádár, Alishahi, and Chrupa\la 2015; Lazaridou, Chrupa\la, Fernández, et al. 2016; Vong and Lake 2020). However, these previous studies only evaluate the semantics of single words in isolation (and sometimes only nouns). In contrast, our work aims at a more comprehensive approach, testing and comparing the acquisition of both word-level meanings (including adjectives and verbs) and sentence-level semantics.

There has been some effort to test sentence-level semantics in a XLS settings. For example, Chrupała, À. Kádár, and Alishahi (2015) also introduces a model that learns from a large-scale dataset of naturalistic images with corresponding texts. To evaluate sentence-level semantics, the model's performance was tested in a cross-modal retrieval task, as commonly used to evaluate image-sentence ranking models (Hodosh, Young, and Hockenmaier 2013). They show that sentence to image retrieval accuracy decreases when using scrambled sentences, indicating that the model is sensitive to word order. In a subsequent study, Á. Kádár, Chrupała, and Alishahi (2017) introduces *omission scores* to evaluate the models' selectivity to certain syntactic functions and lexical categories. Another evaluation method for sentence-level semantics is to compare learned sentence similarities to human similarity judgments (e.g. Merkx and S. L. Frank 2019).

Nevertheless, these previous studies only explored broad relationships between sentences and pictures, they did not test the models' sensitivity to finer-grained

phenomena such as dependencies between predicates (e.g., adjectives and verbs) and arguments (e.g., nouns) or semantic/ roles in detail.

5.2. Methods

5.2.1. Data

We used the Abstract Scenes dataset 1.1 (Zitnick and Parikh 2013; Zitnick, Parikh, and Vanderwende 2013), which contains 10K crowd-sourced images each with 6 corresponding short descriptive captions in English. Annotators were asked to "create an illustration for a children's story book by creating a realistic scene" given a set of clip art objects (Zitnick and Parikh 2013). The images contain one or two children engaged in different actions involving interactions with a set of objects and animals. Further, the children can have various emotional states depicted through a variety of facial expressions. The corresponding sentences were collected by asking annotators to write "simple sentences describing different parts of the scene" [1] (Zitnick, Parikh, and Vanderwende 2013).

While some studies have used larger datasets with more naturalistic images (e.g. Lin, Maire, Belongie, et al. 2014; Plummer, L. Wang, Cervantes, et al. 2015), here we used the Abstract Scenes dataset since it contains many similar scenes and sentences, allowing us to create balanced test sets (as described in the following section). In other words, the choice of the dataset was a trade-off between the naturalness of the images on the one hand and their partial systematicity, on the other hand, which we needed to design minimally different pairs of images to evaluate the model.

For the following experiments, we split the images and their corresponding descriptions into training (80%), validation (10%) and test set (10%).

5.2.2. Model

We use a modeling framework that instantiates XSL from images and texts in the dataset. To learn the alignment of visual and language representations, we employ an approach commonly used for the task of image-sentence ranking (Hodosh, Young, and Hockenmaier 2013) and other multimodal XSL experiments (Chrupała, Gelderloos, and Alishahi 2017; Vong, Orhan, and Lake 2021).

The objective is to learn a joint multimodal embedding for the sentences and images, and to rank the images and sentences based on similarity in this space. State-of-the-art models extract image features from Convoluatinal Neural Networks (CNNs) and use LSTMs to generate sentence representations, both of which are projected into a joint embedding space using a linear transformation (Karpathy and Fei-Fei 2015; Faghri 2018).

As commonly applied in other multimodal XSL work (Chrupała, À. Kádár, and Alishahi 2015; Khorrami and Räsänen 2021), we assume that the visual system of the

1. The annotators were asked to refer to the children by the names "Jenny" and "Mike".

learner has already been developed to some degree and thus use a CNN pre-trained on ImageNet (Russakovsky, Deng, Su, et al. 2015) (but discard the final classification layer) to encode the images. Specifically, we use a ResNet 50 [2] (K. He, Zhang, Ren, et al. 2016) to encode the images and train a linear embedding layer that maps the output of the pre-final layer of the CNN into the joint embedding space.

The words of a sentence are passed through a linear word embedding layer and then encoded using a one-layer LSTM (Hochreiter and Schmidhuber 1997). Using a linear embedding layer, the hidden activations of the last timestep are then transformed into the joint embedding space.

The model is trained using a max-margin loss [3] which encourages aligned image-sentence pairs to have a higher similarity score than misaligned pairs, by a margin α:

$$\mathscr{L}(\theta) = \sum_a [\sum_b max(0, \gamma(i_a, s_b) - \gamma(i_a, s_a) + \alpha) + \sum_b max(0, \gamma(i_b, s_a) - \gamma(i_a, s_a) + \alpha)] \quad (5.1)$$

$\gamma(i_a, s_b)$ indicates the cosine similarity between an image i and a sentence s, (i_a, s_a) denotes a corresponding image-sentence pair. The loss is calculated for each mini-batch, negative examples are all examples in a mini-batch for which the sentence does not correspond to the image.

We train the model on the training set until the loss converges on the validation set. Details about hyperparameters can be found in the appendix.

5.2.3. Evaluation Method

In order to evaluate the model's acquisition of visually-grounded semantics, we used a two-alternative forced choice design, similar to what is typically done to evaluate children's knowledge in laboratory experiments (Bergelson and Swingley 2012; Noble, Rowland, and Pine 2011; Gertner and Fisher 2012). Each test trial consists of an image, a target sentence and a distractor sentence: (i, s_t, s_d). We measure the model's accuracy at choosing the correct sentence given the image.

Crucially, we design the test tasks in a way that allows us to control for linguistic biases. Consider the example trial on the left in Figure 5.1. The model could posit that, say, Jenny (and not Mike) is the agent of an action even without considering the image, and only because Jenny may happen to be the agent in most sentences in the training

2. We also tried the more recent ResNet 152, but found results to be inferior. Also, we did not attempt to fine-tune the parameters of the CNN for the task, which could improve performance further.

3. In preliminary experiments we also applied a max-margin loss with emphasis on hard negatives (Faghri 2018), but observed a performance decrease. This could be due to the fact that our dataset contains many repeating sentences and semantically equivalent scenes, and consequently we could find "hard negatives" that should actually be positive learning examples (because they are semantically equivalent) in many situations.

Figure 5.1. – Counter-balanced evaluation of visually-grounded learning of semantics: Each test trial has a corresponding counter-example, where target and distractor sentence are flipped.

data. To avoid such linguistic biases, we paired each test trial with a counter-balanced trial where the target and distractor sentence were flipped (cf. Figure 5.1, right side), in such a way that a language model without any visual grounding can only perform at chance level (50%).

More precisely, we made the tasks as follows. First we searched in the heldout test set for image-sentence pairs $[(i_x, s_x), (i_y, s_y)]$ with *minimal differences* in the sentences given the phenomenon under study. For example, to study the acquisition of noun meanings, we look for pairs of sentences where the difference is only one noun such as s_x = "jenny is wearing a crown" and s_y = "mike is wearing a crown" (the corresponding images i_x and i_y depict the corresponding scenes, as shown in Figure 5.1). Second, based on such a minimal pair, we construct two counter-balanced triads: (i_x, s_x, s_y) and (i_y, s_y, s_x). The target sentence in one triad is the distractor in the other triad (and vice-versa). Using such a pair of counter-balanced triads, we test whether a model can both successfully choose the sentence mentioning “Jenny” when presented with the picture of Jenny *and* choose the sentence mentioning “Mike” when presented with the picture of Mike.

In the following we describe in more detail the phenomena of semantics we investigated using this testing setup. We provide an example for each category of task in Figure 5.2.

Figure 5.2. – Examples for the evaluation of word and sentence-level semantics. Each test trial consists of an image, a target and a distractor sentence.

5.3. Tasks

5.3.1. Word-level Semantics

To study the acquisition of word meanings, we collect minimal pairs for the most commonly occurring nouns, adjectives and verbs. An example can be seen in Figure 5.1. Across all word-level categories, we make sure that there is only one referent present in the scene (this could be a child, an animal, or inanimate object, depending on the noun category under study). This ensures that we only evaluate word learning, and not more complex sentence-level semantics. [4]

Nouns We group the nouns into *persons*, *animals* and *objects*. Regarding persons, we consider the two children talked about in the dataset, i.e., *Jenny* and *Mike*. Regarding animals, we consider all 6 animals present in the dataset. [5] Regarding objects, we consider the 12 most frequently occurring words that are describing physical objects. [6]

4. For example, if *Mike* (without a crown) was present in the picture to the left in Figure 5.1, the model would not only need to understand the difference between *Jenny* and *Mike*, but also understand what it means to *wear a crown* in order to correctly judge which sentence is the correct one, that is, which of Mike and Jenny is the one with the crown.

5. ("dog", "cat", "snake", "bear", "duck", "owl")

6. ("ball", "hat", "tree", "table", "sandbox", "slide", "sunglasses", "pie", "pizza", "hamburger", "balloons", "frisbee")

Verbs The category of verbs is a bit tricky to evaluate because verbs are usually followed with an object that is tightly connected to them (e.g. *kicking* is usually connected to a ball whereas *eating* is connected to some food), resulting in a very limited availability of minimally different sentences with respect to verbs in the dataset. To be able to create a reasonable number of test trials, we trimmed the sentences [7] after the target verb and only consider verbs that can be used intransitively, e.g., "Mike is eating an apple" becomes "Mike is eating".

Further, we ensure, that the trials do not contain pairs of target and distractor sentences where the corresponding actions can be performed at the same time. For example, we do not include trials where the target sentence involves *sitting* and the distractor sentence *eating*, because the corresponding picture could be ambiguous: If the child in the picture is *sitting* and *eating* at the same, both the target and distractor sentences could be semantically correct. The resulting set of possible verb pairings is: ("sitting", "standing"), ("sitting", "running"), ("eating", "playing"), ("eating", "kicking"), ("throwing", "eating"), ("throwing", "kicking"), ("sitting", "kicking"), ("jumping", "sitting").

Adjectives The most common adjectives in the dataset are related to mood (e.g., happy and sad) and are displayed in the pictures using varied facial expressions (happy face vs sad face). Due to the lack of other kinds of adjectives [8], we only focused on mood-related adjectives. In addition, as there is no clear one-to-one mapping between each adjective and a facial expression, we only test the broad opposition between rather positive mood (smiling or laughing face) and rather negative mood (all other facial expressions). The resulting set of pairings was: ("happy", "sad"), ("happy", "angry"), ("happy", "upset"), ("happy", "scared"), ("happy", "mad"), ("happy", "afraid"), ("happy", "surprised").

Similar to what we did in the case of verbs, we trimmed the sentences after the target adjective in order to obtain more minimal pairs in our test set.

5.3.2. Sentence-level Semantics

In addition to evaluating the learning of word-level semantics, here we evaluate some (rudimentary) aspects of sentence-level semantics, that is, semantic phenomena where the model needs to leverage *relationships* between words in the sentence to be able to arrive at the correct solution. We focused on the following three cases for which a reasonable number of minimal pairs could be found.

Adjective - Noun Dependency In this task, we test if the model is capable of recognizing not only a given adjective (e.g., sad), but also the person experiencing

7. The trimming was only done for the test trails and not in the training set.

8. In the dataset, most of the properties for objects are fixed (e.g. colors and shapes) and are thus very rarely referred to in the descriptions. Consequently, we did not find minimal pairs for adjectives describing simple properties like color.

	Evaluation task	Accuracy	p (best)	p (worst)	Size
Word-level Semantics	Nouns: Persons	0.78 ± 0.05	< 0.001	< 0.01	50
	Nouns: Animals	0.93 ± 0.02	< 0.001	< 0.001	360
	Nouns: Objects	0.86 ± 0.01	< 0.001	< 0.001	372
	Verbs	0.83 ± 0.05	< 0.001	< 0.001	77
	Adjectives	0.64 ± 0.06	< 0.01	0.25	56
Sentence-level Semantics	Adjective-noun dependencies	0.57 ± 0.01	< 0.05	< 0.05	192
	Verb-noun dependencies	0.72 ± 0.04	< 0.001	< 0.001	400
	Semantic roles	0.75 ± 0.06	< 0.001	< 0.05	50

Table 5.1. – Accuracy, p-values (for the best and for the worst performing model) and evaluation set size (in number of trials) for all semantic evaluation tasks. The high variance in terms of number of trials is caused by the limited availability of appropriate examples in the dataset for some tasks (cf. Footnote 10).

this emotion (i.e. Jenny or Mike). The procedure used here is similar to the one we used to test individual adjectives, except that here the picture contains not only the person experiencing the target emotion but also the other person who is experiencing a different emotion (cf. examples on bottom left in Figure 5.2).

Take the following example: “mike is happy” and its minimally different distractor sentence “mike is sad” associated with a picture where Mike is happy and Jenny is sad (see Figure 5.2). In order to choose the target sentence over the distractor, the model needs to associate happiness with Mike but not with Jenny. In fact, since both persons appear in the picture and the word Mike appears in both sentences, the model cannot succeed by relying only on the individual name “mike” (in which case performance would be at chance). Similarly, it cannot succeed only by relying on the contrast “happy” vs. “sad” since Mike is happy but Jenny is sad (in which case performance would also be at chance).

Moreover, it cannot succeed even if it combines information in the words “mike” and “happiness” without taking into account their dependency in the sentence (say, if it only relied on a bag-of-words representation) because both the sentence and distractor would be technically correct in that case. More precisely, the bag of words of the target sentence {“mike”, “happy”} and of the distractor {“mike”, “sad”} both describe the scene accurately since the latter contains Mike, Happy, and Sad. The model can only succeed if it correctly learns that happiness is associated with Mike in the picture, suggesting that the model learns “happy” as modifier/predicate for “mike” in the sentence.

To construct test trials for this case, we used the same adjectives as for the word-level adjective learning, but we searched for minimal pair sentences with a second child in the scene with the opposite mood compared the target child.

Verb - Noun Dependencies Similar to adjective-noun dependencies, we aim to evaluate learning of verbs as predicate for the nouns they occur with in the sentence. We use the same verbs as in the word-learning setup as well as trim the sentences after the verb. We look for images with a target and distractor child engaged in different actions and construct our test dataset based on these scenes (see example in Figure 5.2, bottom right).

Semantic Roles In this evaluation, we test the model's learning of semantic roles in an action that involves two participants. We test the model's learning of the mapping of nouns to their semantic roles (e.g., agent vs. patient/recipient).

We look for scenes where both children are present and engaged in an action. In this action, one of the children is the agent and the other one is the patient/recipient. For example, in the sentence "jenny is waving to mike" the agent is Jenny and the recipient is Mike (see Figure 5.2, top right). The distractor sentence is constructed by flipping the subject and object in the sentence, i.e., "mike is waving to jenny". To succeed in the task, the model should be able to recognize that Jenny, not Mike, is the one doing the waving. This task is a more challenging version of the verb-noun dependency we described above because, here, Jenny and Mike are not only both present in the picture, they are also both mentioned in the sentences. To succeed, the model has to differentiate between agent and recipient in the sentence. Here again, a null hypothesis that assumes a bag-of-word representation of the sentence would not succeed: We need to take into account how each noun *relates* to the verb.

As with all other evaluation tasks, for each test trial we have a corresponding counterbalanced trial where the semantic roles are flipped.

5.4. Results

To evaluate the learned semantic knowledge, we measure, for each task, the model's accuracy at rating the similarity of the image and the target sentence $\gamma(i, s_t)$ higher than the similarity to the distractor sentence $\gamma(i, s_d)$. We report both final accuracy scores after the model has converged as well as intermediate scores before convergence, which we take as a proxy for the learning trajectory.

To ensure reproducibility, we make the semantic evaluation sets as well as the source code for all experiments publicly available. [9]

5.4.1. Acquisition Scores

We ran the model 5 times with different random initializations and evaluate each converged model using the proposed tasks. Mean and standard deviation of the resulting accuracy scores can be found in Table 5.1. As some of the evaluation sets are

9. https://github.com/mitjanikolaus/cross-situational-learning-abstract-scenes

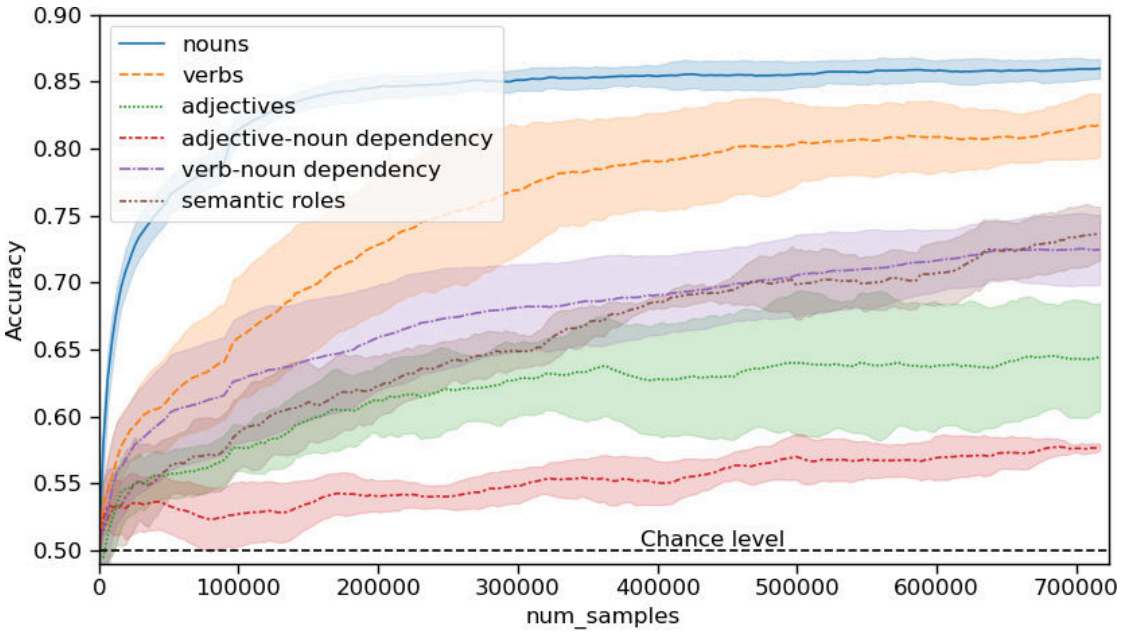

Figure 5.3. – Learning trajectory of the models (mean over 5 runs, shaded areas show standard deviation). Accuracies for all noun categories were averaged. We calculated a rolling average over 30 data points to smooth the curve. The training set contains ~50K examples, which means that the graph displays development over 15 epochs.

rather small [10], we also performed binomial tests to evaluate whether the accuracy in the binary test is significantly above chance level (50%). We report the p-values' significance levels for the best and for the worst performing model [11] for each evaluation task.

The results show that the model has learned the semantics for most nouns very well. The score for verbs is also relatively high. As for adjectives, performance is only slightly above chance level and not always statistically significant, depending on the random initialization (e.g. the worst model is not significantly better than chance).

Regarding sentence-level semantics, the results suggest that the model has learned verb-noun dependencies and semantic roles relatively well. In contrast, Adjective-noun dependencies are not learned very well, which is not surprising given the poor adjective word-learning performance.

10. Some evaluation sets are smaller than others due to the fact that all image-sentence pairs are taken directly from the test set and no new artificial images or sentences were created. This was done to ensure that the tests are performed using data that comes from the same distribution as the training set, i.e. data that the model has been exposed to.

11. Each model corresponds to a different random initialization.

5.4.2. Acquisition Trajectories

In addition to the final evaluation scores, we are also interested in the *learning trajectory* of the model. We calculated the accuracy scores of the model every 100 batches. Figure 5.3 shows how the performance on the semantic evaluation tasks develops during the training of the model.

The model converged after having seen around 700K training examples (around 14 epochs). The trajectories show that the model first learns to discriminate nouns and only slightly later the verbs and then more complex sentence-level semantics.

5.5. Discussion

This work dealt with the question of how children learn the word-world mapping in their native language. As a possible learning mechanism, we investigated XSL, that has received much attention in the literature. While laboratory studies on XSL have typically used very simplified learning situations to test if children are cognitively equipped to learn a toy language in an XSL fashion. The question remains as whether such a mechanism scales up to the learning of real languages where the learning situations can be highly ambiguous.

The novelty of our work is that we were interested not only in the scalability of XSL to learn from more naturalistic input, but also its scalability to the learning of various aspects of semantic knowledge. These include both the meanings of individual words (belonging to various categories such as nouns, adjectives, and verbs) and the meanings of higher level semantics such as the ability to map how words relate to each other in the sentence (e.g., subject vs. object) to the semantic roles of their respective referent in the world (e.g., agent vs. patient/recipient). We were able to perform these evaluations using a simple method inspired from the field of experimental child development and which has usually been used to test the same learning phenomena in children, i.e., the two-alternative forced choice task.

Using this evaluation method, we found that an XSL-based model trained on a large set of pictures and their descriptions was able to learn word-level meanings for nouns and verbs relatively well, but struggles with adjectives. Further, the model seems to learn some sentence-level semantics, especially verb-noun dependencies and semantic roles. Finally, concerning the learning trajectory, the model initially learns the semantics of nouns and only later the semantics of verbs and more complex sentence-level semantics.

Concerning word-level semantics, the fact that the model learns nouns better than (and before) the predicates (adjectives and verbs) resonates with findings in child development about the "noun bias" (Gentner 1982; Bates, V. Marchman, Thal, et al. 1994; Michael C. Frank, Braginsky, Yurovsky, et al. 2021). The model also learns verbs better than adjectives. However, we suspect this finding is caused by the limited availability of adjectives in the dataset. [12] In fact, the verb-related actions (e.g. "sitting"

12. The data contained mostly mood-related adjectives.

vs. "standing") were arguably more salient and easier to detect visually than adjective-related words ("happy" vs. "sad") which require a fine-grained detection of the facial expressions.

