Characterizing children's naturalistic narratives via machine learning

Megan Waller
Carnegie Mellon University, USA

Dhruv Nambiar
New York University, USA

Anthony Tomasic
Madeline Elston
Erik Thiessen
Carnegie Mellon University, USA

Abstract: Storytelling is a fundamental human behavior, and narrative skill is predictive of success in many domains, such that the development of narrative skill is an important process in childhood and adolescence. The goal of this research was to create and assess a novel method for identifying narratives produced in conversations between children and adults. To do so, we crafted a coding manual with a set of concrete rules and examples for identifying utterances containing narrative behavior. Then we asked trained coders to apply these rules to a set of transcribed conversations between adults and children, drawn from CHILDES. Our results indicated that coders could apply these rules with high inter-rater reliability. The utterances that coders identified had many of the characteristics of narrative defined by prior empirical literature. Further, we trained a version of a large language model (LLM) (GPT-4o-mini) to apply these rules and found that the model could successfully mimic human judgments. These results suggest that it is possible to automatize judgments of narrative behavior for rapid analysis of transcribed conversations at scale. This provides a novel avenue for investigation of the development of storytelling abilities, one which has the potential to generate new insights about the acquisition and use of narrative skill in social contexts.

Keywords: storytelling; narrative development; large language model; classifier model; machine learning

Corresponding author(s): Erik Thiessen, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. Email: thiessen@andrew.cmu.edu

ORCID ID(s): Megan Waller: https://orcid.org/0000-0002-9442-9996; Dhruv Nambiar https://orcid.org/0009-0003-8047-8900; Anthony Tomasic https://orcid.org/0000-0001-7864-0364; Madeline Elston https://orcid.org/0009-0005-8669-8036; Erik Thiessen https://orcid.org/0000-0002-2563-032X

Citation: Waller, M., Nambiar, D., Tomasic, A., Elston, M. & Thiessen, E. (2026). Characterizing children's naturalistic narratives via machine learning. *Language Development Research, 6*(1), 51–124. http://doi.org/10.34842/ldr2026-946

Introduction

Since before the days of Gilgamesh and Enkidu, stories have been an integral part of human experience. Stories play an important role not only in defining shared cultural histories but also in making sense of daily life and the natural world. This function of narrative is so automatic and pervasive that stories have been proposed as a compelling and plausible candidate for an atomic, indivisible, innate unit of cognition or communication (e.g., Boyd, 2009; Grishakova & Sorokin, 2016; Herman, 2003). One piece of evidence that is consistent with this claim is the observation that individual differences in storytelling abilities (both production and comprehension) are predictive of success in a wide variety of domains (e.g., Babayiğit et al., 2020; Davidson et al., 2017; Gardner-Neblett & Iruka, 2015). Similarly, an array of cognitive deficits ranging from neurodevelopmental disorders to dementia are diagnosed by atypicality in narrative communication. In cognitive science, narrative has been thought to provide a template for representing information that is widely generalizable and generative, operating in a manner consistent with a schema (e.g., Baldassano et al., 2018; Bartlett, 1932; Bower et al., 1979). Though it may be impossible to resolve questions about the innateness of narrative, the plausibility of the hypothesis illustrates the depth of the integration between narrative and the human experience and highlights the power of narrative for making a psychological impact. Whether storytelling is an innate aspect of human cognition, it is also an ability shaped by learning. This can be seen, for example, by the notable differences in storytelling conventions across cultures (e.g., Imada, 2010; McCabe, 1997) Similarly, the profound developmental changes in storytelling ability suggest a role for learning in the acquisition of narrative skills (e.g., Montanari, 2004; Pinto et al., 2018; Zanchi & Zampini, 2020). That said, theoretical accounts of the role (or lack thereof) of learning in narrative development are limited by a paucity of relevant data. Much of what is known about the development of storytelling skills comes from experimental settings. Our goal in this research project is to extend and supplement that knowledge by providing an analysis of storytelling in naturalistic conversation between parents and children. This is especially critical to questions about the role of learning in the development of narrative skills, as conversations (especially with parents) are likely to provide a rich source of information for learning.

Children's storytelling abilities change over the course of childhood. Storytelling is not obviously apparent from a young child's first one-word or two-word utterances but becomes more readily detectable in the third and fourth year of life (e.g., Bruner, 1975; McCabe & Peterson, 1991), and by age five their stories include enough information that they can be understood outside the immediate context (e.g., Beliavsky, 2003). Children's early stories are brief and structured around free association as much (or more) than sequential or causal structure (e.g., Sperry & Sperry, 1996). Among four-year-olds, many stories lack discernible relations between events or

statements and are often simply descriptions of ideas or images (Marjanovic-Umek et al., 2002). With age, their stories show more evidence of being structured in an adult-like fashion with a clear beginning, middle, and end (e.g., Berman et al., 1994; Bliss et al., 1998). By age 6, children's stories are much more likely to be based on temporally or causally ordered sequences, and by age 8 are likely to contain descriptions of character motivation, thematic continuity, and narrative resolution (Hill et al., 2025; Marjanovic-Umek et al., 2002). Though children's stories become more complex with age, their length may not follow a linear pattern. Among English-speaking children, the length of stories appears to show a U-shaped curve. The initial stories told by young children are quite short and grow longer until around 7 or 8 years of age, at which point they plateau and then actually become shorter due to children being able to convey information more concisely (e.g., Esposito et al., 2020). Similarly, older children tell stories that are less likely to revolve around the here and now, and more likely to incorporate abstract ideas (e.g., Severing & Verhoeven, 2001).

The vast majority of evidence about the development of storytelling ability in childhood is drawn from studies in which children are presented with some stimulus that prompts a story, and then to tell that story to an experimenter. A commonly used example of this approach involves presenting children with the wordless picture book "Frog, where are you?" (Mayer, 1969) and asking them to narrate a story based on the 24 pictures in the book (Berman et al., 1994). An alternative approach commonly used is story-retelling (e.g., Gazella & Stockman, 2003), in which the experimenter conveys a story to the child, and then asks the child to repeat the story after some interval. The results from these and other paradigms have been largely (though not entirely; see Pesco & Gagné, 2017) consistent, allowing researchers to make confident claims about developmental milestones and changes across age (e.g., Mar et al., 2021; Merritt & Liles, 1989).

While this approach has been tremendously successful in identifying the general trajectory of storytelling ability across developmental time, it suffers from two notable weaknesses. First, while these studies provide us with a snapshot of performance at specific ages, they provide limited information about how change occurs between ages. This limitation can be overcome to a certain extent by designing careful longitudinal or interventional studies, but this is costly and is most appropriate for questions with specific hypotheses under investigation. Our ability to formulate specific questions, in turn, is hindered by the second major limitation of experimental approaches: a lack of ecological validity that makes it challenging to connect these results to naturalistic storytelling. In all these experimental designs, children are provided with a story, or story scaffolding, before they begin to tell their story; this differs dramatically from a self-generated story that unfolds over the course of a conversation. Low ecological validity is not unique to experiments that investigate storytelling, and experiments with low ecological validity can still be of tremendous value. However, we believe that the ecological validity concern is of special importance in the

study of storytelling, because of the fundamentally communicative and interactive nature of stories.

The claim that stories communicate information is self-evident, but an appreciation of their interactivity may require closer scrutiny. Unlike the fairy tale stories aimed at children, or the novels and short stories that adults read, storytelling in the social environment typically involves give and take between a speaker and conversational partners, rather than a sole narrator regaling an audience with a tale. This give and take, or interactivity, can take many forms. One frequent avenue of interactivity is provided by what linguists refer to as "back-channeling", the optional vocal and non-vocal signals that signal that the listener is engaged in the ongoing discussion (Ward & Tsukahara, 2000). These backchannel signals can also be used to request additional information, emphasize statements, and coordinate turn-taking in conversation (e.g., Clark & Krych, 2004; Clark & Murphy, 1982). These backchannel interactions allow conversational partners to steer the ongoing narrative, even when they are not the primary narrator, and shape the comprehension of listeners (Dideriksen et al., 2023; Tolins & Fox Tree, 2014; 2016). Questions provide another opportunity for interaction between conversational partners to shape ongoing narratives. Questions can be used to signal (lack of) understanding (e.g., Marcus, 1993).