Concerning sentence-level semantics, the model performed surprisingly well on verb-noun dependency task where the model assigned a semantic role to one participant and on the similar but (arguably) more challenging task of assigning semantic roles to two participants. Further, the fact that the model shows a rather late onset of understanding of semantic roles, only after a set of nouns and verbs have been acquired (cf. Figure 5.3) mirrors children's developmental timeline. Indeed, children become able to assign semantic roles to nouns in a sentence correctly when they are around 2 years and 3 months old (Noble, Rowland, and Pine 2011), at an age when they have already acquired a substantial vocabulary including many lexical categories such as nouns and verbs (Michael C. Frank, Braginsky, Yurovsky, et al. 2021)

In this work, we used artificial neural networks to study how properties the input can (ideally) inform the learning of semantics. Our modeling did not purport to account for the details of the cognitive processes that operate in children's minds nor did it take into account limitations in children's information-processing abilities. Thus, this work is best situated at the computational level of analysis (Marr 1982), which is only a first step towards a deeper understanding of the precise algorithmic implementation. That said, we can speculate about the internal mechanisms used by the model to succeed in the tasks and about their potential insights into children's own learning. For example, it is very likely that the model leverages simple heuristics to recognize the agent in a sentence, e.g., it may have learned to associate the first appearing noun in the sentence to the agent of the action. Research on child language suggest that children also use such heuristics (e.g. Gertner and Fisher 2012). This suggests that the model, like children, might use partial representations of sentence structure (i.e., rudimentary syntax) to guide semantic interpretation.

Exploiting structural properties of the input (e.g., order of words in a sentence) may be insightful when it mirrors genuine learning heuristics in children. However, a neural network model may also capitalize on idiosyncratic biases in the dataset (that do not reflect the natural distribution in the world) to achieve misleadingly high performance. [13] For example, a misleading bias in the linguistic input is if a certain noun (e.g., Jenny) occurs more frequently in the dataset as agent, leading the model to, say, systematically map "Jenny" to agent. Similarly, an example of a misleading bias in visual data is if the agent is always depicted on the left or right side of the image, leading the model to capitalize on this artificial shortcut.

In the current work, we controlled for linguistic biases by counter-balancing all testing trials. As for the visual bias, we ruled out some artificial biases such as the agent spatial order in the images. Indeed, investigation of our semantic roles test set shows that the agent occurs roughly equally on the right (52%) and left sides, which means that a model exploiting such a bias could only perform around chance level.

13. For example, Goyal, Khot, Summers-Stay, et al. (2017) finds that grounded language models trained on a visual question answering task are exploiting linguistic biases of the training set.

There could be other biases we are not aware of and which require performing further controls. That said, this is an open question for all research using neural networks as models of human learning. More generally, our understanding of language acquisition would greatly benefit from further research on the interpretation of neural network learning, revealing the content of these black box models. This would allow us to tease apart genuine insights about realistic heuristics that could be used by children and artificial shortcuts that only reflect biases in the learning datasets.

In future work, we plan to study visual datasets with even more naturalistic scenes such as COCO (Lin, Maire, Belongie, et al. 2014). In this regard, maybe closer to our work is the study by Shekhar, Pezzelle, Herbelot, et al. (2017) and Shekhar, Pezzelle, Klimovich, et al. (2017) who used COCO to create a set of distractor captions to analyze whether vision and language models are sensitive to (maximally difficult) single-word replacements. Our goal is to go beyond these analysis to test specific semantic phenomena as we did here with the Abstract Scenes dataset. Another step towards more naturalistic input is the use speech input instead of text (Chrupała, Gelderloos, and Alishahi 2017; Khorrami and Räsänen 2021).

Finally, this work focused on testing how XSL scales up to natural language learning across many semantic tasks. Nevertheless, children's language learning involves more than the mere tracking of co-occurrence statistics: They are also social beings, they actively interact with more knowledgeable people around them and are able to learn from such interactions (Tomasello 2010). Future modeling work should seek to integrate both statistical and social learning skills for a better understanding of early language learning.

6. Modeling Interactions between Statistical Learning and CF

This chapter is based on the article "Modeling the Interaction Between Perception-Based and Production-Based Learning in Children's Early Acquisition of Semantic Knowledge" (Nikolaus and Fourtassi 2021b), published in the *Proceedings of the 25th Conference on Computational Natural Language Learning*.

Existing modeling effort of first language acquisition has mostly focused on mechanisms that leverage statistical regularities in the input, such as cross-situational learning (e.g., Roy and Pentland 2002; Fazly, Alishahi, and Stevenson 2010; Abend, Kwiatkowski, N. J. Smith, et al. 2017; Kachergis, V. A. Marchman, and Michael C. Frank 2021; Khorrami and Räsänen 2021; Nikolaus and Fourtassi 2021a), sometimes integrating also non-verbal social cues (Yu and Ballard 2007), and the ability for pragmatic inference in ambiguous learning situations (Michael C. Frank, N. D. Goodman, and J. B. Tenenbaum 2009).

In comparison, little has been done to model language acquisition in a context where an artificial child agent learns in an interactive context and from Communicative Feedback. The major difficulty impeding progress in this direction is the requirement for a dynamic model of the interlocutor's Communicative Feedback, which in turn requires a model of the interlocutor that is able to "understand" what the child agent is saying and responds in an appropriate manner. The design and implementation of such models remains an open challenge, especially in the context of spontaneous conversations involving natural language spanning multiple turns.

In this chapter, we propose a model integrating both input-based and feedback-based learning using artificial neural networks which we train on a large corpus of crowd-sourced images with corresponding descriptions. The Communicative Feedback-based learning mechanism was implemented using reinforcement learning. We train models in varying learning setups and evaluate them using the paradigms for word-level and sentence-level semantics designed in the previous chapter (5). We found that feedback-based learning improves performance above and beyond input-based learning across a wide range of semantic tasks including both word- and sentence-level semantics. In addition, we documented a synergy between these two mechanisms, where their alternation allows the model to converge on more balanced semantic knowledge.

While this proof-of-concept implementation showed promising results, the proposed model should be applied to more realistic input and feedback signals in future work. One main limitation remains the implementation of the reward value, which

was approximated by comparing generated sentences to a set of ground truth captions. More realistic scenarios would require a dynamic model of a reward function. Progress on this end would eventually also allow for a more direct comparison of the model's predictions with human behavioral data.

Another important avenue for future research could be to apply such modeling techniques for the study of the effects of Communicative Feedback on intelligibility and grammar learning (as analyzed in Chapters 3 and 4). The current analyses investigated learning effects only by analyzing the direct follow-up turns of the children. However, it is very likely that many effects on learning only become apparent in later interactions, as a result of repeated short interactions. Training and evaluating computational models that leverage such noisy feedback signals could highlight the potential of such more long-term effects on learning.

6.1. Introduction

An important aspect of language acquisition is learning how to map linguistic forms to meanings. This involves both mapping individual word forms (e.g., "dog") to concepts of the world (e.g. the category DOG) and mapping the relationship between words in a sentence (e.g., "the dog chases the ball") to a given event configuration in the world (i.e., that the dog is the agent performing the act of chasing on the ball, the semantic patient). Children manage to learn this mapping in their native language at an impressive speed (Fisher and L. R. Gleitman 2002; Roberta Michnick Golinkoff, Ma, Song, et al. 2013; Michael C. Frank, Braginsky, Yurovsky, et al. 2021) and despite the high ambiguity of this task in the natural context where language learning occurs (Quine 1960).

Input-based learning

Much modeling effort has focused on learning from the multimodal input that children perceive around them. These models are based on Cross-Situational Learning (hereafter XSL): While a single word-world mapping situation is ambiguous, being exposed to many situations allows the learner to narrow down, over time, the set of possible associations. This kind of learning has been demonstrated using toy situations in controlled laboratory testing with children (L. Smith and Yu 2008). It has also been shown to scale up to more realistic learning contexts using a combination of NLP and computer vision tools (Chrupała, À. Kádár, and Alishahi 2015; Vong and Lake 2020; Vong, Orhan, and Lake 2021; Vong and Lake 2022; Nikolaus, Alishahi, and Chrupała 2022).

Feedback-based learning

Learning from perceived multimodal input is an important mechanism, especially in the early stages of development. Nevertheless, an additional mechanism comes into play as soon as children start to produce language themselves, thus becoming able to receive *feedback* from more linguistically knowledgeable interlocutors (e.g., caregivers) (Warlaumont, Richards, Gilkerson, et al. 2014; Eve V. Clark 2018; Tsuji, Cristia, and Dupoux 2021).

One specific form of feedback that has received much attention is when the caregiver provides explicit reformulation to the child's inadequate use of words (R. Brown 1973; Chouinard and Eve V. Clark 2003; Saxton, Houston–Price, and Dawson 2005; Hiller and Fernandez 2016). Nevertheless, explicit reformulation is not the only way children can get useful feedback on their early productions. For instance, the feedback that signals communicative success/failure to the child – even in an implicit form – can also play a role. Below we elaborate on the nature and potential usefulness of this – more general – mechanism which we call **Communicative Feedback** (hereafter CF).

When children start to talk, they immediately start putting words to use in social interaction to try and establish coordinated communication. This coordination aims

at achieving various goals such as directing the interlocutor's attention (e.g., "A duck!") or requesting something (e.g., "I am thirsty!"), among many other communicative intents that children demonstrate very early in life (C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. 1996; Casillas and Hilbrink 2020; Nikolaus, Maes, Auguste, et al. 2022).

Importantly, children are sensitive to when coordination appears to *break down* without necessarily requiring explicit correction or even a verbal response from the caregiver. In fact, the child might feel misunderstood merely by not getting the reaction she expected (e.g., a puzzled or a still face) or by not getting the exact object she requested (e.g. Tronick, Als, Adamson, et al. 1978; Markova and Legerstee 2006). On such occasions, the child may not be offered the correct linguistic form as in reformulation-based (or corrective) feedback, but communication breakdown represents in and of itself a negative feedback, a cue to the child that her way of using words was not correct and that it should be revised for communication to be re-established or "repaired" (Eve V. Clark 2018; Eve V. Clark 2020). Vice versa, successful coordination (i.e., a contingent response or action from the caregiver) signals to the child that her use of words was probably adequate, encouraging (or reinforcing) this use in future conversations.

Compared to explicit corrective feedback, CF relies on sensitivity to broad coordination and mis-coordination cues that are fundamental to reach shared understanding in any linguistic exchange (see "communicative grounding" (H. H. Clark 1996)). It is, thus, arguably more pervasive in child-caregiver conversations and less dependent on parenting styles, Socioeconomic Status (SES) or culture (Childers, Vaughan, and Burquest 2007; Mesman, Minter, Angnged, et al. 2018).

Previous experimental research has explored a simple form of Communicative Feedback and how it can help with language acquisition, especially regarding the emergence of speech-related vocalization (Oller 2000). When the child produces a sound that contains speech-related vocalization (as opposed to other non-speech types of vocalization such as cry or laugh), the child is more likely to receive an immediate, positive response from the caregiver than if the produced sound is not speech-related. Critically, the fact of receiving a response from the caregiver (that is contingent of the production of speech) encourages the child to subsequently produce more speech-related vocalizations (K. Bloom 1988; Goldstein, King, and West 2003; Goldstein and Schwade 2008; Warlaumont, Richards, Gilkerson, et al. 2014).

To the best of our knowledge, no previous modeling work has investigated the role that CF could play in *semantic learning* or how CF may interact with the – more studied – class of semantic learning mechanisms that are based on perception alone such as XSL.

6.1.1. The current study

This work aims at providing a comprehensive account of early semantic learning combining both input-based learning through XSL and feedback-based learning through CF. The learning account we propose is very similar to – and in fact, can be

seen as a computational instantiation of – the "original word game" proposed by R. Brown (1968):

> *"The original word game is the operation of linguistic reference in first language learning. At least two people are required: One who knows the language (the tutor) and one who is learning (the player) ... The tutor names things in accordance with the semantic customs of the community. The player forms hypotheses about the categorical nature of the things named. He tests his hypotheses by trying to name new things correctly. The tutor compares the player's utterances with his own anticipations of such utterances and, in this way, checks the accuracy of fit between his own categories and those of the player. He improves the fit by correction."* [1]

Here we focus on a simple case of semantic learning where the meaning can be derived from concrete visual scenes. We use an integrated model to characterize the child's learning both during the input-based and during the feedback-based learning phase. In the input-based learning phase, the model optimizes the generative probability of the tutor's utterances given the visual scenes. This probability is refined thanks to exposure to several situations (i.e., XSL).

In the feedback-based learning phase, the same language model is now used to generate utterances given a scene. The adequacy of the utterance is evaluated against the gold standard descriptions of the scene (representing the tutor's superior knowledge). The adequacy value is a continuous number we use to characterize the valence of the Communicative Feedback: The higher the adequacy, the more likely the child receives signals of communication success from the tutor (e.g., a positive, contingent reaction). Vice versa, the lower the value, the more likely the child receives signals of communication breakdown (e.g. a puzzled face or a non-contingent reaction). The model gets updated via Reinforcement Learning (RL) using the adequacy value as a reward.

Using this computational framework, we study the role of CF in early semantics acquisition. In addition, we investigate how CF interacts with XSL. We evaluate and compare these two mechanisms in terms of how they fare on a wide range of semantic tasks including both word-level (nouns, adjectives, and verbs) and sentence-level meaning acquisition (e.g. semantic roles).

Combining some kind of (weakly) supervised learning model with reinforcement learning is not a new technique. Such a setup has been used in previous NLP work (Ranzato, Chopra, Auli, et al. 2016; Rennie, Marcheret, Mroueh, et al. 2017). The novelty of our work is to use these tools to instantiate new hypotheses about early language acquisition and to test these hypotheses using a benchmark of language acquisition tasks, similar to the tasks used to study children's semantic learning in laboratory experiments.

This chapter is organized as follows. First we present the cross-modal dataset we use in this work and introduce the modeling framework. We explain how we instantiate

1. Note that in our work, the fit of the semantic knowledge is not necessarily improved by *correction*, but rather by broad cues about the success or failure of the communicative coordination.

both the input-based mechanism (XSL) and the feedback-based mechanism (CF) using tools from NLP and computer vision. Next, we present the experiments we run: each representing a learning scenario, including scenarios combining both input and feedback-based mechanisms. Next, we test the extent to which these models learn various aspects of semantics. Finally, we discuss the results in the light of the literature on early language learning.

To ensure reproducibility, we make the source code for the model and all experiments publicly available. [2]

6.2. Methods

6.2.1. Data

We used the Abstract Scenes dataset 1.1 (Zitnick and Parikh 2013; Zitnick, Parikh, and Vanderwende 2013), which contains 10K crowd-sourced images each with 6 corresponding short descriptive captions in English. The images are clip-art scenes involving one or two children engaged in different actions involving a set of different objects and animals. [3] The corresponding captions were crowd-sourced from a different set of annotators. [4] Two example scenes along with descriptions can be found in Figure 5.1.

We use this dataset as it allows us to evaluate the learning of visually-grounded semantics on the word-level and sentence-level, using recently proposed evaluation tasks by (Nikolaus and Fourtassi 2021a) (see also Section 6.2.4). Other studies on XSL have used larger dataset with naturalistic images (e.g. Lin, Maire, Belongie, et al. 2014; Plummer, L. Wang, Cervantes, et al. 2015). However, there is currently no similar evaluation method available for these datasets that allows for detailed examination of the learned visually grounded semantics. We divide the data into training (80%), validation (10%) and test splits (10%) as proposed in Nikolaus and Fourtassi (2021a).

6.2.2. Modeling framework

We develop an integrated modeling framework that can both learn from pairs of images and sentences in the context of XSL *and* and to produce its own sentences given an image to learn using rewards (CF). This framework will allow to assess various learning scenario, including ones that combine both XSL and CF.

Some previous work in NLP has used image-sentence ranking models (Hodosh, Young, and Hockenmaier 2013) to learn the alignment of visual and language representations, and thus to model cross-modal XSL (Chrupała, Gelderloos, and Alishahi

2. https://github.com/mitjanikolaus/perception-and-production-based-learning

3. Annotators were asked to "create an illustration for a children's story book by creating a realistic scene" given a set of clip art objects (Zitnick and Parikh 2013).

4. Annotators were asked to write "simple sentences describing different parts of the scene". They were asked to refer to the children by the names "Jenny" and "Mike" (Zitnick, Parikh, and Vanderwende 2013).

2017; Vong, Orhan, and Lake 2021; Nikolaus and Fourtassi 2021a). However, these models are not designed to *produce* new utterances given an image.

As we are here interested in both input-based and feedback-based learning, we use a different computational framework borrowed from studies on image captioning (Vinyals, Toshev, Bengio, et al. 2015; K. Xu, Ba, Kiros, et al. 2015; Anderson, X. He, Buehler, et al. 2018). This framework is based on a language model conditioned on the image. Just like the image-sentence ranking models, here the model is trained using pairs of images and captions, instantiating learning in a XSL fashion. In addition, the same language model can be used to generate sentences given an image, which we used to instantiate the feedback-based mechanism CF. Since the goal is not to produce a state-of-art image captioning model, we consider a basic implementation close to that used in Anderson, X. He, Buehler, et al. (2018).

To process the images, we use ResNet 50 (K. He, Zhang, Ren, et al. 2016) pre-trained on ImageNet (Russakovsky, Deng, Su, et al. 2015), assuming that the visual system of the child has already been developed to some degree allowing her to process visual scene. [5] We discard the final classification layer and fine-tune the remaining layers of this CNN during the training progress to encode the images in our dataset.

Conditioned on this image encoding, an autoregressive language model learns to produce utterances word by word: The words of a sentence are passed through a linear word embedding layer and then fed, together with the encoded image features [6], into a one-layer LSTM (Hochreiter and Schmidhuber 1997).

6.2.3. Model Training

Input-based learning is realized by training the model using a cross-entropy loss. The model is given pairs of images with corresponding sentences and uses these to learn a mapping from the visual to the language domain. Given an image i and a target ground-truth sentence s consisting of the words $w_1, \ldots, w_T$, the loss is defined as:

$$\mathscr{L}_{XSL}(\theta) = -\sum_{t=1}^{T} \log p_\theta(w_t | w_{<t}; i) \tag{6.1}$$

Feedback-based learning is instantiated by training the model using REINFORCE (Williams 1992). To operationalize the Communicative Feedback (i.e., the reward), we calculate the BLEU score (Papineni, Roukos, Ward, et al. 2001) between the produced sentence and all 6 reference descriptions/captions from the dataset, taking into

5. As commonly applied in other multimodal XSL work (Chrupała, À. Kádár, and Alishahi 2015; Khorrami and Räsänen 2021).

6. While Vinyals, Toshev, Bengio, et al. (2015) fed the image features only at the first timestep into the LSTM, here we feed it at every timestep as this showed to improve performance on our evaluation substantially. An explanation could be that when feeding the image features only at the first timestep the model gradually *forgets* about the input, and relies more on the language modeling task of next-word prediction, which does not aid the learning of *visually-grounded* semantics.

account both the quality of semantics as well as word order (n-gram sequences).[7] Crucially, the BLEU score takes into account the fact that there is not only one correct sentence for each image, but rather a range of equally adequate ways to describe the same scene. In particular, if the model produces an exact imitation of one of the reference sentences, it obtains the highest BLEU score, even if the other 5 reference sentences are very different.

Given an image i, the sampled sentence from the model $s_m = w_1, \ldots, w_T$ and the 6 reference sentences $S_{ref} = s_1, \ldots, s_6$, the loss is defined as follows:

$$\mathscr{L}_{CF}(\theta) = -\sum_{t=1}^{T} r(s_m, S_{ref}) \cdot \log p_\theta(w_t) \tag{6.2}$$

where $r(s_m, S_{ref}) = BLEU(s_m, S_{ref})$.

More details on model hyperparameters can be found in Appendix C.1.

6.2.4. Model Evaluation

In order to evaluate the model's acquisition of visually-grounded semantics, we use an evaluation method proposed by Nikolaus and Fourtassi (2021a). It is based on a two-alternative forced choice design, similar to what is typically done to evaluate children's knowledge in laboratory experiments (Bergelson and Swingley 2012; Noble, Rowland, and Pine 2011; Gertner and Fisher 2012). Note that the models are not trained to optimize these tasks. The tasks are only used during the evaluation phase and they test if the models learn various aspects of semantics as a "side product" of XSL and CF. Indeed, when we evaluate children's knowledge in the lab, we do not suppose they have acquired their knowledge by being trained on lab tasks.

These tasks test the model's learning of grounded semantics on the word level (nouns, adjectives, verbs) and sentence level (adjective-noun dependencies, verb-noun dependencies, semantic roles). A task involves multiple test trials, each consists of an image, a target sentence and a distractor sentence: (i, s_t, s_d). Critically, each test trial is *counter-balanced* to control for linguistic biases (e.g., that Jenny occurs most frequently as semantic agent and Mike more as a semantic patient), in a way that a language model that does not have access to the image data performs at chance (see also Figure 5.1, more examples are shown in Figure 5.2).[8]

The model's accuracy at choosing the correct sentence s_t given the image i indicates how well it has learned visually grounded semantics for the phenomenon under study. We operationalize the model's choice for a trial by calculating both the perplexity of

7. While the BLEU score only measures the adequacy of the children's produced sentence, we used it here as a proxy for adults' Communicative Feedback. The assumption being that the degree to which adults provide positive, contingent responses (i.e., cues of coordination success) depends closely on children's production adequacy as was shown previously, though in a different context, by Warlaumont, Richards, Gilkerson, et al. (2014). We return to this assumption in the Discussion.

8. Besides controlling for linguistic biases, the evaluation sets also control for some potential visual biases, e.g., that the semantic agent may occur more frequently on the left side of the image (see Nikolaus and Fourtassi (2021a) for more details).

the target sentence s_t given the image i and the perplexity of the distractor sentence s_d given i. If the perplexity of the target sentence s_t is lower, the trial has been successfully completed.

6.3. Analyses

6.3.1. Comparing learning scenarios

We study and compare four different learning scenarios:

`XSL`**: Pure input-based learning** In this scenario, the model learns only using XSL. It represents our baseline against which we compare configurations including CF.

`Alt`**: Alternating between input-based and feedback-based learning** Here, the model switches between the XSL and CF objectives throughout the entire learning process.

`XSL+CF`**: First pure input-based learning, then pure feedback-based learning** We train the model until convergence using XSL, and afterwards we fine tune the model using CF.

`XSL+Alt`**: First pure input-based learning, then alternation** The model is first trained until convergence using XSL, but afterwards, we alternate between XSL and CF. This scenario is intuitively the most plausible one: Once the language learner starts to speak (i.e. produce their own utterances), this does not mean that they stop to engage in input-based learning. Rather, they continue learning using both mechanisms.

Accuracies for the four different learning scenarios are reported in Table 6.1.[9] The scenario `XSL` learns word-level and sentence-level semantics relatively well compared to the other scenarios. It only appears to struggle with the verbs and the verb-noun dependencies. This fact highlights the role of XSL as a major learning mechanism. When looking at the results of `Alt`, we can conclude that combining XSL and CF from the start deteriorates the performance (compared to `XSL` alone) of all metrics. This deterioration was observed regardless of the frequency of alternation between XSL and CF (for direct comparison with `XSL+Alt` we only report results using one XSL update every 10 CF updates in Table 6.1, but see Appendix C.2 for results with other alternation frequencies).