The interactive nature of conversational storytelling provides multiple opportunities for learning. For example, the nature of the responses to a child's statements - such as clarifying questions, reformulations, or suggestions to go on - can provide useful information about whether listeners understand the story, and what kind of story structures they expect (e.g., Pratt et al., 1982). Similarly, the cooperative nature of conversational communications allows adults to scaffold storytelling through questions, prompts, and examples (Vygotsky, 1978). The claim that children learn from this kind of conversational interaction is not especially unexpected; multiple sources of evidence indicate that infants and young children learn about many aspects of language through conversations and interactions with others in their environment (e.g., Elmlinger et al., 2025; Marcus, 1993; Weisleder & Fernald, 2013). However, the experimental paradigms that provide so much information about developmental milestones in storytelling are, by virtue of their elegant experimental control, relatively uninformative about how narrative abilities are learned and changed in the course of conversation. While these experiments often scaffold children's storytelling, they do so in standardized ways that are necessarily artificial and insensitive to an individual child's abilities.

If we wish to learn how conversational interactions shape narrative abilities, we must study conversational interactions. Interactions between individuals are notoriously more difficult to study in experimental paradigms than designs involving a single participant, because the participants influence each other in ways that make it hard to assess causality (e.g., Creswell, 2020; Grossen, 2009; Nesbit & Hadwin, 2006). While

this makes experimental designs challenging to use in this domain, there are often valuable insights to be gained - especially in relatively novel areas of study - by relying on observation instead of manipulation of independent variables (e.g., Lorenz, 1950). There is very little literature assessing the role of conversation in the development of storytelling ability (though see Abbeduto et al., 1995; Burdelski & Fukuda, 2019; Haden et al., 1997). As such, the use of observational techniques - while they are sharply limited in their ability to make causal inferences - can still be very productive in advancing our understanding of these phenomena. However, even observational studies of conversations can be quite challenging, because it is impossible to know in advance when storytelling will occur or the frequency and duration of observation necessary to capture narrative behavior. Therefore, the use of previously recorded conversations offers several advantages for researchers.

One of the premier repositories of recorded and transcribed child conversations is the CHILDES database, which contains over 20 million words (MacWhinney, 1991). As part of the TalkBank project, the CHILDES repository is now available for online access, enabling modern database tools for parsing transcripts (MacWhinney, 2000). This database contains transcripts (and often the original recordings) of conversations between adults and children, from infancy into adolescence, primarily in English but with samples from several other languages. These transcripts arise from a variety of sources, including those that are derived from experimental settings. One relevant example is the "Frogs" corpora, which resulted from recordings of experiments in which children were presented with a picture book containing images of cartoon frogs, and asked to tell a story that connected the pictures (Mayer, 1969). Of particular interest for our purposes, however, are those transcripts that are collected from naturalistic conversational settings, in which a recording device is unobtrusively present as a child interacts with their caregivers at home or other adults as they go about their daily business. These conversations can provide us with an unvarnished look at storytelling "in the wild," and allow us to examine how adults scaffold and support the development of narrative abilities.

However, the use of naturalistic conversations presents an interpretive challenge that is not present in an experiment like the Frogs studies. In an experiment where children are asked to tell a story, most (if not all) of their subsequent utterances can be assumed to be story-related. The story begins when the child starts speaking, and it ends when the child is finished. The beginnings and endings of stories are not so clearly demarcated in conversation, nor does a speaker typically have the opportunity to narrate a story uninterrupted. Moreover, the interlocutors in naturalistic conversation (unlike an experimenter trying to elicit a story) cannot be relied upon to interject story-relevant material, or to help maintain the structure of the story. In many cases, an ongoing narrative can be interrupted by a conversation related to a mundane task like making breakfast or cleaning, only to be resumed minutes later. As such, transcripts of conversations must be laboriously analyzed by human coders, to

determine whether each line of dialogue is (or is not) part of an ongoing story. The daunting nature of the coding task undoubtedly contributes to the dearth of literature on storytelling in naturalistic conversations. That said, we believe that the advantages of developing a corpus of naturalistic narrative is worth the effort, because of the insight it can provide into adult interactions with children's ongoing development of narrative skills.

There are two noteworthy pragmatic advantages to developing a corpus of naturalistic narratives embedded in conversations, in addition to the theoretical insight one might gain into learning. The first of these pragmatic advantages is that such a corpus would enable a novel mode of comparison between narrative language and child language more generally. There are age norms for both general language skills like mean length of utterance (e.g., Rice et al., 2010), as well as age norms for specifically narrative language (e.g., Berman et al., 1994; Moore Channell et al., 2018). However, these norms are necessarily derived from different children performing somewhat artificial tasks. By contrast, comparing between narrative and non-narrative utterances is a within-subjects comparison that assesses performance in the same children in the same setting. In particular, such a dataset can tell us about the proportion of children's linguistic input and output that includes narrative content. Intuitively, the number of stories that children tell, and the amount of time they spend telling stories, changes over developmental time. Currently, there is no dataset that allows us to measure this intuition.

The second pragmatic advantage of developing a corpus of naturalistic storytelling is that modern machine learning tools might be able to model human coding well enough to automate some or all of the necessary coding for the analysis of subsequent transcripts, thus unlocking a much wider range of data for examination. More broadly, a high-quality model of storytelling would enable a suite of new applications based on real-time analysis of storytelling in conversations. Training on only a small subset of the transcripts in CHILDES may enable a machine learning system to train the remaining transcripts at a pace impossible for human coders to match. Labeling a massive array of transcripts with tags indicating the presence of storytelling would allow researchers to investigate many novel questions about the development of storytelling.

Our goal in Study 1 is to test and validate a system for coding the presence of narrative in naturalistic conversation. In particular, we will assess children's mean length of utterance (MLU), and the amount of material children contribute to conversational stories. Prior work indicates that storytelling is associated with more verbal expression, such that children's MLU is longer in stories than out of stories (e.g., Losi et al., 2022; Potratz et al., 2022; Wagner et al., 2000). If our story coding scheme is successful, we should replicate this result in conversational narrative. Relatedly, we make the novel prediction that - due to increases in MLU with age - children should contribute

more overall words to stories as they get older. We will additionally conduct a series of pre-registered analyses to assess developmental differences in conversational narrative behaviors across ages, and differences between narrative and non-narrative utterances. While our sample size (only 60 transcripts) in this study is too small to make broad developmental claims, these analyses serve as a proof of concept and source of hypotheses for subsequent high-powered investigations. In Study 2, we will assess whether this coding scheme can be learned and automated by large language models, which would in turn enable analyses of much larger datasets.

Study 1

To determine whether a transcript of a recorded conversation contains a narrative, one must be able to define the narrative in a way that is precise, valid, and reliable. In terms of precision and validity, we are fortunate to be able to draw from prior literature that defines narrative and studies narrative development (e.g., Esposito et al., 2010; Gazella & Stockman, 2003). While individual researchers often use distinct terminology, a clear consensus emerges that narratives entail some idea of 'plot', in which events are related (often defined in terms of chronological, causal, or thematic relations) and evaluated from a speaker's point of view (e.g., Burdelski, 2019; Crawshaw et al., 2020; Yabe et al., 2018). Drawing from this literature, we defined a narrative statement as a section of speech that contains at least two ***interconnected*** statements, at least one of which is an ***action*** or ***event***. An ***action*** describes an utterance in which a subject is described as performing a behavior, often to achieve an aim. "The little duck goes swimming" would be an action. An ***event*** is an utterance describing how something changed from one state to another. "It began to rain" would be an example of an event. Other statements such as preferences, desires, and general descriptions, were not considered actions or events; however, if these types of sentences are *interconnected* with another action or event, they may be considered part of the narrative. From this general definition of narrative statements, we created a set of rules to identify how they might be instantiated and related in conversational speech (see Method and Appendix A for a complete description of this process). As this is a novel coding scheme, various questions about its reliability and validity must be investigated. Our first goal in Study 1 is to assess the inter-rater reliability of our coding scheme, operationalized in terms of Cohen's Kappa (McHugh, 2012).

Our second goal is to assess construct validity. To do so, we will analyze the utterances that our coding scheme highlights as narrative. If these statements are truly parts of stories, then they should show characteristics that are broadly consistent with what we know about children's storytelling abilities. In particular, we will focus on mean length of utterance (MLU) and overall story length, because both of these characteristics yield clear hypotheses. Prior research indicates both that MLU should be greater in narrative utterances than in non-narrative utterances, and also that there should be notable changes in the length of children's stories across age. Therefore, we will

provide descriptive statistics about mean length of utterance (MLU) and overall story length and assess the extent to which these results are consistent (or not) with narrative development more broadly.

Our third goal was to illustrate the value of this type of naturalistic dataset by conducting some preliminary analyses to demonstrate the value of this dataset, such as assessing the concreteness of children's storytelling or how parents respond to children's story utterances with questions or repetitions. We recognize that this dataset is too small to make generalizing conclusions, but these preliminary results will motivate the effort to continue this methodological development. In Study 2, we will make use of the transcripts labeled in Study 1 and attempt to train a machine learning model to replicate the judgments of human coders. If this attempt is successful, it will then be possible to generate a large enough dataset to assess more generalizable claims about narrative development.

Method

Materials/Transcripts

All transcripts were drawn from the CHILDES database (MacWhinney, 2000). The contexts in which these recordings were made varies widely, from completely naturalistic overheard conversations to structured interviews with researchers. For the purposes of the current study, we selected transcripts that involved conversations between a single child and one or more adults. For discussion of the transformation of 'raw' downloaded CHILDES transcripts into the format seen by our coders, see Appendix B.

Transcript Selection

Sixty (60) transcripts were selected from the English North American and English United Kingdom collections from the CHILDES database: 20 from children with ages 6 years - 6 years 11 months, 20 from children with ages 7 years to 7 years 11 months, and 20 from children with ages 8 years to 9 years 6 months, balanced for gender within each group. These transcripts were selected without fully reading the content of these transcripts but instead were chosen based on whether they were at least 200 lines, contained sufficient speech from the child, and constituted a dialog between a child and an adult rather than utterances only from the child. It is important to note that due to these restrictions, we could not evenly sample across months within each age bracket. As a result, our analyses are better suited to detecting broader age-related differences rather than fine-grained developmental changes that may occur within narrower monthly intervals. Information about which corpora each transcript was pulled from, and a description of the corpora's sampling methods can be found in Appendix C.

From the selected transcripts, a subset of up to 300 lines were selected to be coded by two coders. If the transcript was shorter than 300 lines, the entire transcript was used. There were a total of 16,559 lines across the 60 transcript subsets.

Coding Rules for Labeling Stories

Before reading the content of the 60 transcripts, a coding book had been developed for defining whether a given line would be coded as part of a story ("in") or unrelated ("out") to an ongoing narrative within the conversation. As part of the training process, the two coders read through the manual, and then labeled approximately five CHILDES transcripts that were not part of the preregistered set of 60 transcripts (See Appendix A for coding manual). For these training sets, they compared their judgements and discussed disagreements to ensure understanding of the coding manual. In the manual, the two coders were instructed to identify whether each line of the story was part of a narrative. As stated above, we define a narrative as a section of speech that contains at least two **interconnected** statements, at least one of which is an **action** (a character performing a behavior) or **event** (a change of state**)**. This includes both fictional and nonfictional statements and can be in any tense (past, present, future, conditional, hypothetical, etc.). Narratives also transcended individual speakers, as multiple speakers could contribute to explaining and expanding on a narrative.

Commands, questions, backchanneling, or unrelated interjections are not considered part of the narrative. However, statements of preference, descriptions, or desires could be considered part of a narrative if they were *interconnected* with an action or event. Utterances are considered interconnected if they are linked either temporally, causally, or thematically. Take the example shown in Table 1. Here either line 35 or line 36 on their own would not constitute a narrative, but when connected thematically to the action/event on line 37, both become part of a narrative.

Table 1. ***Excerpt from CHILDES transcript used as training set. The initial question is not coded as part of the story, but the child's response, all these lines would be part of a story due to the action "the boy felled into the wall", and the thematically connected description of characters.***

Utterance #	Story Coding	Transcript
34	out	INV: what's [//] can you tell me a little about that?
35	in	CHI: there's-uh these two Spy Kids .
36	in	CHI: this one who's the girl the girl was pink .
37	in	CHI: and the boy felled into the wall !

Utterances that could not be conclusively categorized as "in" or "out" of an ongoing narrative were marked as "ambiguous" and were not included in the analyses ($n = 14$). We report these utterances in Appendix D. Utterances were tagged as ambiguous if statements were too vague or ungrammatical to fully identify their relevance to the narrative, or sound effects whose meaning or relevance was unclear to the coder were marked as ambiguous and were not considered in the analyses. Additionally, any utterances that were solely marked as xxx (unintelligible) or only included noises like coughs or laughter were removed from analysis (n = 719).

Of the set of 60 transcripts, 18 transcripts were coded by both coders, and these transcripts were spread out across age brackets. Interrater reliability was assessed using Cohen's κ calculated at the utterance level, based on coders' independent classification of each transcript line as in-story, out-of-story, or ambiguous (McHugh, 2012). We report a weighted κ using equal weights (implemented in the kappa 2 function in R). Their agreement was strong (κ = 0.803), and the coders met to resolve any disagreements on those 18 transcripts. The other 42 transcripts were coded only by one coder individually. The labeling results are then integrated into a label file for each transcript (Table 2).

Table 2. ***Human coding label result for a transcript coded by both Coders. If only one coder coded the transcript, only one of the name columns, the in/out/ambiguous column, and the Transcript column were included.***

Utterance #	Coder 1	Coder 2	Agreement	Final Judgement	Transcript
96	out	out	TRUE	out	INV: how about the lake ?
97	out	in	FALSE	in	CHI: yeah that was at Findley .
98	out	out	TRUE	out	INV: hm: ?
99	out	in	FALSE	in	CHI: that was at Findley .
100	out	out	TRUE	out	INV: oh .
101	in	in	TRUE	in	CHI: I went over to watch the boats go out and [/] and the lifeguard caught me .

Validation Analyses Procedures

The preregistration for our analyses can be found here: https://doi.org/10.17605/OSF.IO/CGZS6. This preregistration was submitted after transcript coding was completed but prior to conducting any statistical analyses. The preregistration served to document, prior to analysis, the hypotheses and analytic

approach that were theoretically relevant to our research questions.
For each utterance, we algorithmically coded a number of characteristics for the purposes of illustrating the language patterns of dyadic storytelling. To assess the development of sentence structures, we used the NLTK library parts of speech tagger (https://www.nltk.org/; Bird et al., 2009), which classifies words into the Penn Treebank (Marcus et al., 1993; found here: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html). For each cleaned transcript line, we apply the NLTK parts of speech tagger to each word in the line, generating a count of the total number of coordinating conjunctions (CC tag) and prepositional or subordinating conjunctions tag (IN tag), as well as the number of unique conjunctions within an utterance.

As specified in our preregistration, we initially defined the length of the story as the number of utterances by the same speaker in a row that are coded as "in-story". A story starts when a line is marked as "in" the story and ends when 3 lines are marked as "out" of the story. We assessed whether this heuristic reflects human judgments of story length by asking two coders to label 12 transcripts (separate from the repeated utterance judgments) for the beginning and end of stories. The coder's judgments did not align with the model (*Mean* F1 Story Start = 0.369, *Mean* F1 Story End = 0.312). Therefore, instead of this heuristic, we had those coders label all 60 transcripts for when stories start and end by labeling whether a given line was a beginning of a story, an end of a story, or both. Every line marked as "in-story" is contained within the boundaries of a beginning and an end. Agreement on story boundaries was assessed per utterance, with coders indicating yes/no for each line as a story beginning or ending; κ was computed separately for each (McHugh, 2012). The coders strongly agreed on identifying beginnings and endings on the 20 transcripts they double-coded (beginnings: $\kappa = 0.734$, endings: $\kappa = 0.727$). Similar to the procedures for the in/out story coding, any discrepancies between the two coders were resolved in a meeting to establish final judgments for all transcript lines.