Moving to the more plausible scenarios (where feedback-based learning comes into play only after a phase of pure input-based learning), we found that for `XSL+CF`, we have, on the one hand, an increase in performance (compared to the baseline

9. Note that the results are not directly comparable to the results for the cross-situational learner in Nikolaus and Fourtassi (2021a), see Appendix C.4 for more detail.

	Evaluation task	Accuracy			
		XSL	Alt	XSL+CF	XSL+Alt
Word-level Semantics	Nouns: Persons	**0.87** ± 0.03	0.51 ± 0.01	0.79 ± 0.03	**0.87** ± 0.04
	Nouns: Animals	**0.99** ± 0.01	0.53 ± 0.05	0.98 ± 0.01	**0.99** ± 0.00
	Nouns: Objects	0.94 ± 0.01	0.51 ± 0.01	0.94 ± 0.00	**0.95** ± 0.00
	Verbs	0.55 ± 0.05	0.50 ± 0.00	**0.77** ± 0.04	0.73 ± 0.05
	Adjectives	0.75 ± 0.02	0.50 ± 0.01	0.81 ± 0.03	**0.82** ± 0.02
Sentence-level Semantics	Adj-noun dependencies	0.61 ± 0.03	0.50 ± 0.00	0.62 ± 0.02	**0.63** ± 0.03
	Verb-noun dependencies	0.55 ± 0.03	0.50 ± 0.00	**0.72** ± 0.05	0.68 ± 0.02
	Semantic roles	**0.65** ± 0.07	0.50 ± 0.01	0.61 ± 0.05	0.61 ± 0.07
	Average	0.74 ± 0.01	0.51 ± 0.01	0.78 ± 0.01	**0.79** ± 0.01

Table 6.1. – Accuracy (mean and standard deviation over 5 runs with different random initializations) for all semantic evaluation tasks for different learning scenarios.

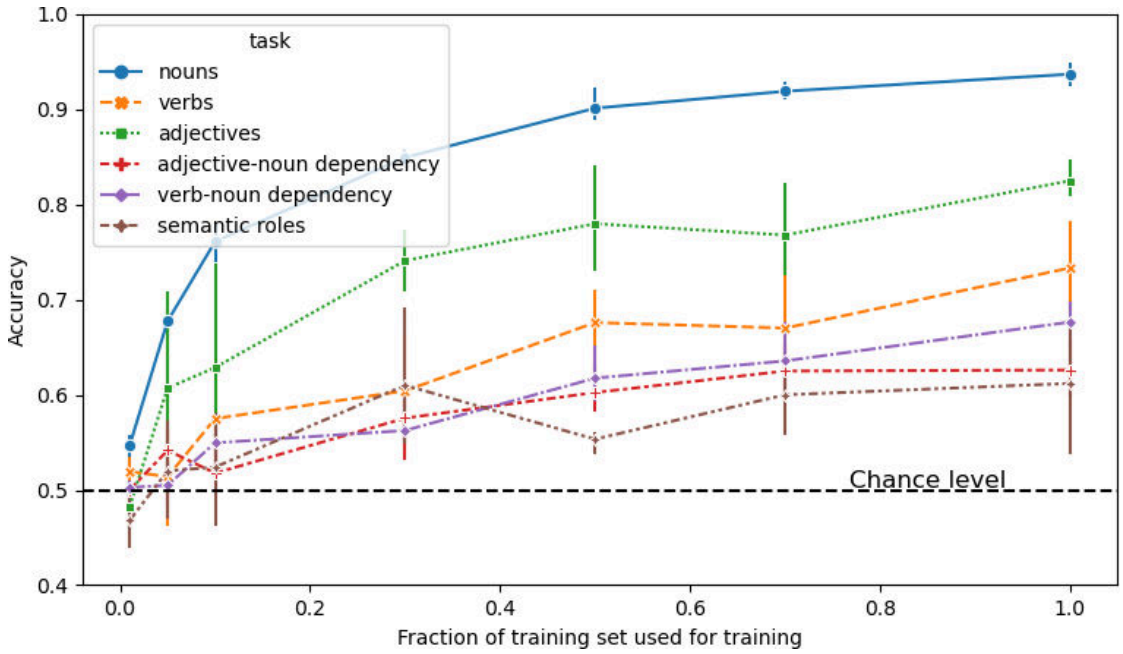

Figure 6.1. – Accuracy as a function of training set size for best performing learning setup (XSL+Alt). Vertical bars indicate the standard deviation over 5 runs. Accuracies for all noun categories were averaged.

XSL) in some categories like "verbs," "adjectives,", and "verb-noun dependencies." On the other hand, we observe a decrease in other categories, especially the category "persons." Finally, the scenario XSL+Alt leads to the best overall results except for verbs and semantic roles, but the difference is within the margin of error. Here we

only show results of XSL+Alt using one XSL update every 10 CF updates (which seems to optimize performance), but other – both lower and higher – ratios only marginally change the model's behavior and the conclusions remain the same (see Appendix C.2).

Appendix C.3 contains a comparison of the BLEU scores (our measure of utterance adequacy) for the different learning scenarios. Consistent with our semantic evaluation results, XSL+Alt leads to the highest BLEU score.

6.3.2. Developmental Trajectories

Results in Table 1 show evaluation scores after the model has converged on the entire dataset. Here we test the developmental trajectories in each semantic category using different data sizes as a proxy for progression in time. Figure 6.1 shows the accuracy for different tasks when the best-performing model XSL+Alt is trained on different training data sizes. Already with very small training data (10% of the original training set, 800 examples), nouns and adjectives are learned to a high degree. Verbs and sentence-level semantics are learned only with larger training set sizes.

6.3.3. Effect of the data size used for XSL pre-training

In the best performing configuration, XSL+Alt, the model was first pre-trained on the entire dataset using XSL, and then trained further using XSL and CF, using again the entire dataset. However, in real life, children spend only a fraction of their learning time (generally the first year of their life) doing pure input-based learning. Thus, here we test how different fractions of pre-training data influence performance.

Figure 6.2 shows the average task accuracy (cf. last row in Table 6.1) for XSL+Alt models that are pre-trained until convergence on training datasets of different size, and then trained in alternation between XSL and CF on the full training dataset until convergence. While the results indicate that more pre-training data is better, we observe a steep gain in average task accuracy starting from pre-training only 5% of the data (up from chance level with 0% pre-training, a limit case that corresponds to the scenario of Alt alone), indicating that even a small amount of input-based training is useful to initiate a successful learning trajectory.

6.4. Discussion

How do children learn the meanings of words and sentences in their native language? Previous modeling effort has largely focused on input-based learning mechanisms such as XSL. However, children do not learn only by mere exposure to the perceptual cross-modal input, they also practice their early – albeit rudimentary– knowledge and receive feedback from caregivers, which allows them to correct/refine this knowledge (Eve V. Clark 2018; Eve V. Clark 2020). Here we investigated one possible feedback mechanism on children's early production (CF), that relies on general coordination and mis-coordination cues, and does not necessarily require the caregiver

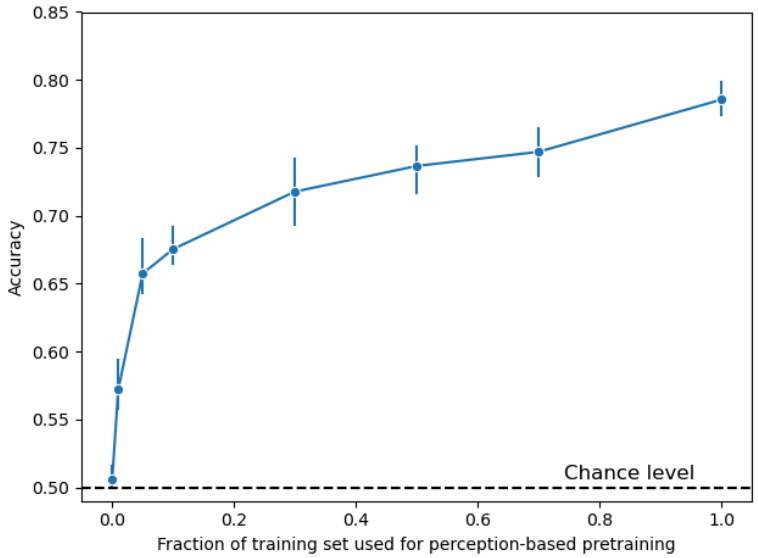

Figure 6.2. – Average accuracy as a function of amount of input-based pre-training for the best performing learning setup (`XSL+Alt`). Vertical bars indicate the standard deviation over 5 runs.

providing an explicit correction.

We proposed a computational model that integrates both XSL and CF, allowing us to study how these two mechanisms could interact in early semantic learning. The same model learns both from perceptual input and from feedback on production through reinforcement. We tested various learning scenarios that varied in their plausibility given our understanding of how children's learn language. Crucially, we found that the most plausible learning scenario (i.e., `XSL+Alt`) – where the model first learns by leveraging the linguistic *input*, and second through alternating input-based and feedback-based learning – is also the one that leads to the best overall performance on most semantic tasks.

The fact that `XSL+Alt` performed better than `XSL` alone confirms the main hypothesis of this work: CF plays a role in semantic learning above and beyond XSL. In addition, the fact that `Alt` – which alternates input-based and feedback-based learning from the start – hurts performance compared to XSL, suggests that for CF to be effective, it requires a first phase of learning through input, which is an intuitive finding since the model has first to be exposed to enough linguistic/semantic input to be able to start producing – at least partially – meaningful utterances (for which RL is more useful). This finding also corresponds to children's learning trajectory where they only start producing words (and receiving feedback on them) after a period of pure input-based learning. [10]

10. Children do not generally utter their first words until they are about 10 months old (Michael C. Frank, Braginsky, Yurovsky, et al. 2021) while they already understand certain words well before that age (Bergelson and Swingley 2012), indicating that they engage in a input-based learning well before

Interactions between input-based and feedback-based learning Another interesting finding of this work is that XSL+Alt (e.g., alternating XSL and CF after a period of pure XSL) performs better than XSL+CF (i.e., using CF alone after a period of pure XSL). This finding means that when CF is combined with XSL, it leads to improvement in performance compared to when either XSL or CF operates alone or in a sequential fashion. In other words, we found that XSL and CF interact *synergistically* to improve performance. In what follows, we examine this observed synergy in more details.

Results in Table 3 show that while XSL+CF improved performance on "verbs" compared to XSL, it also led to a significant drop in the category "persons." [11] We speculate that by using reinforcement learning alone, XSL+CF explores the hypothesis space and picks short utterances that lead to a high reward signal and continues (re)producing them. While this behavior could lead to improvement for the parts of the language that are well covered by this local space (e.g., verbs), it can also lead to a drop in performance for the other aspects. In particular, here the difference between Jenny and Mike in the category "persons" may become forgotten.

Qualitative and quantitative investigation of the model's behavior supports our speculation. For example, when we sample sentences randomly from the productions of XSL+CF and XSL+Alt given images in the validation set, we observed that while XSL+CF produces a variety of verbs (similar to XSL+Alt), it tends to produce systematically shorter utterances involving disproportionately only one person (see Table 11 in Appendix C.5).

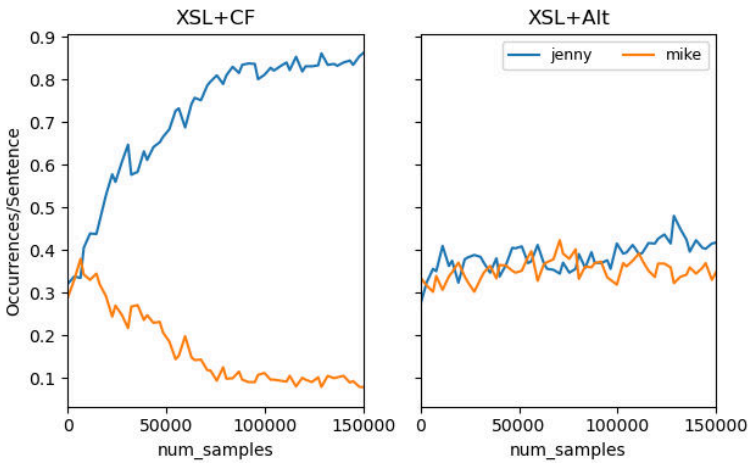

Figure 6.3. – Comparison of the fraction of occurrences of persons ("jenny" and "mike") in sentences produced during training of the XSL+CF (left) and XSL+Alt (right) training setups. The graphs only display the second training step, not the pre-training using XSL.

starting to produce their own utterances.

11. The drop in "persons" could explain the slight drop in "semantic roles," (as distinguishing the persons is a prerequisite to understand semantic roles) however this slight drop is within the margin of error, so we could not draw strong conclusions about the difference with XSL for this category.

Figure 6.3 confirms this observation quantitatively: XSL+CF increasingly produces sentences involving Jenny, but decreasingly sentences involving Mike. This fact leads to the situation where the model gets less feedback on the difference between Jenny and Mike and, therefore, *unlearns* this distinction to some degree.

For XSL+Alt, the fraction of sentences involving Jenny and Mike remains largely constant, thus avoiding the problem faced by XSL+CF. At the same time, XSL+Alt keeps a balanced coverage of verbs allowing it to maintain the good scores achieved by XSL+CF on this category (see Appendix C.5 for a quantitative analysis comparing the production of verbs in both models).

The conclusion we draw from comparing XSL+CF and XSL+Alt is that, even after a period of pure XSL, continuing to learn through XSL from time to time while doing reinforcement on production helps the model not to get biased towards a subset of the language it is supposed to learn. Similar phenomena of "language drift" – due to reinforcement learning operating alone – have been observed in another line of work studying emergent communication systems (Lowe, Gupta, Foerster, et al. 2020; Lazaridou, Potapenko, and Tieleman 2020).

Learning Trajectories The best performing model, i.e. XSL+Alt, not only instantiates – intuitively – the most plausible learning scenario in early childhood, it also recapitulates some specific findings in the language development literature about the timeline of semantic learning. For example, it learns nouns before predicates (adjectives and verbs), resonating with previous findings about the "noun bias" (Gentner 1982; Bates, V. Marchman, Thal, et al. 1994; Michael C. Frank, Braginsky, Yurovsky, et al. 2021). That said, the models' performance on verbs (relative to other parts of speech) should be interpreted with caution given the fact that we only used static images in both training and testing. In real life, children learn verbs from dynamic actions and some experimental studies also evaluate verb learning use videos instead of static images (Roberta Michnick Golinkoff, Kathryn Hirsh-Pasek, Cauley, et al. 1987; Gertner, Fisher, and Eisengart 2006).

The model shows a rather late onset of understanding sentence-level semantics such as semantic roles, only after a sizable lexicon has been acquired. This fact mirrors, e.g., the finding that children show evidence of recognize semantic roles in a sentence during their second year of life (Roberta Michnick Golinkoff, Ma, Song, et al. 2013), that is, at an age when they have already acquired a substantial vocabulary (Michael C. Frank, Braginsky, Yurovsky, et al. 2021). Note that the model's performance on sentence-level semantics remains relatively low compared to word-level semantics even when learning from the entire dataset. It is difficult, based only on the current results, to conclude whether more data will lead to improvement in sentence-level semantics or whether the model has already reached its ceiling performance due to structural limitations (e.g., the lack of higher-level conceptual knowledge about semantic agency).

Limitations and future research directions While our modeling work has allowed us to test crucial hypotheses about semantic learning, it used – like any modeling work – simplifying assumptions about the phenomenon under study. For example, here we used an integrated model for both perception and production. This choice was primarily motivated by parsimony. While it allowed us to provide a direct comparison of XSL and CF, it abstracted away limitations in children's production abilities compared to perception (e.g., due to immature motor/articulatory skills) and from difficulties that children face when trying to coordinate production with perception (e.g. E V Clark and Hecht 1983). In addition, we did not account for constraints on children's information processing abilities during the learning process (e.g., limited attention span and working memory), and how these constraints may, for example, translate in the learner focusing on specific parts of the input (Gelderloos, Kamelabad, and Alishahi 2020).

More generally, the current work focused on investigating the input-output mapping problem for semantic learning and how Communicative Feedback can help such learning. It did not intend to account for the exact cognitive processes that operate in children's mind nor did it take into account specific cognitive limitations and constraints when trying to achieve this mapping. Thus, this work is best situated at the computational level of analysis (Marr 1982), which is a necessary first step towards a deeper understanding of the cognitive implementation.

Another simplifying assumption of this work was the use of the BLEU score as a reward to the model when learning through reinforcement. In other words, we used a measure that only evaluates the extent to which the learner's utterance is correct as a proxy for how the teacher would react. While this assumption is grounded in previous experimental work showing that adults' responses are contingent on children's type of vocalization (Warlaumont, Richards, Gilkerson, et al. 2014), here we went beyond the broad distinction studied in this previous work (speech vs. non-speech) and assumed that adults' responses are also contingent on the adequacy of speech itself. That is, immediate, positive reaction from adults is more likely to follow correct/adequate speech from the child, which would encourage the re-use of adequate (but not inadequate) speech in subsequent conversations.

Note that the BLEU score feeds the model with ideal information whereas the feedback that children receive in real life is highly dynamic, multimodal and noisy. While, as we said above, the current work took a computational level of analysis approach that only studied learning under optimal conditions, future work is required to (1) estimate the quality and frequency of Communicative Feedback in child-caregiver conversations (CHILDES (MacWhinney 2014)) and (2) use these findings to assess the scalability of the current proposal to account for child's language use and development in the real world.

In conclusion, this work provides a quantitative proof of concept about the role feedback-based learning can play in semantic knowledge acquisition together with input-based learning. An important finding was that combining both mechanisms leads to synergistic learning. One question for future experimental work is whether such synergy can be observed in controlled behavioral experiments.

Part IV.

Discussion and Conclusion

Summary

7. Discussion and Conclusion

The following sections cover discussions of multiple areas within the broader context of this thesis. Additionally, most sections propose possible avenues for future research in the respective directions. Finally, we summarize the major results of the thesis in the conclusion.

7.1. Behavioral Experiments

While the results of the corpus studies highlight the potential of Communicative Feedback both for learning to produce intelligible as well as grammatical utterances, the performed analysis only provided *correlational evidence* for such mechanisms. In order to draw more definite conclusions on the role of Communicative Feedback on Language Acquisition, future work could include experimental studies that test the possible effects of these communicative signals in more controlled conditions. Possible setups could involve simple word-learning paradigms with young infants, as commonly employed in the field (e.g., Woodward and E. M. Markman 1998; Roberta M. Golinkoff, Kathy Hirsh-Pasek, Bailey, et al. 1992; Pruden, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, et al. 2006; Schafer 2005; Woodward, E. M. Markman, and Fitzsimmons 1994). In these paradigms, children are usually exposed multiple times to a target novel word-referent mapping before their comprehension of the novel words is assessed. To test the possible role of Communicative Feedback in word learning, some of the exposure trials could be replaced by production trials that are responded by feedback in the form of acknowledgements or clarification requests from an interlocutor. [1] By manipulating the contingency of the feedback on the correctness of the child's production, the effect of such feedback signals on learning could be investigated in a highly controlled setup.

7.2. On the Role of Acknowledgements

Amongst all kinds of Communicative Feedback signals that can be provided by the listener (Figure 1.2), "Acknowledgement" stood out as the feedback mechanism that has received the least attention regarding its potential role in fine-tuning children's linguistic knowledge. This lack of research is even more surprising given the high

1. Another promising direction could be to attempt to improve *retention* of newly learned words by means of production and feedback trials (see also Horst and Samuelson 2008).

frequency of such explicit positive feedback in naturalistic conversations (Dideriksen, Fusaroli, Tylén, et al. 2019; Dingemanse and Liesenfeld 2022) and presence in different languages around the world (Cutrone 2005; Maynard 1990; Liesenfeld and Dingemanse 2022).

One issue with existing studies is that they explored the role of acknowledgements only within a set of other communicative devices (e.g., as part of a set of interjections, narrative-eliciting behaviors, or repetitions in general), the only exception being our corpus study presented in Chapter 4 (Nikolaus, Prévot, and Fourtassi 2023). More controlled studies are required to test the specific role of acknowledgments using different methodologies and for different aspects and stages of language acquisition.

7.3. Implicit Communicative Feedback: Contingency

In order to enable studies of more implicit Communicative Feedback signals in the form of contingency, researchers are currently still missing reliable measures to estimate conversational contingency in dialog.

One main challenge for the development of improved measures is the fact that judging contingency from a third point of view requires inferring the speaker's communicative intent and interpreting the listener's response. [2]

Ideally, the endeavor to improve measures of content-contingency should be pursued within a computational agenda that aims at automatizing them as well. This is important to avoid subjective biases in human annotation, facilitate cross-lab and large-scale comparison, leading to more cumulative science on this question.

An automatic measure should, at a minimum, be able to evaluate the similarity of pairs of utterances while also capturing their complementarity at the speech act level as in the case of adjacency pairs (Schegloff and Sacks 1973; Nikolaus, Maes, Auguste, et al. 2021).

To achieve this goal, the child developmental community would benefit from ongoing effort in natural language processing methods on the evaluation of coherence in dialog systems (Dziri, Kamalloo, Mathewson, et al. 2019; Cervone, Stepanov, and Riccardi 2018; Cervone and Riccardi 2020; Higashinaka, Meguro, Imamura, et al. 2014).

One shortcoming of automatic measures is that they usually over-emphasize internal discourse coherence (e.g., the extent that two turns are semantically or "logically" related) rather than subtle context-dependent pragmatics. However, in conversations, meaning is usually constructed in a highly incremental and inter-subjective fashion (e.g., Fusaroli, Rączaszek-Leonardi, and Tylén 2014), and thus, a deep understanding of the discourse as whole as well as the interlocutors' common ground is required to judge the contingency of a turn. That said, we suspect most of child-caregiver interactions in early childhood would still be reasonably captured by rather simple measures of contingency. As children's conversations become longer and more sophisticated,

2. See also Section 1.5.3

more advanced methods for measuring contingency and their role as Communicative Feedback will have to be developed.