Once the coders began the task of coding the length of the story, we determined that our original decision that we should analyze utterances made by the same speaker only obscures the dyadic nature of storytelling in these naturalistic conversations. Therefore, story length will be analyzed separately from *turn-length,* which specifies how many utterances or words are spoken before a speaker changes. Note that in our preregistration, we also planned to analyze the number of embedded clauses that a speaker produces in the course of storytelling. We attempted to use the NLTK library's chart parsing module to detect sentence phrases but were unable to obtain consistently valid results. We thus abandoned this plan because we could not effectively implement an automated approach to embedded clause detection.

Preliminary Narrative Characteristic Analyses Procedures

To provide further support for the value of this naturalistic dataset, we also planned to conduct preliminary analyses about the 60 transcripts included here. The following variables were extracted from the transcripts themselves to serve these analyses:

For establishing the concreteness of children's storytelling and non-storytelling utterances, we created a dictionary using the dataset provided by Brysbaert et al. (2013). This publication contains a spreadsheet of 37,058 English words and an associated mean concreteness score (measured on the dataset of the paper). The minimum mean concreteness score is 1.04 for the words "eh" and "essentialness" and the maximum mean concreteness score is 5.0 for 280 words including "apple," "goat," and "saxophone". We also use the standard deviation of the concreteness score from this dataset. For each cleaned transcript line, we generate a sum of concreteness scores of the words that occur in the dictionary. We also compute the geometric mean of the concreteness standard deviation scores. The concreteness score and geometric mean score is added to each transcript line in the dataset.

For determining self as subject, for each cleaned transcript line, we check if any of the words in the line are first-person pronouns. If one exists, then that line is marked as one where the self is the subject. The length of utterance was coded by counting the number of words in each line. Questions were identified by checking if the last symbol of the raw transcript line included a question mark.

To detect whether a caregiver repeated a given utterance, we generate a score representing how much of an adult's utterance is a repeat of what the child previously said. To do this, we start with a cleaned transcript line. If the line is spoken by the child, we give it a score of 0, since we are only measuring repeats in adult utterances. If the line is not spoken by the child, then we look back at the 3 most recently spoken lines by the child. We then calculate the proportion of words in the adult's utterance that appear anywhere in the 3 previous child lines. We calculate this proportion for every non-child transcript line. As an informal validation of the selection of a window size of 3 previous child lines, we asked two coders to label a random set of 12 transcripts for whether or not any adult utterance repeats a child's utterance, without providing any specific window size instructions. These judgments strongly agreed with the window size of 3 (*Mean* Accuracy = 94.7%, *Mean* F1 = 0.509). Given the high agreement, rather than use the binary judgements, we decided to use the proportion of repetitions in the prior three lines as a continuous variable to better capture variability in the data.

Statistical Analyses

All analyses were carried out using mixed effect models with the *lme4* and *lmerTest* packages (Bates et al., 2015; Kuznetsova et al., 2017) in R (version 4.2.1, R Core Team, 2022). For the majority of analyses, the model structure involves predicting the

relevant dependent variable (e.g. utterance concreteness score or whether or not the utterance contains first person) from the interaction between the child's age in months (centered on sample mean) and whether or not utterance was part of a story (all categorical predictors are sum-coded as –1 and 1). Some other analyses involve other predictors, such as the speaker, or presence of questions. The details of all model structures and results can be found in Appendix E. All models were initially fit with a random intercept and slope for the relevant categorical variable (such as story coding) by transcript. Following recommendations by Barr et al. (2013), we initially aimed to include the maximal random effect structure and reduced the random effect structure if the models did not converge. In the results below, for clarity, we simply report the significant results of the maximal model that would converge.

Model estimates from linear models are reported as unstandardized regression coefficients (β), which reflect the expected change in the outcome for a one-unit change in the predictor. For generalized (logistic) models, coefficients are exponentiated and reported as odds ratios (ORs) to facilitate interpretation. For all models, we report 95% confidence intervals (CIs). For linear mixed-effects models, we report 95% profile-likelihood CIs. For logistic models, we report Wald CIs due to numerical instability of profile-likelihood intervals in these models. These CIs are also converted to odds ratios for easier interpretation. A 95% CI provides a range of values that are compatible with the observed data under the model; narrower intervals indicate greater precision in estimating the effect size (Greenland et al., 2016). Because the sample size was determined by feasibility constraints rather than an a priori power analysis, the reported 95% confidence intervals are especially informative, as they quantify the precision of our effect size estimates and the degree of uncertainty associated with the present sample. For linear models, CIs that include 0 indicate that the data are compatible with both positive and negative effects. For logistic models, because the CIs have been converted to odds ratios, CIs that include 1 indicate that the data are compatible with both increased and decreased odds. In both cases, intervals that exclude 0 (linear models) or 1 (logistic models) indicate that the estimated effect differs from the null value at the $\alpha = .05$ level. The simple slopes for any significant interactions were investigated using the *emmeans* package (Lenth, 2025).

As suggested by an anonymous reviewer, we also fit Bayesian versions of each model corresponding to our research questions using the *brms* package in R (Bürkner, 2017). Consistent with our preregistration, the frequentist models remain our primary analyses; the Bayesian models were included to evaluate the robustness of the findings under an alternative inferential framework. Bayesian estimation provides several advantages, including full posterior distributions for all parameters, direct probability statements about effects, and partial pooling of random effects that can improve estimation stability (Gelman et al., 1995). Overall, the Bayesian analyses yielded effects that were similar in direction and magnitude to those obtained from the frequentist

models; any notable deviations are described in the manuscript. Priors, model specifications, and results for these analyses can be found in Appendix E.

Results

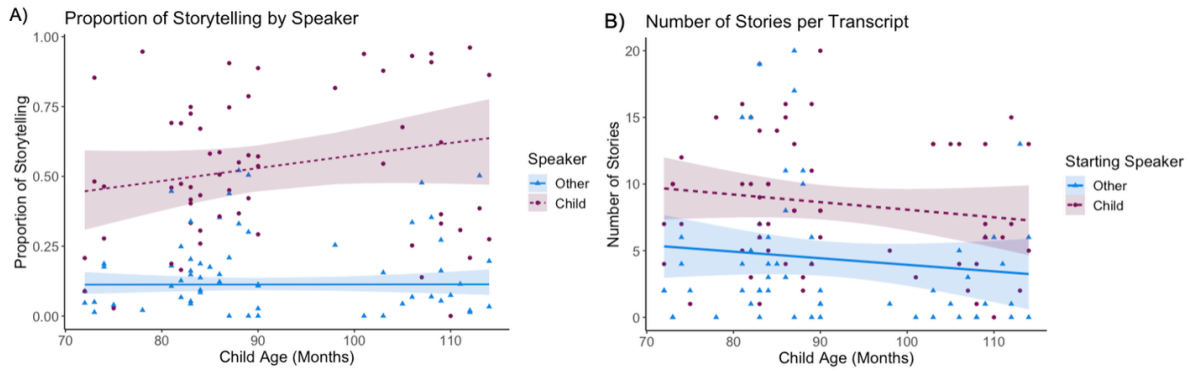

Figure 1. ***A) Proportion of utterances coded as "in-story" within each transcript, separated by speaker and plotted by child age (months). "Other" speakers were primarily parents and interviewers (96%), with the remainder consisting of grandparents, friends, or siblings of unknown age. Proportions were calculated separately within each speaker (i.e., the proportion of that speaker's utterances coded as in-story), rather than as a proportion of the total transcript; therefore, child and other proportions are not complementary. B) Number of story beginnings per transcript by age and speaker who started the story. Note that our coding scheme allowed for multiple speakers to contribute to the same story.***

In these transcripts, children contributed 51.1% of the utterances overall in these transcripts, so the speakers were evenly split between adults and children. Regardless of speaker, 37.7% of utterances were coded as in-story. Using the beginning and end of story coding, a total of 779 individual stories were detected in this sample. On average, these stories have a length of 7.67 utterances, but this length is quite variable (*SD* = 9.19 utterances).