7.4. Multimodal Communicative Feedback

Most work that has been performed within the framework of Communicative Feedback (including the work performed as part of this thesis) focused solely on studying *verbal* Communicative Feedback signals. However, many feedback signals are communicated non-verbally in face-to-face conversations (Paggio and Navarretta 2013; Allwood, Cerrato, Jokinen, et al. 2007; Brunner 1979; Dittmann and Llewellyn 1968; Bodur, Nikolaus, Prévot, et al. 2023). Further, measuring implicit feedback in the form of action contingency (e.g., directly providing a requested object in place of a verbal response; see also section 1.5.3.3) requires access to the multimodal environment of the interlocutors. Especially in early childhood, where a large part of communication is of referential nature (C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. 1996), verbal-only analyses of conversations are far from providing an exhaustive picture of the interaction.

As a consequence, current studies on the role of Communicative Feedback for language acquisition are likely *underestimating* effects on learning, as they are missing a large quantity of signals.

Analyses on multimodal conversational corpora are required to obtain a more complete picture of Communicative Feedback behavior in interaction. Recently, a range of suitable corpora have been recorded and should be leveraged for more comprehensive analyses: ChiCo (Bodur, Nikolaus, Kassim, et al. 2021), ECOLANG (Shi, Gu, and Vigliocco 2022), and CANDOR (Reece, Cooney, Bull, et al. 2023).

7.5. Cross- and within-cultural variability

One motivation behind our focus on Communicative Feedback as a mechanism for language learning is that it relies on what is generally assumed to be universal principles of human communication. Therefore, the learning mechanism is more likely to be universal than mechanisms that require parents to adopt explicit teaching strategies towards children (e.g., corrections).

While, as we mentioned earlier, there is evidence that Communicative Feedback is used across many cultures in adult-adult conversations, there is surprisingly very few studies capitalizing on this potential to understand how CF plays out in the context of children's first interactions and to investigate how it influences language development across cultures, including in non-WEIRD[3] ones (Henrich, Heine, and Norenzayan 2010). For example, the role of CF for children in conversation with interlocutors other than their caregivers, such as with older siblings, could play a more important role in cultures with relatively less frequent adults' child-directed speech (Shneidman and

3. Western, Educated, Industrialized, Rich, and Democratic.

Goldin-Meadow 2012; Casillas, P. Brown, and Levinson 2020; Ochs and Schieffelin 1984). Most recently, Cristia, Gautheron, and Colleran (2023) analyzed recordings of children growing up on Malakula island, Vanuatu and found that children's own vocalization counts are more highly correlated with the vocalization counts of their peers, than with the vocalization counts of their adult caregivers.

The study of variability, not only across but also within cultures is crucial. Indeed, many aspects of conversational dynamics have been shown to vary depending on the conversational partners, contexts, and languages. For example, child-child conversations are on average shorter and less coherent than child-caregiver conversations (Dunn and C. Kendrick 1982; Barton and Tomasello 1994), and the use of certain communicative signals varies between affiliative and task-oriented conversations (Dideriksen, Christiansen, Tylén, et al. 2020) as well as between languages even in culturally similar communities (Dideriksen, Christiansen, Dingemanse, et al. 2022).

More research studying how Communicative Feedback plays out in a wider range of contexts (including with various conversational partners, such as peers) is needed to shed light on possibly universal mechanisms supporting the acquisition of language across languages and cultures.

7.6. Communicative Feedback in Later Stages of Language Acquisition

Most of the studies we reviewed have investigated the role of Communicative Feedback in the pre-verbal stage or for children producing their first words. It is unclear how Communicative Feedback would play out in later stages of language development and more future work is required to address this question in detail.

On the one hand, we speculate that as children become more competent speakers, the role of CF would diminish for the learning of some aspects of *form* such as syntax and morphology. If the child makes mistakes that do *not* impede understanding (e.g., "go-ed" instead of "went"), CF may not provide a useful learning signal as children can still receive explicit or implicit signals of communicative success (see also Marcus 1993; R. Brown and Hanlon 1970, regarding the role of corrective feedback). CF is more useful regarding mistakes that are big enough to risk impeding the transmission of the child's communicative intents (e.g., "I bit dog" instead of "Dog bit me"). Such big mistakes naturally occur more in the earlier stages. The results from our corpus analysis in Chapter 4 indicate however that feedback signals are given to almost *all* kinds of grammatical errors with comparable reliability. Still, future work is required to disentangle the effects of *communicative* and *corrective* feedback (in the form of recasts) in such studies, and in order to aggregate more conclusive evidence.

On the other hand, CF should continue to play a role regarding the acquisition of *meaning* throughout the learning process; errors in meaning often impede successful communication (e.g., when the child requests "ball" but they mean DOLL). In addition to meaning, we believe CF would continue help children refine their mastery of language *use*: A communicative intent can be phrased in various ways, and very often,

the choice of the correct phrasing depends on the context. In other words, even when the form and (literal) meaning of the utterance is sound, its use in a specific context could still be correct or incorrect (e.g., using a verb in present tense when talking about the past), leading to signals of communicative success or failure from the listener that the child can pick up on.

7.7. Computational Models of Communicative Feedback in Language Acquisition

As motivated in Chapter 6, computational models can serve as a useful tool to study long-term effects on learning in a controlled environment. The results of the computational simulations revealed insightful evidence for possible interactions between input-based and feedback-based learning.

As possible areas of future work the issue of dynamic feedback modeling (see Chapter 6) remains the most challenging. One way to circumvent this issue is to train models using aggregated conversational data from static corpora. A possibility to actually *address* the issue is to study language learning in controlled environments with a dynamically responding interlocutor agent. As a first step, one could study the acquisition of language in simple communication games (e.g., "Lewis signaling games"), where agents are learning to communicate as a means for coordination to solve well-defined problems/tasks such as a referential game (Lewis 1969).

Recently, this approach has been used in computational models to study how language is *emerging* in increasingly complex interactive contexts (Kirby and Hurford 2002; Mordatch and Abbeel 2018; Lazaridou, Peysakhovich, and M. Baroni 2017; Lazaridou and M. Baroni 2020; Galke, Ram, and Raviv 2022). In these studies, agents are usually updating their linguistic knowledge about form-meaning mappings using *Reinforcement Learning* (RL, Sutton and Barto 2018): The speaker agent is given a positive reward if the game outcome was successful, and negative otherwise. This reward signal can be seen as an instantiation of CF in that it provides the speaker with signals about communication success (or failure) that may have caused (or impeded) the successful accomplishment of the coordination task.

While such models have studied language creation/emergence, very similar computational tools can be used, in principle, to study language *acquisition*. In fact, some studies in this same literature successfully incorporated a language transmission component (from a pre-trained "teacher" to an untrained "student") in their multi-generational emergent communication frameworks (F. Li and Bowling 2019; Cogswell, J. Lu, S. Lee, et al. 2020; Y. Lu, Singhal, Strub, et al. 2020; Dagan, Hupkes, and Bruni 2021). That said, the goals of these studies has been still the study of language emergence across generations rather than the study of language acquisition of a child in an interactive context. [4]

4. Another line of work has studied the learning of natural language instructions, typically in game-like setups (Goldwasser and Roth 2014; Branavan, Chen, L. S. Zettlemoyer, et al. 2009; Misra, Langford, and Artzi 2017; X. Wang, Q. Huang, Celikyilmaz, et al. 2019; Hill, S. Clark, Hermann, et al. 2019; Hill,

We believe that computational models that specifically aim at modeling Communicative Feedback as a mechanism of language acquisition, even in a simplified context, are much needed. Besides, such models should not focus exclusively on feedback-based learning. Children learn both from the statistical regularities of the input and from social interaction; a helpful model of language acquisition should ideally integrate and contrast both components, as implemented in the model proposed in Chapter 6 and other methodologically related studies (Lazaridou, Potapenko, and Tieleman 2020; Lowe, Gupta, Foerster, et al. 2020).

7.8. Conclusion

This thesis described a novel framework for studying aspects of language acquisition in social interaction: Communicative Feedback, which underlies general principles of human communication, can support children to acquire their mother tongue. This hypothesis was explored by re-evaluating past research within the new framework (Section 1.5 in Chapter 1), two corpus studies (Chapters 3 and 4), as well as by developing computational models of the learning process (Chapter 6).

Overall, we find multiple pieces of evidence for the role of Communicative Feedback mechanisms in language acquisition. Such feedback signals have been found to be in principle useful for learning language in multiple stages of development. Still, the findings from the discussions in the preceding sections suggest that this thesis only forms a starting point for research on Communicative Feedback in language acquisition, many questions remain open for future investigations.

Several tools have been developed as part of this thesis and are shared for the community for future use on scalable research on language acquisition in social interaction. This includes models for the automatic annotation of speech acts in child-caregiver conversations (Chapter 2), models for the annotation of repetition-based clarification requests and acknowledgements in child-caregiver conversations (Chapter 4), as well as methods for the evaluation of semantic learning in multimodal neural networks (Chapter 5).

7.9. Outlook

In this final section, I will outline possible future research directions based on the findings of this thesis, as well as my personal skills and experience.

One main limitation of the corpus studies executed within the scope of this thesis (Chapters 3 and 4) is the lack of investigation of possible long-term effects of Communicative Feedback on language learning. Any evidence of learning beyond the scope of effects shown in the immediate follow-up utterance of the child were not

Tieleman, von Glehn, et al. 2020; Hill, S. Clark, Hermann, et al. 2020). In these studies, models learn to *understand* linguistic instructions with a task-dependent feedback signal. However, though interactive, these models do not instantiate the CF-based mechanism since agents do not produce language: The feedback they receive is rather about the behavior/actions they perform in the task.

explored as part of the analyses based on micro-conversations. As the learning signals under investigation are highly noisy, it is likely that certain learning effects can only be observed on the longer term. Somewhat related, it would be important to explore how such feedback can be leveraged when a language learner is exposed to it in alternation with other learning signals.

One way to start investigating these questions is to use computational modeling based on the feedback statistics found in the corpus analyses. The model will learn from the linguistic input (supervised learning) as well as the Communicative Feedback on the children's own utterances.

Such models will therefore combine supervised learning with reinforcement learning (RL) (Sutton and Barto 2018) techniques, and could take inspiration from the methods proposed in the growing literature on fine-tuning language models with reinforcement learning from human feedback (Stiennon, Ouyang, Jeff Wu, et al. 2020; Sokolov, Kreutzer, Lo, et al. 2016; Ouyang, Jeffrey Wu, Jiang, et al. 2022).

The proof-of-concept implementation of such a model presented in Chapter 6 showed that learning based on Communicative Feedback can improve performance above and beyond supervised learning across a wide range of semantic tasks, including both word- and sentence-level semantics. However, the findings of this work are limited to a small and artificial dataset, and the implementation of the reward signal was not based on realistic data. By drawing on results from the above mentioned corpus studies, it will be possible to model more realistic reward signals based on actual child-caregiver interactions.

More specifically, a first step in this project would be to train language models on corpora of child-directed speech (CHILDES; MacWhinney 2014). Then, a reward model will be trained based on the valence of received feedback to children's utterances (e.g., clarification request: negative reward; acknowledgement: positive reward). Finally, the language model is fine-tuned with reinforcement learning using the learned reward model. The models will be evaluated using syntactic tasks (e.g., Warstadt, Parrish, Liu, et al. 2020; Huebner, Sulem, Cynthia, et al. 2021): We will measure whether the RL-based fine-tuning improves the models' performance on a range of syntax evaluation metrics when compared to models that only underwent a supervised training phase. Observing such an improvement would provide strong evidence for the hypothesis that Communicative Feedback signals are indeed useful for learning the grammar of one's mother tongue.

A second approach would be to follow an analogous procedure to train and fine-tune *multimodal* models that learn from paired auditory and visual input (e.g., Sullivan, Mei, Perfors, et al. 2022). Such input more closely resembles the multimodal learning environment of a child: Instead of learning from transcriptions of child-directed speech, children are actually exposed to spoken language, grounded in the physical world and their experiences. In addition to the aforementioned syntactic tasks, such models can be evaluated using tasks that measure the acquisition of grounded semantics (e.g., Chrupała, À. Kádár, and Alishahi 2015; Nikolaus and Fourtassi 2021a; Nikolaus, Alishahi, and Chrupała 2022).

By modeling the learning process based on such realistic input data we can start

developing more well-informed theories about the actual processes underlying the acquisition of language.

Part V.

Bibliography

Bibliography

Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, et al. “Bootstrapping Language Acquisition”. In: *Cognition* 164 (July 2017), pp. 116–143. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2017.02.009 (cit. on pp. 25, 93, 108).

Mary S. Ainsworth and John Bowlby. “An Ethological Approach to Personality Development”. In: *American Psychologist* 46.4 (1991), pp. 333–341. ISSN: 1935-990X. DOI: 10.1037/0003-066X.46.4.333 (cit. on p. 31).

Nameera Akhtar, Frances Dunham, and Philip J. Dunham. “Directive Interactions and Early Vocabulary Development: The Role of Joint Attentional Focus*”. In: *Journal of Child Language* 18.1 (Feb. 1991), pp. 41–49. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000900013283 (cit. on p. 30).

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, et al. “The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena”. In: *Language Resources and Evaluation* 41 (2007), pp. 273–287 (cit. on p. 127).

Peter Anderson, Xiaodong He, Chris Buehler, et al. “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6077–6086 (cit. on p. 114).

Dina Anselmi, Michael Tomasello, and Mary Acunzo. “Young Children’s Responses to Neutral and Specific Contingent Queries*”. In: *Journal of Child Language* 13.1 (Feb. 1986), pp. 135–144. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000900000349 (cit. on p. 29).

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, et al. “VQA: Visual Question Answering”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2425–2433 (cit. on p. 96).

Giovanna Axia and Maria Rosa Baroni. “Linguistic Politeness at Different Age Levels”. In: *Child Development* (1985), pp. 918–927 (cit. on p. 56).

Mireille Babineau, Naomi Havron, Isabelle Dautriche, et al. “Learning to Predict and Predicting to Learn: Before and beyond the Syntactic Bootstrapper”. In: *Language Acquisition* 0.0 (June 2022), pp. 1–24. ISSN: 1048-9223. DOI: 10.1080/10489223.2022.2078211 (cit. on pp. 25, 26).

Adrian Bangerter and Herbert H. Clark. “Navigating Joint Projects with Dialogue”. In: *Cognitive Science* 27.2 (2003), pp. 195–225. DOI: 10.1207/s15516709cog2702_3 (cit. on p. 28).

Michelle E. Barton and Michael Tomasello. "The Rest of the Family: The Role of Fathers and Siblings in Early Language Development". In: *Input and Interaction in Language Acquisition*. New York, NY, US: Cambridge University Press, 1994, pp. 109–134. DOI: 10.1017/CBO9780511620690.007 (cit. on p. 128).

Elizabeth Bates. *The Emergence of Symbols: Cognition and Communication in Infancy*. Academic Press, 1979. ISBN: 978-1-4832-6730-2 (cit. on p. 22).

Elizabeth Bates, Luigia Camaioni, and Virginia Volterra. "The Acquisition of Performatives Prior to Speech". In: *Merrill-Palmer Quarterly* 21.3 (1975), pp. 205–226. ISSN: 1535-0266 (cit. on pp. 20, 25).

Elizabeth Bates, Virginia Marchman, Donna Thal, et al. "Developmental and Stylistic Variation in the Composition of Early Vocabulary". In: *Journal of Child Language* 21.1 (1994), pp. 85–123. DOI: 10.1017/S0305000900008680 (cit. on pp. 105, 121).

Janet Bavelas, Jennifer Gerwing, and Sara Healing. "Doing Mutual Understanding. Calibrating with Micro-Sequences in Face-to-Face Dialogue". In: *Journal of Pragmatics* 121 (2017), pp. 91–112. ISSN: 0378-2166. DOI: 10.1016/j.pragma.2017.09.006 (cit. on p. 79).

Elika Bergelson and Richard N. Aslin. "Nature and Origins of the Lexicon in 6-Mo-Olds". In: *Proceedings of the National Academy of Sciences* 114.49 (2017), pp. 12916–12921. DOI: 10.1073/pnas.1712966114 (cit. on p. 90).

Elika Bergelson and Daniel Swingley. "At 6–9 Months, Human Infants Know the Meanings of Many Common Nouns". In: *Proceedings of the National Academy of Sciences* 109.9 (2012), pp. 3253–3258. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1113380109 (cit. on pp. 94, 95, 98, 115, 119).

Claire Bergey, Zoe Marshall, Simon DeDeo, et al. "Learning Communicative Acts in Children's Conversations: A Hidden Topic Markov Model Analysis of the CHILDES Corpora". In: *Topics in Cognitive Science* 14.2 (2022), pp. 388–399 (cit. on p. 57).

Kathleen Bloom. "Distinguishing between Social Reinforcement and Social Elicitation". In: *Journal of Experimental Child Psychology* 38.1 (1984), pp. 93–102. ISSN: 0022-0965. DOI: 10.1016/0022-0965(84)90020-1 (cit. on pp. 30, 60).

Kathleen Bloom. "Quality of Adult Vocalizations Affects the Quality of Infant Vocalizations*". In: *Journal of Child Language* 15.3 (Oct. 1988), pp. 469–480. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000900012502 (cit. on pp. 30, 60, 64, 111).

Kathleen Bloom, Ann Russell, and Karen Wassenberg. "Turn Taking Affects the Quality of Infant Vocalizations*". In: *Journal of Child Language* 14.2 (June 1987), pp. 211–227. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000900012897 (cit. on p. 30).

Lois Bloom and Margaret Lahey. *Language Development and Language Disorders*. 1978 (cit. on pp. 25, 38, 55, 57).

Kübra Bodur, Mitja Nikolaus, Fatima Kassim, et al. "ChiCo: A Multimodal Corpus for the Study of Child Conversation". In: *Proceedings of the 23rd International Workshop on Corpora and Tools for Social Skills Annotation (ICMI)*. Montreal QC Canada: ACM, Oct. 2021, pp. 158–163. ISBN: 978-1-4503-8471-1. DOI: 10.1145/3461615.3485399 (cit. on pp. 58, 127).

Kübra Bodur, Mitja Nikolaus, Laurent Prévot, et al. "Backchannel Behavior in Child-Caregiver Video Calls". In: *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. 2022. DOI: 10.31234/osf.io/cvnqm (cit. on p. 28).

Kübra Bodur, Mitja Nikolaus, Laurent Prévot, et al. "Using Video Calls to Study Children's Conversational Development: The Case of Backchannel Signaling". In: *Frontiers in Computer Science* (2023) (cit. on pp. 90, 127).

John N. Bohannon and Laura B. Stanowicz. "The Issue of Negative Evidence: Adult Responses to Children's Language Errors". In: *Developmental Psychology* 24 (1988), pp. 684–689. ISSN: 1939-0599. DOI: 10.1037/0012-1649.24.5.684 (cit. on pp. 79, 88, 89).

Manuel Bohn and Michael C. Frank. "The Pervasive Role of Pragmatics in Early Language". In: *Annual Review of Developmental Psychology* 1.1 (2019), pp. 223–249. DOI: 10.1146/annurev-devpsych-121318-085037 (cit. on pp. 20, 25, 62, 78).

Marc H. Bornstein, Catherine S. Tamis-LeMonda, Joseph Tal, et al. "Maternal Responsiveness to Infants in Three Societies: The United States, France, and Japan". In: *Child Development* 63.4 (1992), pp. 808–821. ISSN: 1467-8624. DOI: 10.1111/j.1467-8624.1992.tb01663.x (cit. on pp. 27, 64).

Francesca M. Bosco, Monica Bucciarelli, and Bruno G. Bara. "Recognition and Repair of Communicative Failures: A Developmental Perspective". In: *Journal of Pragmatics*. Focus-on Issue: The Pragmatics of Failure and Success 38.9 (Sept. 2006), pp. 1398–1429. ISSN: 0378-2166. DOI: 10.1016/j.pragma.2005.06.011 (cit. on p. 29).

Nadège Bourvis, Magi Singer, Catherine Saint Georges, et al. "Pre-Linguistic Infants Employ Complex Communicative Loops to Engage Mothers in Social Exchanges and Repair Interaction Ruptures". In: *Royal Society Open Science* 5.1 (2018), p. 170274. DOI: 10.1098/rsos.170274 (cit. on p. 30).

John Bowlby. *Attachment and Loss: Attachment*. Basic Books, 1969 (cit. on p. 31).

Mika Braginsky, Daniel Yurovsky, Virginia A. Marchman, et al. "From Uh-Oh to Tomorrow: Predicting Age of Acquisition for Early Words across Languages." In: *CogSci*. 2016 (cit. on p. 42).

Martin DS Braine. "On Two Types of Models of the Internalization of Grammars". In: *The ontogenesis of grammar* 1971 (1971). Ed. by D. I. Slobin, pp. 153–186 (cit. on p. 21).

S. R. K. Branavan, Harr Chen, Luke S. Zettlemoyer, et al. "Reinforcement Learning for Mapping Instructions to Actions". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - ACL-IJCNLP '09*. Vol. 1. Suntec, Singapore: Association for Computational Linguistics, 2009, p. 82. ISBN: 978-1-932432-45-9. DOI: 10.3115/1687878.1687892 (cit. on p. 129).

Bonnie Brinton, Martin Fujiki, Diane Frome Loeb, et al. "Development of Conversational Repair Strategies in Response to Requests for Clarification". In: *Journal of Speech, Language, and Hearing Research* 29.1 (Mar. 1986), pp. 75–81. DOI: 10.1044/jshr.2901.75 (cit. on p. 29).

Roger Brown. *Words and Things*. 1st Edition. Free Press, Nov. 1968. ISBN: 978-0-02-904810-8 (cit. on p. 112).

Roger Brown. *A First Language: The Early Stages*. Harvard University Press, 1973. ISBN: 978-0-674-73246-9 (cit. on pp. 63, 110).

Roger Brown and Camille Hanlon. "Derivational Complexity and Order of Acquisition in Child Speech". In: *Cognition and the development of language* (1970). Ed. by John R. Hayes (cit. on pp. 21, 78, 128).

Jerome Bruner. "Child's Talk: Learning to Use Language". In: *Child Language Teaching and Therapy* 1.1 (1985), pp. 111–114. ISSN: 0265-6590, 1477-0865. DOI: 10.1177/026565908500100113 (cit. on pp. 20, 62, 78).

Lawrence J. Brunner. "Smiles Can Be Back Channels". In: *Journal of Personality and Social Psychology* 37 (1979), pp. 728–734. ISSN: 1939-1315. DOI: 10.1037/0022-3514.37.5.728 (cit. on p. 127).

Thea Cameron-Faulkner. "The Development of Speech Acts". In: *Pragmatic development in first language acquisition* (2014), pp. 37–52 (cit. on pp. 38, 58).

Thea Cameron-Faulkner and Tina Hickey. "Form and Function in Irish Child Directed Speech". In: (2011) (cit. on p. 46).

Ana M. Carmiol, Danielle Matthews, and Odir A. Rodríguez-Villagra. "How Children Learn to Produce Appropriate Referring Expressions in Narratives: The Role of Clarification Requests and Modeling". In: *Journal of Child Language* 45.3 (May 2018), pp. 736–752. ISSN: 0305-0009, 1469-7602. DOI: 10.1017/S0305000917000381 (cit. on p. 29).

Malinda Carpenter, Katherine Nagell, Michael 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* 63.4 (1998), pp. i–174. ISSN: 0037-976X. DOI: 10.2307/1166214. JSTOR: 1166214 (cit. on p. 30).

Marisa Casillas, Penelope Brown, and Stephen C. Levinson. "Early Language Experience in a Tseltal Mayan Village". In: *Child Development* 91.5 (2020), pp. 1819–1835. ISSN: 1467-8624. DOI: 10.1111/cdev.13349 (cit. on pp. 27, 128).

Marisa Casillas and Elma Hilbrink. “Communicative Act Development”. In: *Developmental and clinical pragmatics* 13 (2020), p. 61 (cit. on pp. 38, 55, 111).

Alessandra Cervone and Giuseppe Riccardi. “Is This Dialogue Coherent? Learning from Dialogue Acts and Entities”. In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, July 2020, pp. 162–174 (cit. on p. 126).

Alessandra Cervone, Evgeny Stepanov, and Giuseppe Riccardi. “Coherence Models for Dialogue”. In: *Interspeech 2018*. ISCA, Sept. 2018, pp. 1011–1015. DOI: 10.21437/Interspeech.2018-2446 (cit. on p. 126).

Jane B. Childers, Julie Vaughan, and Donald A. Burquest. “Joint Attention and Word Learning in Ngas-speaking Toddlers in Nigeria”. In: *Journal of Child Language* 34.2 (2007), pp. 199–225. DOI: 10.1017/S0305000906007835 (cit. on p. 111).

Michelle M. Chouinard and Eve V. Clark. “Adult Reformulations of Child Errors as Negative Evidence”. In: *Journal of Child Language* 30.3 (2003), pp. 637–669. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000903005701 (cit. on pp. 21, 27, 63, 78, 110).

Anne Christophe, Séverine Millotte, Savita Bernal, et al. “Bootstrapping Lexical and Syntactic Acquisition”. In: *Language and Speech* 51.1-2 (Mar. 2008), pp. 61–75. ISSN: 0023-8309. DOI: 10.1177/00238309080510010501 (cit. on p. 25).

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. “Representations of Language in a Model of Visually Grounded Speech Signal”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 613–622. DOI: 10.18653/v1/P17-1057 (cit. on pp. 97, 107, 113).

Grzegorz Chrupała, Àkos Kádár, and Afra Alishahi. “Learning Language through Pictures”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, 2015, pp. 112–118. DOI: 10.3115/v1/P15-2019 (cit. on pp. 95–97, 110, 114, 131).

Andy Clark. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, Oct. 2015. ISBN: 978-0-19-021702-0 (cit. on p. 26).

E V Clark and B F Hecht. “Comprehension, Production, and Language Acquisition”. In: *Annual Review of Psychology* 34.1 (1983), pp. 325–349. ISSN: 0066-4308, 1545-2085. DOI: 10.1146/annurev.ps.34.020183.001545 (cit. on p. 122).

Eve V Clark. *First Language Acquisition*. Cambridge University Press, 2016 (cit. on pp. 20–22, 25).

Eve V. Clark. “Conversation and Language Acquisition: A Pragmatic Approach”. In: *Language Learning and Development* 14.3 (2018), pp. 170–185. ISSN: 1547-5441. DOI: 10.1080/15475441.2017.1340843 (cit. on pp. 20, 22, 25, 55, 62, 78, 110, 111, 118).

Eve V. Clark. "Conversational Repair and the Acquisition of Language". In: *Discourse Processes* 57.5-6 (2020), pp. 441–459. ISSN: 0163-853X. DOI: 10.1080/0163853X.2020.1719795 (cit. on pp. 29, 63, 64, 78, 88, 111, 118).

Eve V. Clark and Marie-Catherine de Marneffe. "Constructing Verb Paradigms in French: Adult Construals and Emerging Grammatical Contrasts". In: *Morphology* 22.1 (Feb. 2012), pp. 89–120. ISSN: 1871-5656. DOI: 10.1007/s11525-011-9193-6 (cit. on p. 29).

Herbert H. Clark. *Using Language*. Cambridge University Press, 1996. ISBN: 978-1-316-58260-2 (cit. on pp. 21–23, 27, 28, 64, 73, 78, 111).

Herbert H. Clark and Edward F. Schaefer. "Contributing to Discourse". In: *Cognitive Science* 13.2 (Apr. 1989), pp. 259–294. ISSN: 0364-0213. DOI: 10.1016/0364-0213(89)90008-6 (cit. on pp. 22, 23).

Michael Cogswell, Jiasen Lu, Stefan Lee, et al. *Emergence of Compositional Language with Deep Generational Transmission*. 2020. DOI: 10.48550/arXiv.1904.09067. arXiv: 1904.09067 (cit. on p. 129).

Juliette Corrin. "Maternal Repair Initiation at MLU Stage I: The Developmental Power of 'Hm?'" In: *First Language* 30.3-4 (Aug. 2010), pp. 312–328. ISSN: 0142-7237. DOI: 10.1177/0142723710370526 (cit. on p. 29).

Christopher Martin Mikkelsen Cox, Riccardo Fusaroli, Tamar Keren-Portnoy, et al. *Infant Development as Uncertainty Reduction: Bayesian Insights on Phonological Acquisitio*. 2020. DOI: 10.31234/osf.io/ny6vj (cit. on p. 26).

Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, et al. "Child-Directed Speech Is Infrequent in a Forager-Farmer Population: A Time Allocation Study". In: *Child Development* 90.3 (2019), pp. 759–773. ISSN: 1467-8624. DOI: 10.1111/cdev.12974 (cit. on p. 27).

Alejandrina Cristia, Lucas Gautheron, and Heidi Colleran. "Vocal Input and Output among Infants in a Multilingual Context: Evidence from Long-Form Recordings in Vanuatu". In: *Developmental Science* (2023). DOI: 10.31234/osf.io/bqya7 (cit. on p. 128).

Pino Cutrone. "A Case Study Examining Backchannels in Conversations between Japanese–British Dyads". In: *Multilingua - Journal of Cross-Cultural and Interlanguage Communication* 24.3 (2005), pp. 237–274. ISSN: 0167-8507, 1613-3684. DOI: 10.1515/mult.2005.24.3.237 (cit. on pp. 27, 90, 126).

Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. "Co-Evolution of Language and Agents in Referential Games". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 2993–3004. DOI: 10.18653/v1/2021.eacl-main.260 (cit. on p. 129).

M. J. Demetras, Kathryn Nolan Post, and Catherine E. Snow. “Feedback to First Language Learners: The Role of Repetitions and Clarification Questions*”. In: *Journal of Child Language* 13.2 (1986), pp. 275–292. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000900008059 (cit. on pp. 28, 29, 78, 79, 88, 89).

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. “Word-Minimality, Epenthesis and Coda Licensing in the Early Acquisition of English”. In: *Language and Speech* 49.2 (2006), pp. 137–173. ISSN: 0023-8309. DOI: 10.1177/00238309060490020201 (cit. on p. 81).

Joseph Denby and Daniel Yurovsky. “Parents’ Linguistic Alignment Predicts Children’s Language Development”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. 2019, p. 6 (cit. on p. 32).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423 (cit. on p. 42).

Christina Dideriksen, Morten H. Christiansen, Mark Dingemanse, et al. *Language Specific Constraints on Conversation: Evidence from Danish and Norwegian*. Preprint. PsyArXiv, Apr. 2022. DOI: 10.31234/osf.io/t3s6c (cit. on p. 128).

Christina Dideriksen, Morten H. Christiansen, Kristian Tylén, et al. *Quantifying the Interplay of Conversational Devices in Building Mutual Understanding*. Preprint. PsyArXiv, Oct. 2020. DOI: 10.31234/osf.io/a5r74 (cit. on p. 128).

Christina Dideriksen, Riccardo Fusaroli, Kristian Tylén, et al. *Contextualizing Conversational Strategies: Backchannel, Repair and Linguistic Alignment in Spontaneous and Task-Oriented Conversations*. Preprint. PsyArXiv, May 2019. DOI: 10.31234/osf.io/fd8y9 (cit. on p. 126).

Mark Dingemanse and N. J. Enfield. “Other-Initiated Repair across Languages: Towards a Typology of Conversational Structures”. In: *Open Linguistics* 1.1 (2015). ISSN: 2300-9969. DOI: 10.2478/opli-2014-0007 (cit. on pp. 81, 82).

Mark Dingemanse and Andreas Liesenfeld. “From Text to Talk: Harnessing Conversational Corpora for Humane and Diversity-Aware Language Technology”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022 (cit. on p. 126).

Mark Dingemanse, Seán G. Roberts, Julija Baranova, et al. “Universal Principles in the Repair of Communication Problems”. In: *PLOS ONE* 10.9 (2015), e0136100. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0136100 (cit. on pp. 27, 64, 90).

Allen T. Dittmann. “Developmental Factors in Conversational Behavior”. In: *Journal of Communication* 22.4 (Dec. 1972), pp. 404–423. ISSN: 0021-9916. DOI: 10.1111/j.1460-2466.1972.tb00165.x (cit. on p. 28).

Allen T. Dittmann and Lynn G. Llewellyn. “Relationship between Vocalizations and Head Nods as Listener Responses”. In: *Journal of Personality and Social Psychology* 9 (1968), pp. 79–84. ISSN: 1939-1315. DOI: 10.1037/h0025722 (cit. on p. 127).

Ed Donnellan, Colin Bannard, Michelle L. McGillion, et al. “Infants’ Intentionally Communicative Vocalizations Elicit Responses from Caregivers and Are the Best Predictors of the Transition to Language: A Longitudinal Investigation of Infants’ Vocalizations, Gestures and Word Production”. In: *Developmental Science* 23.1 (2020), e12843. ISSN: 1467-7687. DOI: 10.1111/desc.12843 (cit. on p. 30).

Judy Dunn and Carol Kendrick. “The Speech of Two- and Three-Year-Olds to Infant Siblings: ‘Baby Talk’ and the Context of Communication*”. In: *Journal of Child Language* 9.3 (Oct. 1982), pp. 579–595. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S030500090000492X (cit. on p. 128).

Emmanuel Dupoux. “Cognitive Science in the Era of Artificial Intelligence: A Roadmap for Reverse-Engineering the Infant Language-Learner”. In: *Cognition* 173 (Apr. 2018), pp. 43–59. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2017.11.008 (cit. on p. 25).

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, et al. “Evaluating Coherence in Dialogue Systems Using Entailment”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3806–3812. DOI: 10.18653/v1/N19-1381 (cit. on p. 126).

Fartash Faghri. “VSE++: Improving Visual-Semantic Embeddings with Hard Negatives”. In: *British Machine Vision Conference 2018, {BMVC}*. 2018 (cit. on pp. 97, 98).

Michael J. Farrar. “Negative Evidence and Grammatical Morpheme Acquisition”. In: *Developmental Psychology* 28.1 (1992), pp. 90–98. ISSN: 1939-0599. DOI: 10.1037/0012-1649.28.1.90 (cit. on pp. 21, 78).

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. “A Probabilistic Computational Model of Cross-Situational Word Learning”. In: *Cognitive Science* 34.6 (2010), pp. 1017–1063. ISSN: 1551-6709. DOI: 10.1111/j.1551-6709.2010.01104.x (cit. on pp. 93, 96, 108).

Naomi H. Feldman, Thomas L. Griffiths, Sharon Goldwater, et al. “A Role for the Developing Lexicon in Phonetic Category Acquisition”. In: *Psychological Review* 120.4 (2013), pp. 751–778. ISSN: 1939-1471. DOI: 10.1037/a0034245 (cit. on p. 25).

Anne Fernald. “Intonation and Communicative Intent in Mothers’ Speech to Infants: Is the Melody the Message?” In: *Child development* (1989), pp. 1497–1510 (cit. on p. 57).

Raquel Fernandez and Robert M Grimm. “Quantifying Categorical and Conceptual Convergence in Child-Adult Dialogue”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36. 2014, p. 6 (cit. on p. 31).

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. "Classifying Non-Sentential Utterances in Dialogue: A Machine Learning Approach". In: *Computational Linguistics* 33.3 (2007), pp. 397–427. ISSN: 0891-2017, 1530-9312. DOI: 10.1162/coli.2007.33.3.397 (cit. on p. 82).

Cynthia Fisher and Lila R. Gleitman. *Language Acquisition*. John Wiley & Sons Inc, 2002 (cit. on p. 110).

Michael A. Forrester. "The Emergence of Self-Repair: A Case Study of One Child During the Early Preschool Years". In: *Research on Language and Social Interaction* 41.1 (Mar. 2008), pp. 99–128. ISSN: 0835-1813. DOI: 10.1080/08351810701691206 (cit. on p. 30).

Michael A. Forrester and Sarah M. Cherington. "The Development of Other-Related Conversational Skills: A Case Study of Conversational Repair during the Early Years". In: *First Language* 29.2 (May 2009), pp. 166–191. ISSN: 0142-7237. DOI: 10.1177/0142723708094452 (cit. on p. 29).

Abdellah Fourtassi, Sophie Regan, and Michael C. Frank. "Continuous Developmental Change Explains Discontinuities in Word Learning". In: *Developmental Science* 24.2 (2020). ISSN: 1363-755X, 1467-7687. DOI: 10.1111/desc.13018 (cit. on p. 25).

Ruthe Foushee, Mahesh Srinivasan, and Fei Xu. "Active Learning in Language Development". In: *Current Directions in Psychological Science* (2022). DOI: 10.31234/osf.io/26aer (cit. on p. 27).

Michael C. Frank, Mika Braginsky, Daniel Yurovsky, et al. *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press, 2021 (cit. on pp. 52, 105, 106, 110, 119, 121).

Michael C. Frank, Noah D. Goodman, and Joshua B. Tenenbaum. "Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning". In: *Psychological science* 20.5 (2009), pp. 578–585 (cit. on pp. 93, 108).

Karl Friston. "The Free-Energy Principle: A Rough Guide to the Brain?" In: *Trends in Cognitive Sciences* 13.7 (July 2009), pp. 293–301. ISSN: 1364-6613. DOI: 10.1016/j.tics.2009.04.005 (cit. on p. 26).

Riccardo Fusaroli, Joanna Rączaszek-Leonardi, and Kristian Tylén. "Dialog as Interpersonal Synergy". In: *New Ideas in Psychology* 32 (Jan. 2014), pp. 147–157. ISSN: 0732-118X. DOI: 10.1016/j.newideapsych.2013.03.005 (cit. on p. 126).

Riccardo Fusaroli, Kristian Tylén, Katrine Garly, et al. "Measures and Mechanisms of Common Ground: Backchannels, Conversational Repair, and Interactive Alignment in Free and Task-Oriented Social Interactions". In: *Proceedings for the Annual Meeting of the Cognitive Science Society*. 2017 (cit. on p. 81).

Riccardo Fusaroli, Ethan Weed, Deborah Fein, et al. *Caregiver Linguistic Alignment to Autistic and Typically Developing Children*. 2021. DOI: 10.31234/osf.io/ysjec (cit. on p. 32).

Lukas Galke, Yoav Ram, and Limor Raviv. "Emergent Communication for Understanding Human Language Evolution: What's Missing?" In: *EmeCom (ICLR 2022)*. 2022. arXiv: 2204.10590 (cit. on p. 129).

Tanya M. Gallagher. "Revision Behaviors in the Speech of Normal Children Developing Language". In: *Journal of Speech and Hearing Research* 20.2 (June 1977), pp. 303–318. DOI: 10.1044/jshr.2002.303 (cit. on pp. 29, 64).

Tanya M. Gallagher. "Contingent Query Sequences within Adult–Child Discourse*". In: *Journal of Child Language* 8.1 (Feb. 1981), pp. 51–62. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000900003007 (cit. on p. 29).

Lieke Gelderloos, Alireza Mahmoudi Kamelabad, and Afra Alishahi. "Active Word Learning through Self-supervision". In: *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society*. 2020 (cit. on pp. 27, 122).

Dedre Gentner. *Why Nouns Are Learned before Verbs: Linguistic Relativity Versus Natural Partitioning*. Tech. rep. 1982 (cit. on pp. 105, 121).

Yael Gertner and Cynthia Fisher. "Predicted Errors in Children's Early Sentence Comprehension". In: *Cognition* 124.1 (2012), pp. 85–94 (cit. on pp. 94, 95, 98, 106, 115).

Yael Gertner, Cynthia Fisher, and Julie Eisengart. "Learning Words and Rules: Abstract Knowledge of Word Order in Early Sentence Comprehension". In: *Psychological Science* 17.8 (Aug. 2006), pp. 684–691. ISSN: 0956-7976, 1467-9280. DOI: 10.1111/j.1467-9280.2006.01767.x (cit. on p. 121).

E Mark Gold. "Language Identification in the Limit". In: *Information and Control* 10.5 (May 1967), pp. 447–474. ISSN: 0019-9958. DOI: 10.1016/S0019-9958(67)91165-5 (cit. on p. 21).

Michael H. Goldstein, Andrew P. King, and Meredith J. West. "Social Interaction Shapes Babbling: Testing Parallels between Birdsong and Speech". In: *Proceedings of the National Academy of Sciences* 100.13 (2003), pp. 8030–8035. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1332441100 (cit. on pp. 30, 60, 64, 78, 111).

Michael H. Goldstein and Jennifer A. Schwade. "Social Feedback to Infants' Babbling Facilitates Rapid Phonological Learning". In: *Psychological Science* 19.5 (2008), pp. 515–523. ISSN: 0956-7976. DOI: 10.1111/j.1467-9280.2008.02117.x (cit. on pp. 30, 111).

Dan Goldwasser and Dan Roth. "Learning from Natural Instructions". In: *Machine Learning* 94.2 (Feb. 2014), pp. 205–232. ISSN: 1573-0565. DOI: 10.1007/s10994-013-5407-y (cit. on p. 129).

Roberta M. Golinkoff, Kathy Hirsh-Pasek, Leslie M. Bailey, et al. "Young Children and Adults Use Lexical Principles to Learn New Nouns". In: *Developmental Psychology* 28 (1992), pp. 99–108. ISSN: 1939-0599. DOI: 10.1037/0012-1649.28.1.99 (cit. on p. 125).

Roberta Michnick Golinkoff. "'I Beg Your Pardon?': The Preverbal Negotiation of Failed Messages*". In: *Journal of Child Language* 13.3 (Oct. 1986), pp. 455–476. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000900006826 (cit. on pp. 22, 29).

Roberta Michnick Golinkoff, Kathryn Hirsh-Pasek, Kathleen M. Cauley, et al. "The Eyes Have It: Lexical and Syntactic Comprehension in a New Paradigm*". In: *Journal of Child Language* 14.1 (Feb. 1987), pp. 23–45. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S030500090001271X (cit. on p. 121).

Roberta Michnick Golinkoff, Weiyi Ma, Lulu Song, et al. "Twenty-Five Years Using the Intermodal Preferential Looking Paradigm to Study Language Acquisition: What Have We Learned?" In: *Perspectives on Psychological Science* 8.3 (2013), pp. 316–339 (cit. on pp. 110, 121).

Judith C. Goodman, Philip S. Dale, and Ping Li. "Does Frequency Count? Parental Input and the Acquisition of Vocabulary". In: *Journal of child language* 35.3 (2008), pp. 515–531 (cit. on p. 42).

Noah Goodman, Joshua Tenenbaum, and Michael Black. "A Bayesian Framework for Cross-Situational Word-Learning". In: *Advances in neural information processing systems* (2007) (cit. on p. 96).

Yash Goyal, Tejas Khot, Douglas Summers-Stay, et al. "Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6904–6913 (cit. on p. 106).

H. P. Grice. *Logic and Conversation*. Brill, Dec. 1975. Chap. Speech Acts, pp. 41–58. ISBN: 978-90-04-36881-1. DOI: 10.1163/9789004368811_003 (cit. on p. 38).

Julie Gros-Louis, Meredith J. West, and Andrew P. King. "Maternal Responsiveness and the Development of Directed Vocalizing in Social Interactions". In: *Infancy* 19.4 (2014), pp. 385–408. ISSN: 1532-7078. DOI: 10.1111/infa.12054 (cit. on p. 30).

Gerlind Grosse, Tanya Behne, Malinda Carpenter, et al. "Infants Communicate in Order to Be Understood". In: *Developmental Psychology* 46.6 (Nov. 2010), pp. 1710–1722. ISSN: 1939-0599. DOI: 10.1037/a0020727 (cit. on p. 31).

M.A.K. Halliday. "Learning How to Mean". In: *Foundations of Language Development*. Elsevier, 1975, pp. 239–265. ISBN: 978-0-12-443701-2. DOI: 10.1016/B978-0-12-443701-2.50025-1 (cit. on pp. 20, 22, 25).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778 (cit. on pp. 98, 114).

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. "The Weirdest People in the World?" In: *Behavioral and Brain Sciences* 33.2-3 (June 2010), pp. 61–83. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X0999152X (cit. on p. 127).

Lucille J. Hess and Judith R. Johnston. "Acquisition of Back Channel Listener Responses to Adequate Messages". In: *Discourse Processes* 11.3 (July 1988), pp. 319–335. ISSN: 0163-853X. DOI: 10.1080/01638538809544706 (cit. on p. 28).

Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, et al. "Evaluating Coherence in Open Domain Conversational Systems". In: *Interspeech* (2014) (cit. on p. 126).

Felix Hill, Stephen Clark, Karl Moritz Hermann, et al. *Understanding Early Word Learning in Situated Artificial Agents*. Oct. 2019. DOI: 10.48550/arXiv.1710.09867. arXiv: 1710.09867 (cit. on p. 129).

Felix Hill, Stephen Clark, Karl Moritz Hermann, et al. "Simulating Early Word Learning in Situated Connectionist Agents". In: *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*. 2020 (cit. on p. 130).