Figure 1a depicts how the proportion of storytelling changes as a function of child age and speaker identity. In this sample, children spent 56% of their utterances telling stories, while caregivers only spent 18.5% of their utterances telling stories. To assess how this relationship between time spent storytelling changes with age, we fit a model predicting whether or not a given line was "in-story" from the child's age and who was speaking, with a random slope for the speaker by transcript. There was a significant effect of the speaker, where children were significantly more likely to tell stories than other speakers (*OR* = 0.33, *95% CI* [0.25, 0.43], $p = 8 \times 10^{-16}$). Children had between 56% and 75% higher likelihood of producing in-story utterances than other speakers. There was no significant effect of age (*OR* = 1.00, *95% CI* [0.99, 1.03], *p* = .28) or

significant interaction (*OR* = 0.99, *95% CI* [0.97, 1.01], $p = .40$), meaning that the rate of storytelling for both speaker groups was similar across ages in this sample.

Part 1: Does storytelling in conversation follow the morphosyntactic patterns seen in empirical studies.

The first set of analyses investigates whether storytelling in conversation follows the morphosyntactic patterns found in other empirical studies, where children are asked to tell a story in a laboratory setting. We will first explore how the length of children's stories changes with age, and then we will compare features of in-story utterances to out of story utterances to understand how storytelling differs from other types of communication in these dialogs.

In line with prior research, we expected children's stories to get longer between the ages of 6 and 7 years but decrease when they are 8-9 years old (Esposito et al., 2020). Figure 2a shows the average number of words contributed to a given story by the target child and their conversational partners as a function of the child's age. To assess age-related changes in story length, we compared two models, one with age as a linear predictor, and another with a quadratic form for age to account for the predicted parabolic relationship. In both of these models, only the linear term was significant ($\beta = 3.73$, *95% CI* [0.97, 6.51], $p = .01$; $\beta = 3.65$, *95% CI* [0.47, 6.83], $p = .03$, respectively), and the quadratic model did not fit the data significantly better ($\chi^2 = 0.01$, $p = .91$). Therefore, in our naturalistic dataset, we did not see evidence for the U-shaped curve suggested by prior laboratory research. Instead, our analyses indicated that children's stories on average increased in length between half a word to six and a half words per month.

Analyzing stories in naturalistic conversation allows us to capture how conversational dynamics, such as turn-taking while telling stories, develop across age. While no prior study has analyzed this, as an exploratory analysis, we hypothesized that as children get older, there would be a greater number of words spoken by the child when telling a story before another speaker would jump into the conversation. Figure 2b shows the average length of an in-story conversational turn by both children and adult speakers, as a function of the target child's age. A linear regression model predicting number of words before a speaker turn by the age of the child and who is speaking revealed a significant main effect of age ($\beta = 1.21$, *95% CI* [0.50, 1.93], $p = .002$), where on average, both adult and child in-story conversational turns increased in length between half a word to 2 words per month. The model also revealed a main effect of the speaker group, where adults spoke about 2 to 5 fewer words per turn than children ($\beta = -3.4$, *95% CI* [–5.11, –1.71], $p = .00009$).

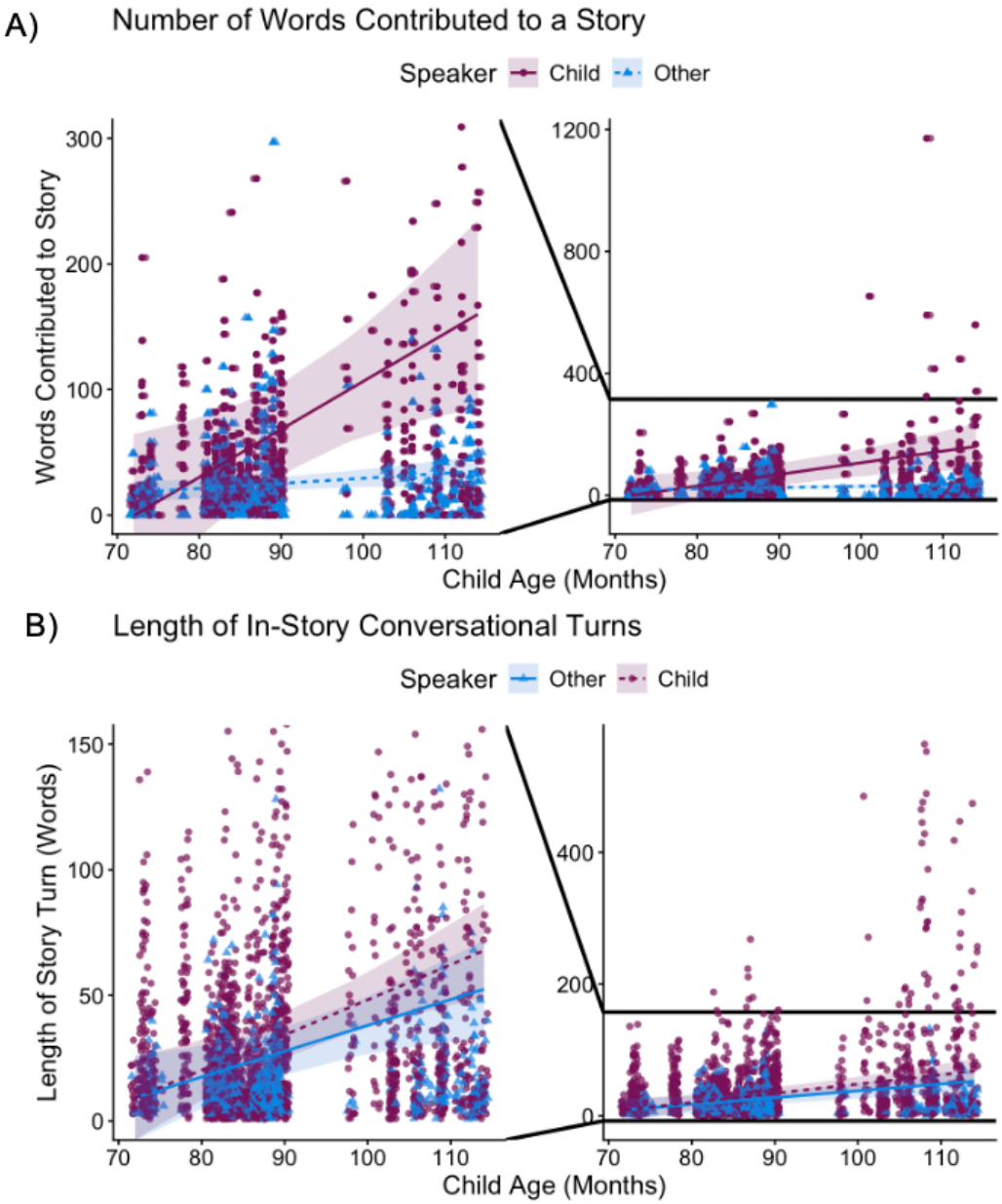

Figure 2. ***A) Average number of words in stories plotted as a function of age and speaker. Each data point represents the average number of words contributed to a story within a given transcript. B) Number of in-story words until the speaker changes, plotted as a function of child age and speaker. In both plots, "Other" speakers are primarily parents and interviewers (96%), while the remainder are grandparents, friends, or siblings of unknown age. For both images, the plot on the right displays the full dataset, while the plot on the left provides a magnified view of a smaller range.***

Finally, there was a significant interaction between age and speaker group where, as children got older, they contributed increasingly more words per story than their conversational partners did (β = –0.19, *95% CI* [–0.34, –0.03], p = .02). This interaction effect was small though, where the confidence band spans just a little more than 0 to a third of a word per month. Simple effects analysis revealed a significant positive trend for the length of children's in-story turns by age (b = 1.40, *95% CI* [0.67, 2.13]). Therefore, our hypothesis is fully supported: as children get older, they increasingly hold the floor longer while storytelling before an adult speaks.

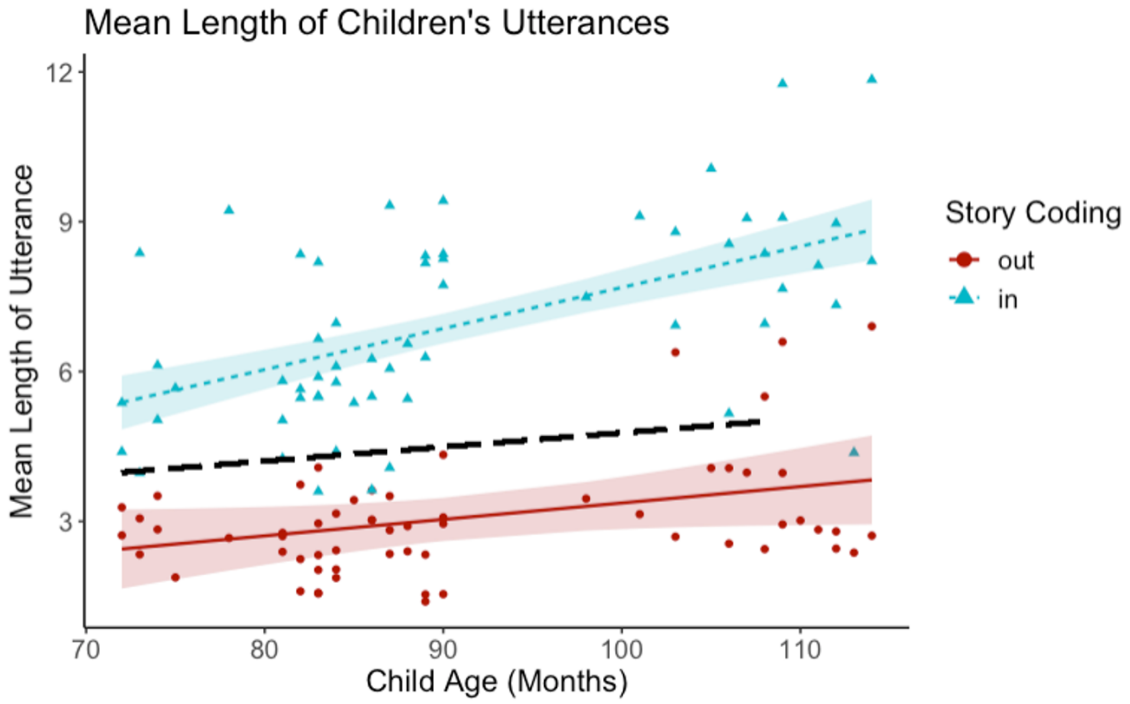

Figure 3. ***Average number of words in children's utterances in a given transcript for in and out of story utterances, plotted as a function of age. The long-dashed line depicts the mean length of utterance norms found by Rice et al. (2010).***

According to prior research, children's utterances (regardless of narrative intent) typically increase from an average of about 4 to 5 words between the ages of 6 and 9 (Rice et al., 2010). This pattern was determined from a monolingual English-speaking US sample, similar to our dataset, where children spoke freely with an investigator during toy play. Figure 3 shows the mean length of "in-" and "out- of" story utterances for each transcript in our sample. This model revealed a main effect of age whereas children get older, their utterances both in- and out-of-story get longer (β = 0.06, *95% CI* [0.03, 0.08], p = .00001). We also found a main effect of story coding (β = 1.93, *95% CI* [1.73, 2.14], $p < 2 \times 10^{-16}$). Across all ages (6 years to 9 years, 6 months), in-story

utterances had an average of 7.6 words (*SD* = 5.4), while out-of-story utterances had an average of 3.9 words (*SD* = 3.2). There was also a significant interaction between Story Coding and Age, such that with age, their storytelling utterance length increases faster with age than non-story utterances (β = 0.02, *95% CI* [0.01, 0.04], *p* = .004). In story utterances are increasing at a rate of 0.08 words per month (*95% CI* [0.05, 0.12]), while out of story utterances are increasing at a rate of 0.03 words per month (*95% CI* [0.01, 0.06]). These data are consistent with prior results demonstrating that MLU is longer in storytelling than when children produce other kinds of utterances (e.g., Potratz et al., 2022; Wagner et al., 2000). Taken together with the results on narrative length, our results are consistent with the suggestion that this coding scheme differentiates narrative utterances from non-narrative utterances. In the next section, we will explore some of the linguistic characteristics of these narrative utterances.

Part 2: What are the characteristics of children's naturalistic stories?

Child Utterance Characteristics. The following set of analyses will investigate how features of children's utterances change as a function of whether they are telling stories or not, and their age. We investigated three hypotheses, that as children age, they will use more variety of conjunctions as their stories contain more advanced structures (Stenning & Michell, 1985), will use more abstract nouns as they discuss more complex ideas (Buccino et al., 2019; Muraki et al., 2023; Ponari et al., 2017, 2020; Reggin et al., 2021; Vigliocco et al., 2009), and more first person pronouns in their stories as they tell less self-centered stories (Fernandez-Baizan et al., 2021; Keysar et al., 1998; Piaget, 1959; Vygotsky, 1987).

In predicting the number of conjunction words in children's stories, the linear regression model revealed a significant effect of the child's age on unique conjunctions, where children used significantly more conjunctions in all utterances as they got older (β = 0.009, *95% CI* [0.005, 0.013], *p* = .0001). There was also a significant effect of Story Coding, where the number of conjunctions was higher for utterances marked as in-story (β = 0.31, *95% CI* [0.27, 0.34], $p < 2 \times 10^{-16}$). There was a significant interaction between storytelling and age, where as children get older, the rate of unique conjunctions increased more rapidly than out-of-story utterances (β = 0.004, *95% CI* [0.0005 ,0.0065], *p* = .02), although note that the magnitude of this interaction is very small (less than one hundredth of an increase per month). When predicting total conjunctions instead of unique conjunctions, the results showed the same pattern. To explore how children use conjunctions throughout their stories, we also evaluated whether the rate of total conjunctions per *story* changed with age, but there was no significant effect at this level (β = 0.0006, *95% CI* [–0.0001, 0.0014], *p* = .14).

To assess the concreteness of children's utterances, the corresponding linear model revealed a significant effect of Story Coding, where in story utterances were more concrete than out of story utterances (β = 0.15, *95% CI* [0.11, 0.20], $p = 2 \times 10^{-9}$). There was no significant effect of age or interaction between age and Story Coding. The relationship between overall concreteness scores and storytelling did not change with age. However, looking at

average concreteness of an utterance obscures whether children are mixing a variety of abstract and concrete words as they get older. For each utterance, we calculated the standard deviation of concreteness scores within that utterance. Here the corresponding regression model revealed a significant effect of age, where the standard deviation of concreteness scores increased with age (β = 0.015, *95% CI* [0.008, 0.021], $p = .00003$), and a significant effect of story coding where story utterances contained more variety of concreteness scores compared to out of story utterances (β = 0.56, *95% CI* [0.50, 0.62], $p < 2 \times 10^{-16}$). Therefore, as children get older, they are more likely to mix concrete and abstract words in an utterance, and stories contain more variety of concreteness scores. However, the interaction was not significant, indicating that the age-related increase in variability was consistent in both in-story and out-of-story utterances.