Felix Hill, Olivier Tieleman, Tamara von Glehn, et al. "Grounded Language Learning Fast and Slow". In: *Proceedings of the International Conference on Learning Representations*. 2020. arXiv: 2009.01719 (cit. on p. 129).

Sarah Hiller and Raquel Fernandez. "A Data-driven Investigation of Corrective Feedback on Subject Omission Errors in First Language Acquisition". In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 105–114. DOI: 10.18653/v1/K16-1011 (cit. on pp. 21, 27, 63, 78, 80, 110).

Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735 (cit. on pp. 98, 114).

Micah Hodosh, Peter Young, and Julia Hockenmaier. "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics". In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 853–899 (cit. on pp. 96, 97, 113).

Erika Hoff. "The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech". In: *Child Development* 74.5 (2003), pp. 1368–1378. ISSN: 1467-8624. DOI: 10.1111/1467-8624.00612 (cit. on p. 31).

Erika Hoff-Ginsberg. "Topic Relations in Mother-Child Conversation". In: *First Language* 7.20 (1987), pp. 145–158. ISSN: 0142-7237. DOI: 10.1177/014272378700702006 (cit. on pp. 31, 89).

Jessica S. Horst and Larissa K. Samuelson. "Fast Mapping but Poor Retention by 24-Month-Old Infants". In: *Infancy* 13.2 (2008), pp. 128–157. ISSN: 1532-7078. DOI: 10.1080/15250000701795598 (cit. on p. 125).

Chiung-chih Huang. "Parental Other-Repetition in Mandarin Parent–Child Interaction". In: *Journal of Pragmatics* 43.12 (Sept. 2011), pp. 3028–3048. ISSN: 0378-2166. DOI: 10.1016/j.pragma.2011.05.014 (cit. on p. 28).

Drew A. Hudson and Christopher D. Manning. "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6700–6709 (cit. on p. 96).

Philip A. Huebner, Elior Sulem, Fisher Cynthia, et al. "BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language". In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, Nov. 2021, pp. 624–646. DOI: 10.18653/v1/2021.conll-1.49 (cit. on p. 131).

Gail Jefferson. "Side Sequences". In: *Studies in social interaction* (1972), pp. 294–338 (cit. on p. 81).

George Kachergis, Virginia A. Marchman, and Michael C. Frank. "Toward a "Standard Model" of Early Language Learning". In: *Current Directions in Psychological Science* (2021), p. 09637214211057836. DOI: 10.1177/09637214211057836 (cit. on pp. 93, 108).

Ákos Kádár, Afra Alishahi, and Grzegorz Chrupa\la. "Learning Word Meanings from Images of Natural Scenes". In: *Traitement Automatique des Langues* 55.3 (2015) (cit. on p. 96).

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. "Representation of Linguistic Form and Function in Recurrent Neural Networks". In: *Computational Linguistics* 43.4 (Dec. 2017), pp. 761–780. ISSN: 0891-2017. DOI: 10.1162/COLI_a_00300 (cit. on p. 96).

Andrej Karpathy and Li Fei-Fei. "Deep Visual-Semantic Alignments for Generating Image Descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3128–3137 (cit. on p. 97).

Kobin H. Kendrick. "Other-Initiated Repair in English". In: *Open Linguistics* 1.1 (2015). ISSN: 2300-9969. DOI: 10.2478/opli-2014-0009 (cit. on p. 82).

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. "Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network". In: *Proceedings of Coling 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 2012–2021 (cit. on p. 40).

Khazar Khorrami and Okko Räsänen. "Can Phones, Syllables, and Words Emerge as Side-Products of Cross-Situational Audiovisual Learning? - A Computational Investigation". In: *Language Development Research* 1.1 (2021). DOI: 10.34842/W3VW-S845 (cit. on pp. 93, 97, 107, 108, 114).

Celeste Kidd, Steven T. Piantadosi, and Richard N. Aslin. "The Goldilocks Effect: Human Infants Allocate Attention to Visual Sequences That Are Neither Too Simple Nor Too Complex". In: *PLOS ONE* 7.5 (May 2012), e36399. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0036399 (cit. on p. 27).

Simon Kirby and James R. Hurford. "The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model". In: *Simulating the Evolution of Language*. Ed. by Angelo Cangelosi and Domenico Parisi. London: Springer, 2002, pp. 121–147. ISBN: 978-1-4471-0663-0. DOI: 10.1007/978-1-4471-0663-0_6 (cit. on p. 129).

Patricia K. Kuhl. "Is Speech Learning 'Gated' by the Social Brain?" In: *Developmental Science* 10.1 (2007), pp. 110–120. ISSN: 1467-7687. DOI: 10.1111/j.1467-7687.2007.00572.x (cit. on pp. 20, 62).

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, et al. "Dialogue Act Sequence Labeling Using Hierarchical Encoder with Crf". In: *Proceedings of the Aaai Conference on Artificial Intelligence*. Vol. 32. 2018 (cit. on pp. 40, 41).

Barbara Landau and Lila R. Gleitman. *Language and Experience: Evidence from the Blind Child*. Harvard University Press, 1985. ISBN: 978-0-674-03989-6 (cit. on p. 25).

Angeliki Lazaridou and Marco Baroni. "Emergent Multi-Agent Communication in the Deep Learning Era". In: *arXiv:2006.02419 [cs]* (July 2020). arXiv: 2006.02419 [cs] (cit. on p. 129).

Angeliki Lazaridou, Grzegorz Chrupa\la, Raquel Fernández, et al. "Multimodal Semantic Learning from Child-Directed Input". In: *Knight K, Nenkova A, Rambow O, Editors. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016 Jun 12-17; San Diegio, California. Stroudsburg (PA): Association for Computational Linguistics; 2016. p. 387–92*. ACL (Association for Computational Linguistics), 2016 (cit. on p. 96).

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. "Multi-Agent Cooperation and the Emergence of (Natural) Language". In: *Proceedings of the 5th International Conference on Learning Representations*. 2017 (cit. on p. 129).

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. "Multi-Agent Communication Meets Natural Language: Synergies between Functional and Structural Language Learning". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 7663–7674. DOI: 10.18653/v1/2020.acl-main.685 (cit. on pp. 121, 130).

David Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969 (cit. on pp. 22, 129).

Fushan Li and Michael Bowling. "Ease-of-Teaching and Language Structure from Emergent Communication". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019 (cit. on p. 129).

Andreas Liesenfeld and Mark Dingemanse. "Bottom-up Discovery of Structure and Variation in Response Tokens ('Backchannels') across Diverse Languages". In: *Proceedings of Interspeech*. Praeger, 2022. DOI: 10.31234/osf.io/w8hpy (cit. on pp. 27, 90, 126).

Elena Lieven, Dorothé Salomo, and Michael Tomasello. "Two-Year-Old Children's Production of Multiword Utterances: A Usage-Based Analysis". In: *Cognitive Linguistics* 20.3 (Aug. 2009), pp. 481–507. ISSN: 1613-3641. DOI: 10.1515/COGL.2009.022 (cit. on p. 65).

Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. "Microsoft COCO: Common Objects in Context". In: *Computer Vision – ECCV 2014*. Vol. 8693. Cham: Springer International Publishing, 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48 (cit. on pp. 97, 107, 113).

Lukas D. Lopez, Eric A. Walle, Gina M. Pretzer, et al. "Adult Responses to Infant Prelinguistic Vocalizations Are Associated with Infant Vocabulary: A Home Observation Study". In: *PLOS ONE* 15.11 (2020), e0242232. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0242232 (cit. on pp. 31, 78).

Ryan Lowe, Abhinav Gupta, Jakob Foerster, et al. "On the Interaction between Supervision and Self-Play in Emergent Communication". In: *Proceedings of the International Conference on Learning Representations*. 2020 (cit. on pp. 121, 130).

Yuchen Lu, Soumye Singhal, Florian Strub, et al. "Countering Language Drift with Seeded Iterated Learning". In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 6437–6447 (cit. on p. 129).

Lyle Lustigman and Eve V. Clark. "Exposure and Feedback in Language Acquisition: Adult Construals of Children's Early Verb-Form Use in Hebrew". In: *Journal of Child Language* 46.2 (2019), pp. 241–264. ISSN: 0305-0009, 1469-7602. DOI: 10.1017/S0305000918000405 (cit. on pp. 29, 90).

Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk. Transcription Format and Programs*. Vol. 1. Psychology Press, 2000 (cit. on p. 41).

Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Third. New York: Psychology Press, 2014. ISBN: 978-1-315-80567-2. DOI: 10.4324/9781315805672 (cit. on pp. 39, 60, 65, 79, 80, 122, 131).

Brian MacWhinney. *Tools for Analyzing Talk Part 1: The CHAT Transcription Format*. 2017. DOI: 10.21415/3MHN-0Z89 (cit. on pp. 66, 80).

Gary F. Marcus. "Negative Evidence in Language Acquisition". In: *Cognition* 46.1 (1993), pp. 53–85. ISSN: 0010-0277. DOI: 10.1016/0010-0277(93)90022-N (cit. on pp. 21, 29, 78, 79, 88, 89, 128).

G. Markova and M. Legerstee. "Contingency, Imitation, and Affect Sharing: Foundations of Infants' Social Awareness". In: *Developmental Psychology* (2006) (cit. on pp. 30, 111).

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, 1982. ISBN: 978-0-262-28898-9 (cit. on pp. 106, 122).

Lillian R. Masek, Brianna T. M. McMillan, Sarah J. Paterson, et al. “Where Language Meets Attention: How Contingent Interactions Promote Learning”. In: *Developmental Review* 60 (June 2021), p. 100961. ISSN: 0273-2297. DOI: 10.1016/j.dr.2021.100961 (cit. on p. 30).

Danielle Matthews. *Pragmatic Development in First Language Acquisition*. Trends in Language Acquisition Research volume 10. John Benjamins Publishing Company, 2014. ISBN: 978-90-272-3480-3 (cit. on pp. 20, 55).

Senko K. Maynard. “Conversation Management in Contrast: Listener Response in Japanese and American English”. In: *Journal of Pragmatics*. Special Issue: ’Selected Papers from The International Pragmatics Conference, Antwerp, 17-22 August, 1987’ 14.3 (1990), pp. 397–412. ISSN: 0378-2166. DOI: 10.1016/0378-2166(90)90097-W (cit. on pp. 27, 90, 126).

Michelle L. McGillion, Jane S. Herbert, Julian M. Pine, et al. “Supporting Early Vocabulary Development: What Sort of Responsiveness Matters?” In: *IEEE Transactions on Autonomous Mental Development* 5.3 (Sept. 2013), pp. 240–248. ISSN: 1943-0612. DOI: 10.1109/TAMD.2013.2275949 (cit. on pp. 30, 63, 66).

Danny Merkx and Stefan L. Frank. “Learning Semantic Sentence Representations from Visually Grounded Language without Lexical Knowledge”. In: *Natural Language Engineering* 25.4 (July 2019), pp. 451–466. ISSN: 1351-3249, 1469-8110. DOI: 10.1017/S1351324919000196 (cit. on p. 96).

Judi Mesman, Tessa Minter, Andrei Angnged, et al. “Universality Without Uniformity: A Culturally Inclusive Approach to Sensitive Responsiveness in Infant Caregiving”. In: *Child Development* 89.3 (2018), pp. 837–850. ISSN: 1467-8624. DOI: 10.1111/cdev.12795 (cit. on p. 111).

Thomas Misiek, Benoit Favre, and Abdellah Fourtassi. “Development of Multi-level Linguistic Alignment in Child-adult Conversations”. In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Online: Association for Computational Linguistics, Nov. 2020, pp. 54–58. DOI: 10.18653/v1/2020.cmcl-1.7 (cit. on p. 31).

Dipendra Misra, John Langford, and Yoav Artzi. “Mapping Instructions and Visual Observations to Actions with Reinforcement Learning”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 1004–1015. DOI: 10.18653/v1/D17-1106 (cit. on p. 129).

Igor Mordatch and Pieter Abbeel. “Emergence of Grounded Compositional Language in Multi-Agent Populations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018 (cit. on p. 129).

James L. Morgan, Katherine M. Bonamo, and Lisa L. Travis. “Negative Evidence on Negative Evidence”. In: *Developmental Psychology* 31 (1995), pp. 180–197. ISSN: 1939-0599. DOI: 10.1037/0012-1649.31.2.180 (cit. on pp. 78, 79).

James L. Morgan and Lisa L. Travis. “Limits on Negative Information in Language Input*”. In: *Journal of Child Language* 16.3 (1989), pp. 531–552. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000900010709 (cit. on pp. 79, 88).

Aliyah Morgenstern, Marie Leroy-Collombel, and Stéphanie Caët. “Self- and Other-Repairs in Child–Adult Interaction at the Intersection of Pragmatic Abilities and Language Acquisition”. In: *Journal of Pragmatics*. The Pragmatic-Discursive Dimension of Grammar Acquisition 56 (2013), pp. 151–167. ISSN: 0378-2166. DOI: 10.1016/j.pragma.2012.06.017 (cit. on p. 30).

Clément Moulin-Frier, Sao Mai Nguyen, and Pierre-Yves Oudeyer. “Self-Organization of Early Vocal Development in Infants and Machines: The Role of Intrinsic Motivation”. In: *Frontiers in Psychology* (2014). ISSN: 1664-1078 (cit. on p. 27).

Maura Jones Moyle, Susan Ellis Weismer, Julia L. Evans, et al. “Longitudinal Relationships Between Lexical and Grammatical Development in Typical and Late-Talking Children”. In: *Journal of speech, language, and hearing research : JSLHR* 50.2 (2007), pp. 508–528. ISSN: 1092-4388. DOI: 10.1044/1092-4388(2007/035) (cit. on p. 81).

Keith E Nelson, Gaye Carskaddon, and John D Bonvillian. “Syntax Acquisition: Impact of Experimental Variation in Adult Verbal Interaction with the Child”. In: *Child Development* 44.3 (1973), pp. 497–504 (cit. on pp. 21, 78).

Elissa L. Newport, Henry Gleitman, and Lila R. Gleitman. “Mother, I’d Rather Do It Myself: Some Effects and Non-Effects of Maternal Speech Style”. In: *Sentence First, Arguments Afterward*. New York: Oxford University Press, 1977. ISBN: 978-0-19-982809-8. DOI: 10.1093/oso/9780199828098.003.0006 (cit. on p. 28).

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, et al. “Compositional Generalization in Image Captioning”. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (2019), pp. 87–98. DOI: 10.18653/v1/K19-1009 (cit. on p. 171).

Mitja Nikolaus, Afra Alishahi, and Grzegorz Chrupała. “Learning English with Peppa Pig”. In: *Transactions of the Association for Computational Linguistics* (2022) (cit. on pp. 94, 110, 131).

Mitja Nikolaus and Abdellah Fourtassi. “Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in Artificial Neural Networks”. In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Online: Association for Computational Linguistics, June 2021, pp. 200–210. DOI: 10.18653/v1/2021.cmcl-1.24 (cit. on pp. 93, 108, 113–116, 131, 171).

Mitja Nikolaus and Abdellah Fourtassi. “Modeling the Interaction Between Perception-Based and Production-Based Learning in Children’s Early Acquisition of Semantic Knowledge”. In: *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*. Online: Association for Computational Linguistics, 2021, pp. 391–407. DOI: 10.18653/v1/2021.conll-1.31 (cit. on pp. 78, 108).

Mitja Nikolaus and Abdellah Fourtassi. "Communicative Feedback in Language Acquisition". In: *New Ideas in Psychology* (2023). DOI: 10.1016/j.newideapsych.2022.100985 (cit. on pp. 19, 78, 88, 89).

Mitja Nikolaus, Eliot Maes, Jeremy Auguste, et al. "Large-Scale Study of Speech Acts' Development Using Automatic Labelling". In: *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Vienna, Austria, 2021 (cit. on p. 126).

Mitja Nikolaus, Eliot Maes, Jeremy Auguste, et al. "Large-Scale Study of Speech Acts' Development in Early Childhood". In: *Language Development Research* 2.1 (2022). DOI: 10.34842/2022.0532 (cit. on pp. 37, 67, 81, 111).

Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. "Communicative Feedback as a Mechanism Supporting the Production of Intelligible Speech in Early Childhood". In: *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. 2022. DOI: DOI:10.31234/osf.io/sg5mv (cit. on pp. 60, 78–80).

Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. "Communicative Feedback in Response to Children's Grammatical Errors". In: *Proceedings for the 45th Annual Meeting of the Cognitive Science Society*. 2023 (cit. on pp. 76, 126).

Anat Ninio and Catherine Snow. "Language Acquisition through Language Use: The Functional Sources of Children's Early Utterances". In: *Categories and Processes in Language Acquisition* (1988) (cit. on p. 20).

Anat Ninio, Catherine E. Snow, Barbara A. Pan, et al. "Classifying Communicative Acts in Children's Interactions". In: *Journal of Communication Disorders* 27.2 (June 1994), pp. 157–187. ISSN: 0021-9924. DOI: 10.1016/0021-9924(94)90039-6 (cit. on pp. 37, 38, 40, 56, 60, 67, 160).

Claire H. Noble, Caroline F. Rowland, and Julian M. Pine. "Comprehension of Argument Structure and Semantic Roles: Evidence from English-learning Children and the Forced-Choice Pointing Paradigm". In: *Cognitive science* 35.5 (2011), pp. 963–982 (cit. on pp. 94, 95, 98, 106, 115).

Neal R. Norrick. "Functions of Repetition in Conversation". In: *Text - Interdisciplinary Journal for the Study of Discourse* 7.3 (Jan. 1987), pp. 245–264. ISSN: 1860-7349. DOI: 10.1515/text.1.1987.7.3.245 (cit. on p. 23).

Elinor Ochs and Bambi Schieffelin. "Language Acquisition and Socialization". In: *Culture theory: Essays on mind, self, and emotion* (1984), pp. 276–320 (cit. on pp. 22, 27, 90, 128).

Naoaki Okazaki. "Crfsuite: A Fast Implementation of Conditional Random Fields (Crfs)". In: (2007) (cit. on p. 41).

D. Kimbrough Oller. *The Emergence of the Speech Capacity*. The Emergence of the Speech Capacity. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2000, pp. xvii, 428. ISBN: 978-0-8058-2628-9 (cit. on p. 111).

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. “Training Language Models to Follow Instructions with Human Feedback”. In: *Advances in Neural Information Processing Systems* 35 (Dec. 2022), pp. 27730–27744 (cit. on p. 131).

Patrizia Paggio and Costanza Navarretta. “Head Movements, Facial Expressions and Feedback in Conversations: Empirical Evidence from Danish Multimodal Data”. In: *Journal on Multimodal User Interfaces* 7.1 (Mar. 2013), pp. 29–37. ISSN: 1783-8738. DOI: 10.1007/s12193-012-0105-9 (cit. on p. 127).

Kishore Papineni, Salim Roukos, Todd Ward, et al. “BLEU: A Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 311. DOI: 10.3115/1073083.1073135 (cit. on p. 114).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. “Scikit-Learn: Machine Learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830 (cit. on p. 41).

Sharon G. Penner. “Parental Responses to Grammatical and Ungrammatical Child Utterances”. In: *Child Development* 58.2 (1987), pp. 376–384. ISSN: 0009-3920. DOI: 10.2307/1130514 (cit. on p. 89).

Carole Peterson, Beulah Jesso, and Allyssa McCabe. “Encouraging Narratives in Preschoolers: An Intervention Study”. In: *Journal of Child Language* 26.1 (Feb. 1999), pp. 49–67. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000998003651 (cit. on p. 28).

Martin J. Pickering and Simon Garrod. *Understanding Dialogue: Language Use and Social Interaction*. First. Cambridge University Press, Jan. 2021. ISBN: 978-1-108-61072-8. DOI: 10.1017/9781108610728 (cit. on pp. 21–23, 27).

Steven Pinker. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT press, 1989 (cit. on p. 95).

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, et al. “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2641–2649 (cit. on pp. 97, 113).

Claire L. Poulson. “Differential Reinforcement of Other-than-Vocalization as a Control Procedure in the Conditioning of Infant Vocalization Rate”. In: *Journal of Experimental Child Psychology* 36.3 (Dec. 1983), pp. 471–489. ISSN: 0022-0965. DOI: 10.1016/0022-0965(83)90047-4 (cit. on p. 30).

Shannon M. Pruden, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, et al. “The Birth of Words: Ten-Month-Olds Learn Words Through Perceptual Salience”. In: *Child Development* 77.2 (2006), pp. 266–280. ISSN: 1467-8624. DOI: 10.1111/j.1467-8624.2006.00869.x (cit. on p. 125).

Matthew Purver, Julian Hough, and Christine Howes. “Computational Models of Miscommunication Phenomena”. In: *Topics in Cognitive Science* 10.2 (2018), pp. 425–451. ISSN: 1756-8765. DOI: 10.1111/tops.12324 (cit. on p. 82).

Matthew Richard John Purver. "The Theory and Use of Clarification Requests in Dialogue". PhD thesis. 2004 (cit. on pp. 28, 63).

Willard Van Orman Quine. *Word and Object*. MIT press, 1960 (cit. on pp. 95, 110).

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, et al. "Sequence Level Training with Recurrent Neural Networks". In: *4th International Conference on Learning Representations, {ICLR}*. arXiv, 2016. DOI: 10.48550/arXiv.1511.06732. arXiv: 1511.06732 [cs] (cit. on p. 112).

Okko Räsänen and Heikki Rasilo. "A Joint Model of Word Segmentation and Meaning Acquisition through Cross-Situational Learning". In: *Psychological Review* 122.4 (2015), pp. 792–829. ISSN: 1939-1471. DOI: 10.1037/a0039702 (cit. on p. 25).

Andrew Reece, Gus Cooney, Peter Bull, et al. "The CANDOR Corpus: Insights from a Large Multimodal Dataset of Naturalistic Conversation". In: *Science Advances* 9.13 (2023), eadf3197 (cit. on p. 127).

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, et al. "Self-Critical Sequence Training for Image Captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7008–7024 (cit. on p. 112).

Amy L. Richman, Patrice M. Miller, and Robert A. LeVine. "Cultural and Educational Variations in Maternal Responsiveness". In: *Developmental Psychology* 28.4 (1992), pp. 614–621. ISSN: 1939-0599. DOI: 10.1037/0012-1649.28.4.614 (cit. on pp. 27, 64).

Pamela Rosenthal Rollins. "Early Pragmatic Accomplishments and Vocabulary Development in Preschool Children with Autism". In: *American Journal of Speech-Language Pathology* 8.2 (1999), pp. 181–190 (cit. on p. 38).