Finally, in predicting the first-person pronoun use, there was no main effect of the child's age on the likelihood of the first person (*OR* = 1.01, *95% CI* [1.00, 1.02], $p = .22$). There was a main effect of story coding, where in-story utterances from children of all ages had between 32% and 78% higher odds to contain first-person pronouns than out-of-story utterances (*OR* = 1.53, *95% CI* [1.32, 1.78], $p = 2 \times 10^{-8}$). Critically, there was a significant interaction between story coding and age (*OR* = 0.99, *95% CI* [0.97, 1.0], $p = .02$). Investigation of the simple slopes of this interaction revealed that while the slope of "in-story" utterances did not change (β = –0.007, *95% CI* [–0.03, 0.01]), "out-of-story" utterances showed a significant increase in first person pronouns (β = 0.02, *95% CI* [0.009, 0.033]). Therefore, our hypothesis is not supported by this data. Children do not show a decrease in first-person pronouns for in-story utterances with age.

Dyadic Nature of Conversations. The following analyses will leverage the dyadic nature of conversations to investigate both how parents respond to children's utterances, and vice versa. We predicted that caregivers would repeat fewer utterances as children age, as caregiver repetitions could signal confirmation of comprehension, or as encouragement, and typically decrease with age (Clark & Bernicot, 2008; Leung et al., 2021). We also predicted that caregivers would ask fewer questions following stories, reflecting better understanding of older children's stories (Yu et al., 2019). Finally, we predicted that caregiver questions would prompt children's storytelling (Gelmini-Hornsby et al., 2011; Schick & Melzi, 2010).

To investigate how the rate of caregiver repetitions changed with age, we fit a linear regression model predicting the proportion of words caregivers repeated from a child's utterance from the interaction between the story coding of the *child's prior utterance* and age in months. There was a significant effect of story coding of the prior utterance in this model, where caregivers repeated more words from children's story utterances than other utterances (β = 0.03, *95% CI* [0.02, 0.04], $p = 5 \times 10^{-7}$). However, there were no significant effects of child age on this relationship, so our hypothesis was not supported.

To investigate whether adults were more likely to ask questions following storytelling

utterances than other types of child utterances, we fit a logistic regression model predicting whether an adult conversational turn contained a question from the story coding of the *preceding child* utterance and the child's age in months. The model included a random slope of the prior line story coding by subject. In the frequentist model, there was a small but statistically significant effect of age (*OR* = 0.98, *95% CI* [0.96, 0.999], *p* = .04), indicating that adults were slightly less likely to ask questions as children got older, regardless of whether the child was telling a story. However, this effect was attenuated in the Bayesian model (*OR* = 0.98, *95% CrI* [0.96, 1.0004]), where the credible interval included 1. There were no other significant effects in this model.

To investigate whether caregiver questions prompt storytelling, we fit a model predicting whether a child's initial utterance following a caregiver turn marked the start of a story. The model included the interaction between the child's age and whether the preceding caregiver conversational turn contained a question. In the frequentist model, there was a small but statistically significant effect of age (*OR* = 1.01, *95% CI* [1.00, 1.02], *p* = .04), indicating that older children were slightly more likely to begin a story following a caregiver turn, regardless of whether it contained a question. However, this effect was attenuated in the Bayesian model with a full random effects structure (OR = 1.01, *95% CrI* [0.99, 1.03]), where the credible interval included 1. In both models, there was clear evidence that caregiver questions predicted the beginning of stories. Stories were more likely to begin following a turn containing a question than following a turn consisting only of statements (frequentist: *OR* = 1.46, *95% CI* [1.28,1.65], $p = 5 \times 10^{-9}$), with questions indicating a 28 to 65% higher odds of a story beginning after a question. There was no evidence of an interaction between age and caregiver questions, suggesting that the association between caregiver questions and story beginnings did not meaningfully change with age.

Discussion

The primary question that motivated this research was to see whether it would be possible to identify narratives in ongoing conversation. Our results indicate that the coding rules developed for this task can serve this purpose, at least to an extent. The coding system is reliable across coders and appears to be appropriate for use across a wide span of developmental time. A second question that motivated the study was to determine whether the utterances that our coding scheme highlighted as narrative showed linguistic or pragmatic characteristics consistent with what is already known about storytelling behavior (e.g., Esposito et al., 2020; Hill et al., 2025). In this regard, our results are promising. In particular, we found that utterances identified as narrative showed a significantly greater mean length of utterance than non-narrative utterances, closely matching developmental trends observed in laboratory settings (e.g., Losi et al., 2022; Potratz et al., 2022; Wagner et al., 2000).

Admittedly, the current dataset does not allow us to verify our coding scheme against an important standard of "ground truth." We do not know if the speakers of these

utterances intended to convey narrative information or considered themselves to be doing so. Because we do not have direct access to the speaker's intent, it is impossible for us to know whether our coding scheme accurately reflects that intent. This is perhaps especially challenging because the data with which our coders worked was available in the form of a transcript, lacking both prosodic and visual information. Given our coding rules, where the minimal standard of narrative is describing two linked actions or events, it seems likely that the bias in our coding scheme tends toward false alarm. That is, there may be some proportion of the utterances that we coded as "in-story" that were in fact not intended to convey narrative information. While it is certainly possible that our coding also misses some utterances that were intended as narrative and incorrectly rates those utterances as "out-of-story," we suspect that the nature of our coding rules makes false alarms more likely.

The permissive bias of our coding scheme is intentional. At several points of debate in creating and refining our coding manual, we erred on the side of inclusivity in our definitions. The reason for this is that the extant literature indicates that when children begin to tell stories, they do so in a relatively simplistic fashion (e.g., Berman et al., 1994; Esposito et al., 2020). If we make our criteria for narrative utterances too strict, we run the risk of excluding some of children's nascent narrative efforts. By adopting a more permissive stance, we will be more likely to capture utterances that are intended as narrative by the youngest talkers in our sample. This gives us the opportunity to more easily ask questions about the origins of narrative behavior than if we only include utterances from those speakers who are already well advanced in the ability to convey narrative intent and information.

As this discussion highlights, a final question that motivated this research is whether it would be possible to assess children's conversational narratives in ways that can inform theories about development. In this regard, several of our analyses are suggestive of further potential. For example, we found that with age, children produced more consecutive utterances during longer conversational turns (e.g., Donnelly & Kidd, 2021). Similarly, we saw the expected increase in use of conjunctions as older children explore more complicated constructions to convey meaning (e.g., Cain et al., 2005; Scionti et al., 2023).

At the same time, our results were in some respects unexpected, suggesting that the context in which stories occur can influence children's narrative behavior. One notable example of this is our results about the length of children's stories, which did not show the U-shaped development consistently demonstrated in laboratory studies of storytelling (e.g., Berman, 1988; Esposito et al., 2020; Ukrainetz et al., 2005), in which children's stories first increase in length, and then decrease. A possible explanation here is the nature of our dataset in comparison to experimental paradigms, which we thank an anonymous reviewer for highlighting. In an experiment, such as those based around wordless picture books, there are typically a predetermined set of episodes,

and a predetermined ending. In this context, the decline in story length among older children can be interpreted as evidence of children's increasing efficiency in relaying information about the same/similar episodes. By contrast, in naturalistic conversation, there is no predetermined endpoint equivalent to the final page of a storybook. Indeed, as children get older, they likely have more episodes to discuss. In this context, we might not expect to see a decrease in overall story length, as we do in experimental settings; in conversation, children's increasing efficiency is offset by their ability (and their conversational partners') ability to continually add more information. If this interpretation is correct, the U-shape of storytelling length observed in prior research is at least partially an artifact of the specific tasks used to assess narrative skill. Another failure to replicate the prior research was observed with children's use of personal pronouns. On the basis of prior results, we anticipated that older children would use fewer first-person pronouns in their stories than younger children (e.g., Fernandez-Baizan et al., 2021; Keysar et al., 1998). Instead, we found little evidence that first-person pronoun usage declined in narrative. These unexpected results may stem from our small sample size, or our imperfect sampling across age. However, our results may also suggest the possibility that some developmental phenomena seen in laboratory settings do not generalize to more naturalistic conversational settings. We suggest that some of the surprising aspects of our results indicate opportunities for investigations of storytelling in conversational discourse that have not been fully incorporated into prior research on narrative development.