Pamela Rosenthal Rollins. "Pathways Early Intervention Program for Toddlers with Autism". In: *Journal of Mental Health & Clinical Psychology* 1.1 (2017) (cit. on p. 38).

Caroline F. Rowland and Sarah L. Fletcher. "The Effect of Sampling on Estimates of Lexical Specificity and Error Rates". In: *Journal of Child Language* 33.4 (2006), pp. 859–877. ISSN: 0305-0009, 1469-7602. DOI: 10.1017/S0305000906007537 (cit. on p. 81).

Deb K. Roy and Alex P. Pentland. "Learning Words from Sights and Sounds: A Computational Model". In: *Cognitive Science* 26.1 (2002), pp. 113–146. ISSN: 1551-6709. DOI: 10.1207/s15516709cog2601_4 (cit. on pp. 93, 108).

Olga Russakovsky, Jia Deng, Hao Su, et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (Dec. 2015), pp. 211–252. ISSN: 1573-1405. DOI: 10.1007/s11263-015-0816-y (cit. on pp. 98, 114).

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. "Statistical Learning by 8-Month-Old Infants". In: *Science* 274.5294 (1996), pp. 1926–1928 (cit. on p. 20).

Matthew Saxton. "Negative Evidence and Negative Feedback: Immediate Effects on the Grammaticality of Child Speech". In: *First Language* 20.60 (2000), pp. 221–252. ISSN: 0142-7237. DOI: 10.1177/014272370002006001 (cit. on pp. 21, 27, 63, 78, 79, 89).

Matthew Saxton, Phillip Backley, and Clare Gallaway. “Negative Input for Grammatical Errors: Effects after a Lag of 12 Weeks”. In: *Journal of Child Language* 32.3 (2005), pp. 643–672. ISSN: 0305-0009, 1469-7602. DOI: 10.1017/S0305000905006999 (cit. on p. 78).

Matthew Saxton, Carmel Houston–Price, and Natasha Dawson. “The Prompt Hypothesis: Clarification Requests as Corrective Input for Grammatical Errors”. In: *Applied Psycholinguistics* 26.3 (2005), pp. 393–414. ISSN: 0142-7164, 1469-1817. DOI: 10.1017/S0142716405050228 (cit. on pp. 29, 64, 79, 80, 90, 110).

Graham Schafer. “Infants Can Learn Decontextualized Words Before Their First Birthday”. In: *Child Development* 76.1 (2005), pp. 87–96. ISSN: 1467-8624. DOI: 10.1111/j.1467-8624.2005.00831.x (cit. on p. 125).

Emanuel A. Schegloff. “Discourse as an Interactional Achievement: Some Uses of ‘Uh Huh’and Other Things That Come between Sentences”. In: *Analyzing discourse: Text and talk* 71 (1982), pp. 71–93 (cit. on p. 28).

Emanuel A. Schegloff. *Interaction: The Infrastructure for Social Institutions, the Natural Ecological Niche for Language, and the Arena in Which Culture Is Enacted*. Routledge, 2006, pp. 70–96. ISBN: 978-1-00-313551-7. DOI: 10.4324/9781003135517-4 (cit. on p. 27).

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. “The Preference for Self-Correction in the Organization of Repair in Conversation”. In: *Language* 53.2 (1977), pp. 361–382. ISSN: 0097-8507. DOI: 10.2307/413107 (cit. on pp. 28, 81).

Emanuel A. Schegloff and Harvey Sacks. “Opening up Closings”. In: *Semiotica* 8.4 (1973), pp. 289–327. ISSN: 1613-3692. DOI: 10.1515/semi.1973.8.4.289 (cit. on pp. 41, 126).

John R. Searle. “A Classification of Illocutionary Acts”. In: *Language in Society* 5.1 (Apr. 1976), pp. 1–23. ISSN: 1469-8013, 0047-4045. DOI: 10.1017/S0047404500006837 (cit. on p. 38).

Atsushi Senju and Gergely Csibra. “Gaze Following in Human Infants Depends on Communicative Signals”. In: *Current Biology* 18.9 (May 2008), pp. 668–671. ISSN: 0960-9822. DOI: 10.1016/j.cub.2008.03.059 (cit. on pp. 20, 57, 62).

Ravi Shekhar, Sandro Pezzelle, Aurelie Herbelot, et al. “Vision and Language Integration: Moving beyond Objects”. In: *IWCS*. 2017 (cit. on p. 107).

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, et al. “FOIL It! Find One Mismatch between Image and Language Caption”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 255–265. DOI: 10.18653/v1/P17-1024 (cit. on p. 107).

Jinyu Shi, Yan Gu, and Gabriella Vigliocco. “Prosodic Modulations in Child-directed Language and Their Impact on Word Learning”. In: *Developmental Science* (2022). ISSN: 1363-755X, 1467-7687. DOI: 10.1111/desc.13357 (cit. on pp. 58, 90, 127).

Laura A. Shneidman and Susan Goldin-Meadow. "Language Input and Acquisition in a Mayan Village: How Important Is Directed Speech?" In: *Developmental Science* 15.5 (2012), pp. 659–673. ISSN: 1467-7687. DOI: 10.1111/j.1467-7687.2012.01168.x (cit. on pp. 27, 127).

H.I. Shwe and E.M. Markman. "Young Children's Appreciation of the Mental Impact of Their Communicative Signals". In: *Developmental Psychology* (1997). DOI: 10.1037/0012-1649.33.4.630 (cit. on p. 31).

Linda Smith and Chen Yu. "Infants Rapidly Learn Word-Referent Mappings via Cross-Situational Statistics". In: *Cognition* 106.3 (2008), pp. 1558–1568. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2007.06.010 (cit. on pp. 20, 95, 110).

Catherine E. Snow, Barbara Alexander Pan, Alison Imbens-Bailey, et al. "Learning How to Say What One Means: A Longitudinal Study of Children's Speech Act Use". In: *Social Development* 5.1 (1996), pp. 56–84. ISSN: 1467-9507. DOI: 10.1111/j.1467-9507.1996.tb00072.x (cit. on pp. 25, 38–40, 42, 43, 47–56, 111, 127, 165).

Artem Sokolov, Julia Kreutzer, Christopher Lo, et al. "Learning Structured Predictors from Bandit Feedback for Interactive NLP". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1610–1620. DOI: 10.18653/v1/P16-1152 (cit. on p. 131).

Robert C. Stalnaker. "Assertion". In: *Pragmatics* (Dec. 1978), pp. 315–332. DOI: 10.1163/9789004368873_013 (cit. on pp. 22, 64).

Nisan Stiennon, Long Ouyang, Jeff Wu, et al. "Learning to Summarize from Human Feedback". In: *Advances in Neural Information Processing Systems 33*. 2020, p. 14 (cit. on p. 131).

Andreas Stolcke, Klaus Ries, Noah Coccaro, et al. "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech". In: *Computational Linguistics* 26.3 (Sept. 2000), pp. 339–373. ISSN: 0891-2017. DOI: 10.1162/089120100561737 (cit. on p. 40).

Chehalis M. Strapp. "Mothers', Fathers', and Siblings' Responses to Children's Language Errors: Comparing Sources of Negative Evidence". In: *Journal of Child Language* 26.2 (June 1999), pp. 373–391. ISSN: 1469-7602, 0305-0009. DOI: 10.1017/S0305000999003827 (cit. on pp. 21, 27).

Jessica Sullivan, Michelle Mei, Andrew Perfors, et al. "SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded From the Infant's Perspective". In: *Open Mind* 5 (2022), pp. 20–29. ISSN: 2470-2986. DOI: 10.1162/opmi_a_00039 (cit. on pp. 58, 94, 131).

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning, Second Edition: An Introduction*. MIT Press, Nov. 2018. ISBN: 978-0-262-35270-3 (cit. on pp. 129, 131).

C. S. Tamis-LeMonda, M. H. Bornstein, and L. Baumwell. "Maternal Responsiveness and Children's Achievement of Language Milestones". In: *Child Development* 72.3 (2001), pp. 748–767. ISSN: 0009-3920. DOI: 10.1111/1467-8624.00313 (cit. on p. 30).

Deborah Tannen. *Talking Voices: Repetition, Dialogue and Imagery in Conversational Discourse*. Cambridge University Press, Nov. 1989. ISBN: 978-0-521-37900-7 (cit. on p. 23).

Michael Tomasello. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, 2003. ISBN: 978-0-674-01764-1 (cit. on pp. 20, 25, 62, 78).

Michael Tomasello. *Origins of Human Communication*. MIT Press, Aug. 2010. ISBN: 978-0-262-26120-3 (cit. on p. 107).

Michael Tomasello, Josep Call, and Andrea Gluckman. "Comprehension of Novel Communicative Signs by Apes and Human Children". In: *Child development* (1997), pp. 1067–1080 (cit. on p. 57).

Michael Tomasello, Malinda Carpenter, Josep Call, et al. "Understanding and Sharing Intentions: The Origins of Cultural Cognition". In: *Behavioral and Brain Sciences* 28.5 (Oct. 2005), pp. 675–691. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X05000129 (cit. on pp. 20, 62).

Edward Tronick, Heidelise Als, Lauren Adamson, et al. "The Infant's Response to Entrapment between Contradictory Messages in Face-to-Face Interaction". In: *Journal of the American Academy of Child Psychiatry* 17.1 (Dec. 1978), pp. 1–13. ISSN: 0002-7138. DOI: 10.1016/S0002-7138(09)62273-1 (cit. on pp. 30, 111).

James P. Trujillo, Irina Simanova, Harold Bekkering, et al. "Communicative Intent Modulates Production and Comprehension of Actions and Gestures: A Kinect Study". In: *Cognition* 180 (2018), pp. 38–51 (cit. on p. 57).

Sho Tsuji, Alejandrina Cristia, and Emmanuel Dupoux. "SCALa: A Blueprint for Computational Models of Language Acquisition in Social Context". In: *Cognition*. Special Issue in Honour of Jacques Mehler, Cognition's Founding Editor 213 (Aug. 2021), p. 104779. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2021.104779 (cit. on p. 110).

Sho Tsuji, Nobuyuki Jincho, Reiko Mazuka, et al. "Communicative Cues in the Absence of a Human Interaction Partner Enhance 12-Month-Old Infants' Word Learning". In: *Journal of Experimental Child Psychology* 191 (Mar. 2020), p. 104740. ISSN: 0022-0965. DOI: 10.1016/j.jecp.2019.104740 (cit. on pp. 20, 62).

Katherine E. Twomey and Gert Westermann. "Curiosity-Based Learning in Infants: A Neurocomputational Approach". In: *Developmental Science* 21.4 (July 2018), e12629. ISSN: 1467-7687. DOI: 10.1111/desc.12629 (cit. on p. 27).

Oriol Vinyals, Alexander Toshev, Samy Bengio, et al. "Show and Tell: A Neural Image Caption Generator". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3156–3164 (cit. on p. 114).

Wai Keen Vong and Brenden M. Lake. "Learning Word-Referent Mappings and Concepts from Raw Inputs". In: *arXiv preprint arXiv:2003.05573* (2020). arXiv: 2003.05573 (cit. on pp. 95, 96, 110).

Wai Keen Vong and Brenden M. Lake. "Cross-Situational Word Learning With Multimodal Neural Networks". In: *Cognitive Science* 46.4 (2022), e13122. ISSN: 1551-6709. DOI: 10.1111/cogs.13122 (cit. on p. 110).

Wai Keen Vong, Emin Orhan, and Brenden M. Lake. "Cross-Situational Word Learning from Naturalistic Headcam Data". In: *4th CUNY Conference on Human Sentence Processing*. 2021 (cit. on pp. 97, 110, 114).

Lev S. Vygotsky. *Thought and Language*. MIT press, 1962. ISBN: 978-0-262-30491-7 (cit. on p. 20).

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, et al. "Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 6622–6631. ISBN: 978-1-72813-293-8. DOI: 10.1109/CVPR.2019.00679 (cit. on p. 129).

Anne S. Warlaumont, Jeffrey A. Richards, Jill Gilkerson, et al. "A Social Feedback Loop for Speech Development and Its Reduction in Autism". In: *Psychological Science* 25.7 (2014), pp. 1314–1324. ISSN: 0956-7976. DOI: 10.1177/0956797614531023 (cit. on pp. 30, 33, 60, 63–67, 69, 78, 79, 110, 111, 115, 122).

Alex Warstadt, Alicia Parrish, Haokun Liu, et al. "BLiMP: The Benchmark of Linguistic Minimal Pairs for English". In: *Transactions of the Association for Computational Linguistics* 8 (July 2020), pp. 377–392. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00321 (cit. on p. 131).

G. J. Whitehurst and M. C. Valdez-Menchaca. "What Is the Role of Reinforcement in Early Language Acquisition?" In: *Child Development* 59.2 (Apr. 1988), pp. 430–440. ISSN: 0009-3920 (cit. on p. 32).

M. Jeanne Wilcox and Elizabeth J. Webster. "Early Discourse Behavior: An Analysis of Children's Responses to Listener Feedback". In: *Child Development* 51.4 (1980), pp. 1120–1125. ISSN: 0009-3920. DOI: 10.2307/1129552. JSTOR: 1129552 (cit. on p. 29).

Ronald J. Williams. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". In: *Reinforcement Learning*. Ed. by Richard S. Sutton. The Springer International Series in Engineering and Computer Science. Boston, MA: Springer US, 1992, pp. 5–32. ISBN: 978-1-4615-3618-5. DOI: 10.1007/978-1-4615-3618-5_2 (cit. on p. 114).

Thomas Wolf, Lysandre Debut, Victor Sanh, et al. "Transformers: State-of-the-art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 38–45 (cit. on p. 42).

Amanda L. Woodward and Ellen M. Markman. "Early Word Learning". In: *Handbook of Child Psychology: Volume 2: Cognition, Perception, and Language*. Hoboken, NJ, US: John Wiley & Sons, Inc., 1998, pp. 371–420. ISBN: 978-0-471-05730-7 (cit. on p. 125).

Amanda L. Woodward, Ellen M. Markman, and Colleen M. Fitzsimmons. "Rapid Word Learning in 13- and 18-Month-Olds". In: *Developmental Psychology* 30 (1994), pp. 553–566. ISSN: 1939-0599. DOI: 10.1037/0012-1649.30.4.553 (cit. on p. 125).

Zhen Wu and Julie Gros-Louis. "Infants' Prelinguistic Communicative Acts and Maternal Responses: Relations to Linguistic Development". In: *First Language* 34.1 (Feb. 2014), pp. 72–90. ISSN: 0142-7237. DOI: 10.1177/0142723714521925 (cit. on p. 30).

Kelvin Xu, Jimmy Ba, Ryan Kiros, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, June 2015, pp. 2048–2057 (cit. on p. 114).

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. "Situation Recognition: Visual Semantic Role Labeling for Image Understanding". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 5534–5542. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.597 (cit. on p. 96).

Victor H. Yngve. "On Getting a Word in Edgewise". In: *Chicago Linguistics Society, 6th Meeting, 1970* (1970), pp. 567–578 (cit. on p. 28).

Chen Yu and Dana H. Ballard. "A Unified Model of Early Word Learning: Integrating Statistical and Social Cues". In: *Neurocomputing*. Selected Papers from the 3rd International Conference on Development and Learning (ICDL 2004) 70.13 (Aug. 2007), pp. 2149–2165. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2006.01.034 (cit. on pp. 93, 96, 108).

Daniel Yurovsky. "A Communicative Approach to Early Word Learning". In: *New Ideas in Psychology* 50 (Aug. 2018), pp. 73–79. ISSN: 0732-118X. DOI: 10.1016/j.newideapsych.2017.09.001 (cit. on pp. 20, 22).

Daniel Yurovsky, Gabriel Doyle, and Michael C Frank. "Linguistic Input Is Tuned to Children's Developmental Level". In: *Proceedings for the Annual Meeting of the Cognitive Science Society*. 2016 (cit. on p. 31).

Daniel Yurovsky and Michael C. Frank. "Beyond Naïve Cue Combination: Salience and Social Cues in Early Word Learning". In: *Developmental Science* 20.2 (2017), e12349. ISSN: 1467-7687. DOI: 10.1111/desc.12349 (cit. on pp. 20, 62).

C. L. Zitnick and Devi Parikh. "Bringing Semantics into Focus Using Visual Abstraction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 3009–3016 (cit. on pp. 97, 113).

C. L. Zitnick, Devi Parikh, and Lucy Vanderwende. "Learning the Visual Interpretation of Sentences". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 1681–1688 (cit. on pp. 97, 113).

Part VI.

ANNEXES

A. Appendix A

A.1. INCA-A Tagset

Speech acts of the INCA-A coding scheme (Ninio, C. E. Snow, Barbara A. Pan, et al. 1994) are listed in Table 1.

Table 1. – Speech acts of the INCA-A tagset.

Speech Act	Description
AA	Answer in the affirmative to yes/no question.
AB	Approve of appropriate behavior.
AC	Answer calls/ show attentiveness to communications.
AD	Agree to carry out an act requested or proposed by other.
AL	Agree to do something for the last time.
AN	Answer in the negative to yes/no question
AP	Agree with proposition or proposal expressed by previous speaker
AQ	Aggravated question expression of disapproval by restating a question
CL	Call attention to hearer by name or by substitute exclamations
CM	Commiserate express sympathy for hearer's distress.
CN	Count.
CR	Criticize or point out error in nonverbal act.
CS	Counter-suggestion/ an indirect refusal.
CT	Correct provide correct verbal form in place of erroneous one.
CX	Complete text if so demanded.
DC	Create a new state of affairs by declaration
DP	Declare make-believe reality.
DR	Dare or challenge hearer to perform an action.
DS	Disapprove scold protest disruptive behavior.
DW	Disagree with proposition expressed by previous speaker.
EA	Elicit onomatopoeic or animal sounds.
EC	Elicit completion of word or sentence.
ED	Exclaim in disapproval.
EI	Elicit imitation of word or sentence by modelling or by explicit command
EM	Exclaim in distress pain.
EN	Express positive emotion.
EQ	Eliciting question (e.g. hmm?).
ES	Express surprise.
ET	Express enthusiasm for hearer's performance.
EX	Elicit completion of rote-learned text.
FP	Ask for permission to carry out act.
GI	Give in/ accept other's insistence or refusal.
GR	Give reason/ justify a request for an action refusal or prohibition
MK	Mark occurrence of event (thank greet apologize congratulate etc.).
NA	Intentionally nonsatisfying answer to question
ND	Disagree with a declaration.
OO	Unintelligible vocalization.
PA	Permit hearer to perform act.
PD	Promise.
PF	Prohibit/forbid/protest hearer's performance of an act

PM	Praise for motor acts i.e for nonverbal behavior.
PR	Perform verbal move in game.
QA	Answer a question with a wh-question.
QN	Ask a product-question (wh-question)
RA	Refuse to answer.
RD	Refuse to carry out an act requested or proposed by other.
RP	Request propose or suggest an action for hearer or for hearer and speaker.
RQ	Yes/no question or suggestion about hearer's wishes and intentions
RR	Request to repeat utterance.
RT	Repeat or imitate other's utterance.
SA	Answer a wh-question with a statement.
SC	Complete statement or other utterance in compliance with request.
SI	State intent to carry out act by speaker.
SS	Signal to start performing an act such as running or rolling a ball
ST	Make a declarative statement.
TA	Answer a limited-alternative question.
TD	Threaten to do.
TO	Mark transfer of object to hearer
TQ	Ask a limited-alternative yes/no question.
TX	Read or recite written text aloud.
WD	Warn of danger.
WS	Express a wish.
XA	Exhibit attentiveness to hearer.
YA	Answer a question with a yes/no question.
YD	Agree to a declaration.
YQ	Ask a yes/no question.
YY	Make a word-like utterance without clear function.

A.2. Model Details

A.2.1. Hyperparameters

The models were trained until convergence on a held-out dev set (10% of the training data). A small set of hyperparameter configurations based on best practices were evaluated in preliminary experiments. The configuration listed in Table 2 led to the best results.

The learning rate for training the BERT-based model is substantially lower than for the other model as this model is already pre-trained and we are only fine-tuning it on the task.

A.2.2. Architecture

A high-level overview of the architecture of the hierarchical LSTM+CRF model can be found in Figure 1.

Table 2. – Model hyperparameters

Hierarchical LSTM + CRF	
vocabulary size	1000
word embeddings size	200
word-level LSTM hidden layer size	200
utterance-level LSTM hidden layer size	100
dropout	0.2
optimizer	Adam
initial learning rate	0.0001
+ BERT	
same as above, except for:	
initial learning rate	0.00001

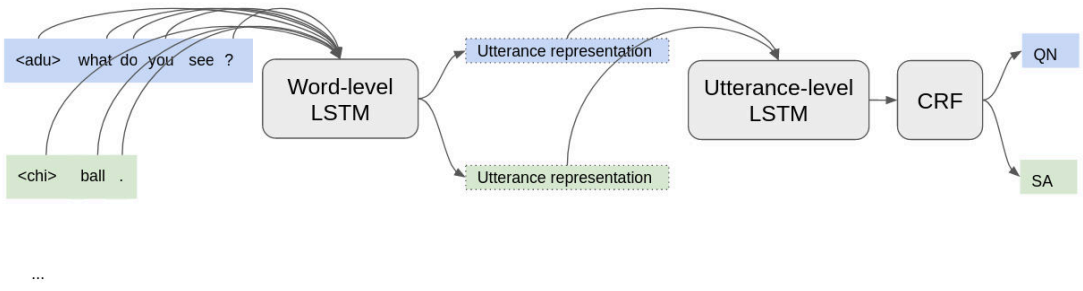

Figure 1. – Architecture of the Hierarchical LSTM + CRF model.

A.3. Error Analysis

Table 3 contains per-label precision, recall, and F1-scores for a model trained on 80% of the New England corpus and tested on the remaining 20%.

Table 3. – Error analysis

	precision	recall	f1-score	support
AA	0.628	0.628	0.628	148
AB	0.690	0.454	0.547	108
AC	0.603	0.527	0.562	245
AD	0.674	0.651	0.662	229
AL	0.000	0.000	0.000	1
AN	0.625	0.571	0.597	35
AP	0.658	0.603	0.629	239
CL	0.800	0.875	0.836	160
CM	0.375	0.231	0.286	13
CN	0.200	0.500	0.286	4
CR	0.000	0.000	0.000	13
CS	0.273	0.086	0.130	35
CT	0.529	0.138	0.220	65
DC	0.750	0.316	0.444	19
DP	0.000	0.000	0.000	8
DS	0.375	0.273	0.316	11
DW	0.633	0.404	0.494	47
EA	0.974	0.884	0.927	43
EC	0.857	0.429	0.571	14
ED	1.000	0.333	0.500	15
EI	0.632	0.800	0.706	15
EM	0.000	0.000	0.000	1
EQ	0.750	0.849	0.796	53
ET	0.739	0.459	0.567	37
EX	0.000	0.000	0.000	1
FP	0.833	0.694	0.758	36
GI	0.375	0.158	0.222	19
GR	0.350	0.226	0.275	31
MK	0.733	0.814	0.772	996
NA	0.000	0.000	0.000	30
ND	0.000	0.000	0.000	1
PA	0.600	0.409	0.486	22
PD	0.800	0.211	0.333	19
PF	0.830	0.702	0.761	272
PM	0.518	0.345	0.414	84
PR	0.769	0.652	0.706	296
QN	0.940	0.958	0.949	1104
RD	0.679	0.494	0.571	77
RP	0.797	0.786	0.791	1689
RQ	0.830	0.848	0.839	506
RR	0.448	0.714	0.550	42
RT	0.467	0.340	0.394	144
SA	0.782	0.662	0.717	417

SC	1.000	0.455	0.625	11
SI	0.551	0.405	0.466	309
SS	0.811	0.664	0.730	116
ST	0.690	0.791	0.737	1620
TA	0.000	0.000	0.000	3
TO	0.333	0.222	0.267	72
TQ	1.000	0.200	0.333	10
TX	0.818	0.863	0.840	73
WD	0.875	0.700	0.778	10
XA	0.671	0.464	0.548	110
YA	0.769	0.408	0.533	49
YD	0.000	0.000	0.000	5
YQ	0.715	0.772	0.742	705
macro avg	0.567	0.446	0.479	10437
weighted avg	0.738	0.725	0.726	10437

A.4. Ages of Acquisition

A.4.1. Regression Plots

The regression plots in Figure 2 and 3 illustrate the proportion of children producing a given speech act (in the case of comprehension, the proportion of contingent responses made by children) across time as well as the best logistic fits used to predict the speech acts' precise age of acquisition. We depict only 6 exemplary speech acts for better readability. The data to create these plots was the original annotation data from C. E. Snow, Barbara Alexander Pan, Imbens-Bailey, et al. (1996).

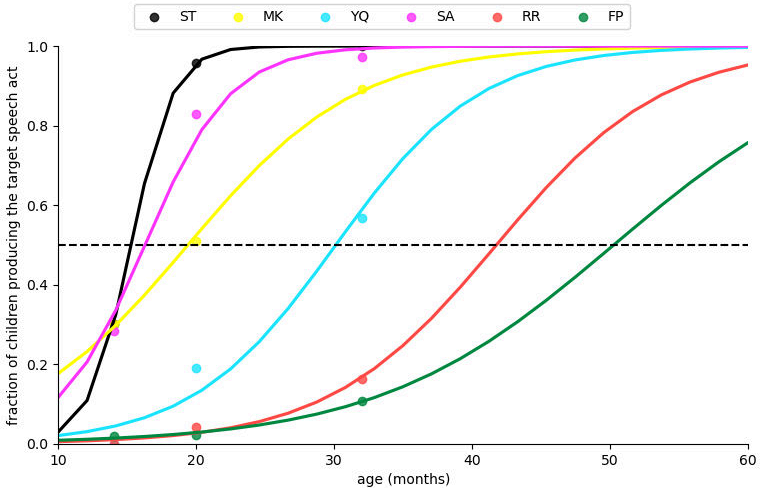

Figure 2. – Regression plot for production.

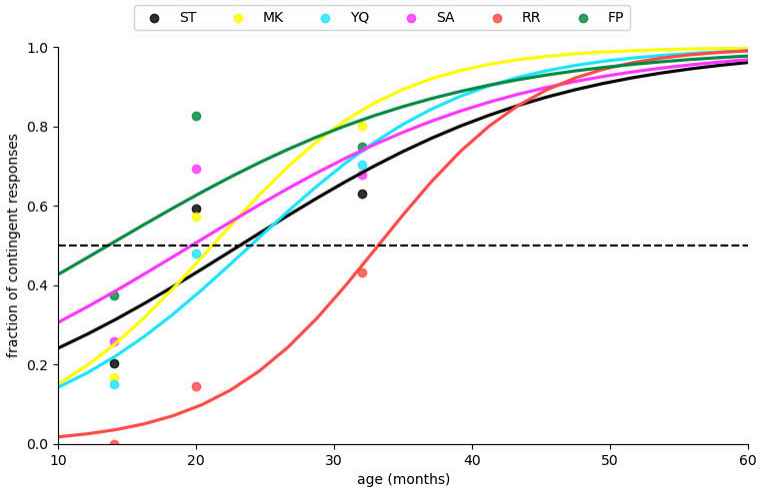

Figure 3. – Regression plot for comprehension.

A.4.2. Predicted Ages of Acquisition

The following tables show the age of acquisition (in months) for speech acts calculated using different data sources ("-" indicates that no age of acquisition could be calculated, i.e. at no observed time the proportion of children producing the speech act surpassed 0.5). We calculated the ages of acquisition in terms of production (Table 4) and comprehension (Table 5).

Table 4. – Predicted ages of acquisition for production.

Speech act	Snow	CRF	CHILDES
AA	20.4	20.4	16.2
AC	30.8	32.9	32.9
AD	20.4	22.5	22.5
AN	35.0	30.8	26.6
AP	39.1	47.5	30.8
CL	41.2	45.4	70.3
CS	99.5	-	39.1
DC	45.4	45.4	53.7
DW	41.2	64.1	35.0
FP	51.6	-	37.1
MK	20.4	18.3	16.2
PA	45.4	45.4	45.4
PF	-	43.3	35.0
PR	28.7	-	-
QN	26.6	24.6	24.6
RD	26.6	24.6	22.5
RP	18.3	20.4	18.3
RR	43.3	39.1	41.2
RT	20.4	20.4	16.2
SA	18.3	16.2	10.0
SC	43.3	53.7	-
SI	22.5	26.6	22.5
ST	16.2	16.2	14.2
TO	-	35.0	37.1
YQ	30.8	28.7	22.5

Table 5. – Predicted ages of acquisition for comprehension.

Speech act	Snow	CRF	CHILDES
AA	26.6	24.6	22.5
AB	24.6	35.0	24.6
AC	24.6	26.6	22.5
AD	20.4	24.6	22.5
AN	-	-	-
AP	14.2	20.4	20.4
AQ	-	-	-

CL	30.8	35.0	30.8
CM	20.4	-	43.3
CN	-	-	-
CR	-	-	-
CS	28.7	30.8	10.0
CT	16.2	16.2	10.0
DC	10.0	-	24.6
DS	-	-	-
DW	30.8	10.0	22.5
EA	10.0	12.1	10.0
EC	-	-	-
EI	10.0	22.5	10.0
EQ	20.4	22.5	18.3
ET	20.4	24.6	26.6
FP	14.2	28.7	10.0
GI	32.9	32.9	49.5
GR	10.0	26.6	20.4
MK	22.5	22.5	20.4
PA	22.5	28.7	30.8
PD	10.0	59.9	10.0
PF	32.9	26.6	30.8
PM	20.4	26.6	26.6
PR	22.5	24.6	24.6
QN	22.5	22.5	10.0
RD	10.0	-	-
RP	32.9	35.0	37.1
RQ	22.5	26.6	28.7
RR	35.0	39.1	99.5
RT	22.5	10.0	10.0
SA	20.4	18.3	22.5
SI	24.6	26.6	16.2
SS	30.8	22.5	30.8
ST	24.6	24.6	18.3
TO	26.6	35.0	24.6
TQ	20.4	12.1	10.0
TX	30.8	28.7	24.6
WD	24.6	-	87.0
XA	24.6	24.6	26.6
YA	26.6	28.7	26.6
YQ	24.6	24.6	26.6

A.4.3. Predicted Ages of Acquisition Including Data of Older Children

Table 6 presents the ages of acquisition in terms of production including data from older children (up to 54 months). We show only speech acts for which the age of acquisition could be calculated, i.e. for which at some age the proportion of children producing the speech act surpassed 0.5 .

Table 6. – Predicted ages of acquisition including older children

Speech act	Age of acquisition
AA	18.3
AC	45.4
AD	32.9
AN	41.2
AP	101.6
AQ	155.7
CL	136.9
CN	95.3
CR	141.1
CS	149.4
DP	107.8
DW	76.6
EA	91.2
EI	164.0
EM	180.6
EQ	93.2
FP	78.7
GR	87.0
MK	16.2
PA	139.0
PD	130.7
PF	84.9
QN	35.0
RD	39.1
RP	10.0
RQ	66.2
RR	66.2
RT	10.0
SA	10.0
SI	20.4
ST	10.0
TA	95.3
TQ	62.0
YA	188.9
YQ	26.6

B. Appendix B

B.1. Model Details

The hyperparameters of the model were chosen on general best-practices and not any further tuned.

Minimum word frequency for vocab	5
Word Embeddings Size	100
Joint Embeddings Size	512
LSTM Hidden Layer Size	512
Optimizer	Adam
Initial Learning Rate	0.0001
Batch size	32
α (margin for loss term)	0.2

C. Appendix C

C.1. Hyperparameters

Model hyperparameters as indicated in Table 7 were chosen based on general best-practices and not any further tuned (except for the frequency of CF updates, see Appendix C.2). During training, we evaluate the model every 100 batches, and stop training if the BLEU score on the held out validation set does not improve for 50 consecutive validations. All models converged within 8 hours when running on a single GPU.

Parameter	**Value**
Minimum word frequency for inclusion vocab	5
Word Embeddings Size	100
LSTM Hidden Layer Size	512
Optimizer	Adam
Optimizer Initial Learning Rate	$1 \cdot 10^{-4}$
Optimizer Initial Learning Rate (Model fine-tuning)	$1 \cdot 10^{-5}$
Dropout	0.2
Batch size	32

Table 7. – Model hyperparameters.

C.2. Varying frequency of CF updates

As the loss terms of the cross-entropy loss used in XSL and the policy gradient loss used in CF can take very different margins, we experiment with different update frequencies of XSL updates with respect one XSL update. An update frequency of 2 indicates that we perform an XSL update every 2 CF updates.

The results as shown in Table 8 show that we obtain the best results (average over all tasks) when performing 1 XSL update every CF update for the model in the `Alt` setup, that is when alternating feedback-based and input-based learning from the start. However, the performance is still worse than for a model trained using `XSL` alone (mainly regarding persons and semantic roles).

For our best performing setup, `XSL+Alt`, we observe a different pattern, displayed in Table 9. In this case it is best to perform an XSL update every 10 CF updates. We hypothesize that this can be explained by the fact that the CF updates are more useful in this setup, as the model has already learned a language model in the first input-based learning phase before starting to produce sentences. In the main text, we report results for both `Alt` and `XSL+Alt` with a frequency of 10 CF updates per XSL update for direct comparison.

	Evaluation task	Frequency of CF updates				
		1	2	5	10	20
Word-level Semantics	Nouns: Persons	0.740	0.660	0.520	0.500	0.480
	Nouns: Animals	0.997	0.978	0.703	0.667	0.500
	Nouns: Objects	0.930	0.858	0.720	0.567	0.497
	Verbs	0.597	0.556	0.542	0.486	0.500
	Adjectives	0.786	0.714	0.643	0.554	0.500
Sentence-level Semantics	Adj-noun dependencies	0.786	0.714	0.643	0.554	0.500
	Verb-noun dependencies	0.565	0.573	0.542	0.510	0.500
	Semantic roles	0.540	0.500	0.480	0.500	0.440
	Average	**0.715**	0.674	0.588	0.537	0.490

Table 8. – Accuracy for all semantic evaluation tasks for varying frequency of CF updates in the `Alt` setup. Note that we only performed one run for each setting, and thus some numbers do not match exactly those in the Table 6.1.

C.3. BLEU Scores

Table 10 shows the BLEU scores for all different learning scenarios. The score was calculated by sampling images from the validation set and comparing generated sentences with the gold sentences. These results are compatible with our observations using the grounded semantics evaluation tasks. Here again `XSL+Alt` performs best.

C.4. Comparison with Nikolaus and Fourtassi (2021)

Our baseline (`XSL`) results differ from the results in Nikolaus and Fourtassi (2021a) for several reasons.

Firstly, their models are trained with a max-margin loss, instead of a cross-entropy objective as we did here. We cannot evaluate our model by directly calculating similarity between images and sentences because it does not learn a multimodal semantic embedding space. Thus, we evaluate it by calculating conditional perplexity for both target and distractor sentences. These factors might explain the drop in performance for some metrics, especially for sentence-level semantics. Future work should investigate how to combine both training objectives (max-margin loss and cross-entropy loss), in order to combine their respective benefits (e.g. Nikolaus, Abdou, Lamm, et al. 2019).

Secondly, we do fine-tune the ResNet of our models, as we observed substantial performance improvements with this change. This might explain the gain in performance for adjectives (the children's emotions), which the model of Nikolaus and Fourtassi (2021a) struggled with (probably due to the inappropriateness of the pre-trained image features, they are largely optimized for recognizing objects in naturalistic scenes, but not clip-art objects).

	Evaluation task	Frequency of CF updates				
		1	2	5	10	20
Word-level Semantics	Nouns: Persons	0.900	0.880	0.880	0.860	0.900
	Nouns: Animals	0.997	0.994	0.997	0.997	0.997
	Nouns: Objects	0.952	0.957	0.954	0.954	0.949
	Verbs	0.722	0.708	0.764	0.778	0.764
	Adjectives	0.750	0.857	0.786	0.839	0.839
Sentence-level Semantics	Adj-noun dependencies	0.646	0.667	0.630	0.594	0.635
	Verb-noun dependencies	0.598	0.593	0.630	0.720	0.708
	Semantic roles	0.620	0.620	0.680	0.660	0.480
	Average	0.773	0.785	0.790	**0.800**	0.784

Table 9. – Accuracy for all semantic evaluation tasks for varying frequency of CF updates in the `XSL+Alt` setup. Note that we only performed one run for each setting, and thus some numbers do not match exactly those in the Table 6.1.

`XSL`	`Alt`	`XSL+CF`	`XSL+Alt`
66.5 ± 0.8	53.9 ± 0.6	70.8 ± 0.2	$\mathbf{72.7} \pm 0.5$

Table 10. – BLEU score on the test set (mean and standard deviation over 5 runs) for different learning setups.

C.5. Analysis of produced sentences

Examples of models' produced sentences (at the end of training) are shown in Table 11.

We further quantitatively compare the produced utterances during the training using `XSL+CF` and `XSL+Alt`. Every 100 batches, we sample sentences from the model for all images in the validation set and analyze these produced sentences for sentence length (Figure 4) as well as occurrences of persons (Figure 6.3) and verbs (Figure 5). There are only 2 persons in the dataset, "jenny" and "mike". We measure occurrence of persons by counting sentences that contain "jenny", but not "mike" (and vice versa). Regarding the verbs, we count occurrences for all verbs that are used in the semantic evaluation tasks.

The examples show that the model produces increasingly short sentences when trained using `XSL+CF`. We also observe a drop in mean sentence length for `XSL+Alt`, but to a substantially smaller degree.

Figure 6.3 shows that the model trained using `XSL+CF` increasingly produces sentences involving "jenny", but decreasingly sentences involving "mike". Thus it might get less feedback on the difference between Jenny and Mike and unlearn this distinc-

XSL+CF	XSL+Alt
jenny is wearing glasses	jenny is crying
an owl is sitting	mike is holding balloons
jenny is holding	mike is kicking the soccer ball
jenny is holding balloons	jenny is holding a ketchup
jenny is flying	jenny is playing in the sandbox
jenny is holding the	jenny has glasses on
jenny is holding	mike is making a pirate
jenny is wearing	jenny is running away from the snake
mike is wearing	the bear is wearing a wizards hat
jenny is angrily	the rain is cooking lightning in the sky

Table 11. – 10 sentences produced by the models for randomly sampled images from the validation set. The model checkpoints used were from the end of training (epoch 19).

tion to some degree. Consequently, it also struggles more to understand semantics roles (distinguishing the persons is necessary to correctly map the semantic roles). For XSL+Alt, the fraction of sentences involving "jenny" and "mike" remains largely constant.

Regarding the presence of verbs, Figure 5 shows a different pattern. While for XSL+Alt the fractions do not vary much, in XSL+CF some verbs are produced increasingly. This might explain the large gain in performance for verbs: The model produces more sentences involving verbs, and thus also receives more valuable feedback to learn meaningful semantic representations.

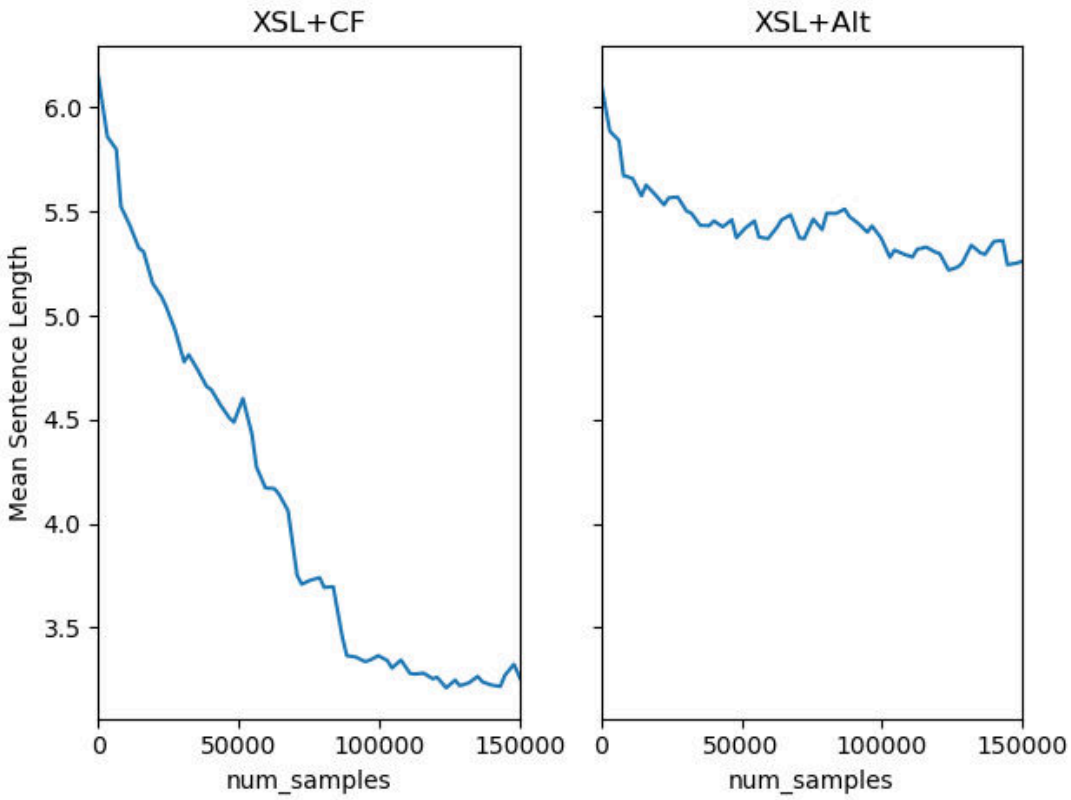

Figure 4. – Comparison of the mean sentence length during training of the `XSL+CF` and `XSL+Alt` training setups. The graphs only display the second training step, not the pre-training using XSL.

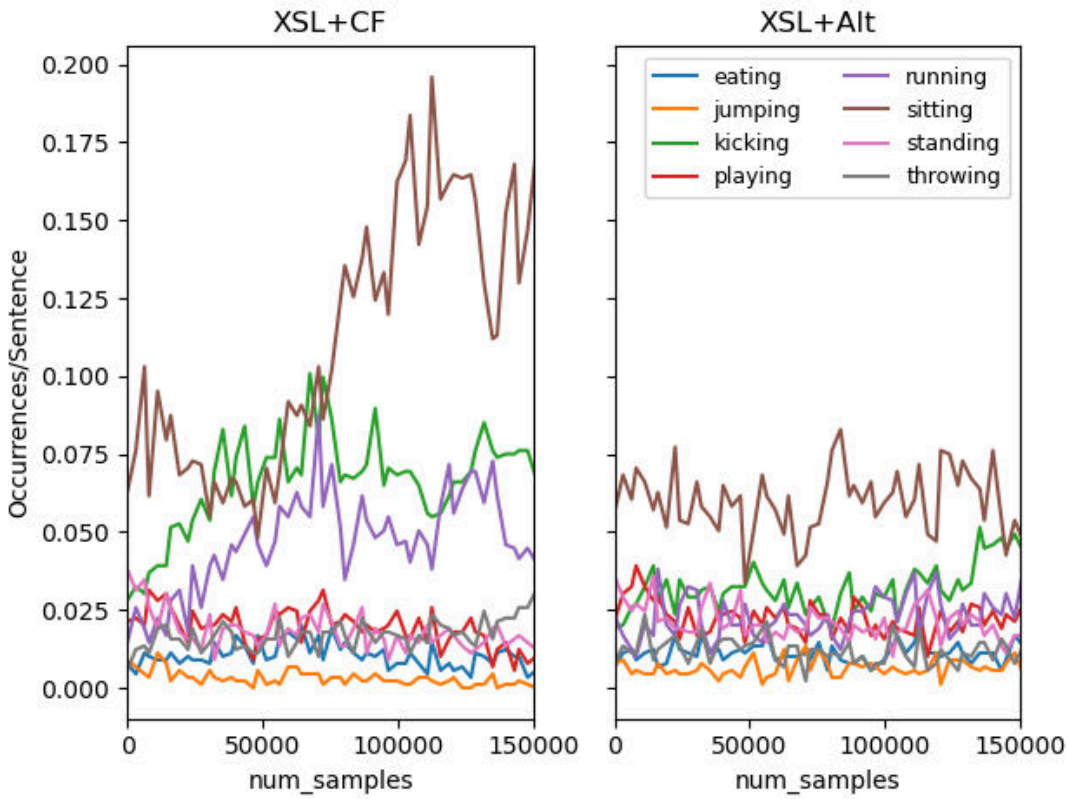

Figure 5. – Comparison of the fraction of occurrences of verbs during training of the `XSL+CF` and `XSL+Alt` training setups. The graphs only display the second training step, not the pre-training using XSL.