Importantly, though, our sample size is much too small to draw generalizable conclusions; the sample includes only 60 children, unevenly distributed across age. To make strong claims about development, or test rigorous hypotheses, it is necessary to have a much larger and more representative sample of conversational narrative. This is especially the case because social interactions are necessarily more variable than the kinds of monologues children can produce in experiments. Characterization of social interactions requires the analysis of a great deal more data, both cross-sectionally and longitudinally, before it will be possible to make confident claims about the fruitfulness of this approach. In Study 2, we will investigate the possibility of automating our story-coding process to yield the scale of data necessary for these types of developmental investigations.

Study 2

The coding rules that we implemented in Study 1 allow us to identify narratives distributed across talkers. However, the work of coding the transcripts is time-consuming both because of the hours required to read and judge each utterance and also because of the time necessary to learn the coding system. Coding the 60 transcripts - approximately 20,000 words - in Study 1 took over 100 hours. At that rate, it would take around 100,000 hours to code all the transcripts in CHILDES. And that is to say nothing of application to other repositories of conversational speech beyond CHILDES,

including some with a clinical focus such as UltraSuite (Eshky et al., 2019). Recent advances in machine learning, especially in large language models, provide a potential opportunity to vastly reduce the hours necessary to code a database. A high-quality Large Language Model (LLM)-based model (Achiam et al., 2023) of our trained coders' judgments of in- and out-of-story utterances could code a database the size of CHILDES in a few days, rather than a few decades.

Our goal in this study was to ascertain whether a LLM can model our story judgment rules, and judge utterances as in or out of story with high accuracy, precision, and recall as judged by trained human coders. That is, when making judgments about whether an utterance is in or out of story, do LLM judgments accurately reflect human coders? If a LLM can learn to mimic our story-coding system, it should agree with human coders about as often as human coders agree with each other. Conversely, if a LLM cannot learn our coding system, this result might indicate some idiosyncrasies either of LLM's representations of language, or of our own story coding system.

Method

Materials/Transcripts

We used the 60 coded transcripts from Study 1, in which every transcript line had been labeled as either "IN" or "OUT" of a story. (The 14 transcript lines labeled ambiguous were dropped from the transcripts.) This dataset was randomly split by transcript into 70% training (42 transcripts), 15% validation (9 transcripts), and 15% test (9 transcripts). Training and validation sets were used for model-training purposes. The test set was used exactly once to measure the accuracy of the model's classification (judgment) accuracy in comparison to human coders.

Model Fine-Tuning and Prediction

Our classification model is a fine-tuning of Open AI GPT-4o-mini pre-trained foundation model (Achiam et al., 2023).[1] The learning algorithm fine-tunes the foundation model to focus on classification, using a simple few-shot (Brown et al., 2020) learning structure that presents the training data in a comma separated value format.

For prediction, the input is 6 lines of coded transcript with conversational marks removed (Table 3). Prediction is accomplished by the model generating the in/out label of the example. (For predictions at the beginning of the transcript, the input is appropriately reduced, so for example the third transcript line uses the first three lines of the transcript only.)

[1] We also assessed the performance of GPT-2, but the performance was too low to be considered useful.

Table 3. ***Example input data to the fine-tuned model for inference. The top part of the table shows the raw transcript. The bottom part of the table shows the data provided for prediction by the model (without the column headers) where six lines are collapsed into a single example for prediction purposes. In the transcript, "INV" stands for investigator and "CHI" stands for child. The fine-tuned model, during inference, provides the predicted label of "in", "out", or "ambiguous".***

Raw Transcript
INV: no , are you really careful with your toys ? CHI: yeah . CHI: last year they all got broken INV: why , what got broken? CHI: John [//] brother , he smash them . INV: how old (i)s he then ?
Model Transcript
INV: no are you really careful with your toys CHI: yeah CHI: last year they all got broken IV: why what got broken CHI: John brother he smash them INV: how old is he then

Model Assessment

After fine-tuning, we presented the model with the test set. These 9 transcripts were completely novel, acquired from conversations with children outside the train and validation sets. That is, the test set transcripts were never presented during the fine-tuning process.

Model assessment occurs in two phases. In the first phase, we randomly select a validation subset of the training data. We then fine-tune the model on the remaining training data. For each utterance in these transcripts, the model made an in/out classification. Finally, we evaluate the model on the validation subset. This select-fine-tune-evaluate process creates a "fold". We then repeat this process five times to compute an average of the metrics across the folds. In the second phase, we fine-tune the entire training set and use the test set to evaluate the model. Since the test set is held out at the beginning of the assessment process, the evaluation of the model on the test set is free from any engineering bias introduced during model assessment.

We compared these classifications with those of human coders using standard machine learning metrics. For each of the N total utterances, the model classifies the utterance as "in" or "out". The classification belongs to one of four categories: true positives (TP) where the model and the human label are both "in", true negatives (TN) where the model and the human label are both "out", false positive (FP) where the model classification is "in" and the human label is "out", and false negative, where the model classification is "out" and the human label is "in". With the counts of results

in these categories, several metrics are computed.

Accuracy is defined as (TP+TN)/N and measures the overall model quality. Precision is defined as TP/(TP+FP) and measures the accuracy and the quality of positive "in" predictions. A high precision value indicates that, if the model makes a positive prediction, it is likely to be accurate. Recall is defined as TP/(TP+FN) and measures how well positive "in" predictions are captured. A high recall value indicates that, if a positive label exists, it is likely to be accurately predicted. Finally, F1 is defined as 2 * Precision * Recall / (Precision + Recall). This metric measures an equally weighted blend of precision and recall, and is used to characterize the overall performance of a model. All four metrics vary between 0 and 1 (Rainio et al., 2024).

Results

The overall accuracy of the model on the validation set is 0.90 (Table 4). This level of performance is very high and provides the opportunity for many practical applications of the model. The confusion matrix of the classifier indicates that prediction errors are balanced across the two classes (Table 5).

Table 4. ***Results of fine-tuned model based on GPT4-o mini for in/out coding. The fold 1-5 validation rows are the results of randomly partitioning the data into a partition five folds, i.e. each fold contains 20% of the data. We then fine-tuned a model on the remaining data, and evaluating the model on the fold validation set. The fold average row contains the average of the five folds. The test row is the results of the evaluation of a model, fined-tuned on the validation data, on the test set.***

In/Out Coding	Accuracy	Precision	Recall	F1
Fold 1	0.8604	0.7384	0.8547	0.7923
Fold 2	0.8959	0.8091	0.7719	0.7901
Fold 3	0.8401	0.8519	0.6103	0.7111
Fold 4	0.8897	0.8425	0.8778	0.8598
Fold 5	0.9090	0.9040	0.9010	0.9025
Fold Average	0.8790	0.8292	0.8031	0.8112
Test	0.9008	0.9105	0.8186	0.8621

Our motivation in conducting this study was to ascertain whether a large language model can apply our story-coding rubric accurately enough to be useful at scale. While the model's performance is below any theoretical optimum performance, it performs well enough for many practical purposes. One would be well advised to interpret any model judgments and attendant developmental conclusions with some degree of caution, as they are rougher approximations of human performance than could be the case. At the same time, the fact that our fine-tuned model produces judgments of narrative behavior that are generally convergent with human coders is a

promising demonstration of its potential applications in detecting and extracting stories from transcripts at scale.

Table 5. ***The confusion matrix for the in/out model.***

Validation Set	Prediction = In	Prediction = Out
True = In	641	142
True = Out	63	1220

Table 6. ***An example ambiguity in human coding and the corresponding error in prediction from the fine-tuned model. The "Transcript" column lists the utterances. The "CODER 1" and "CODER 2" columns are the independent judgments of the coders for these utterances. The coders disagree on lines 2 and 4. The "TRUE" column is the final judgment of the line after a discussion between the coders. The "Predicted" column is the prediction result of the fine-tuned model. Note that the prediction is incorrect on the third line of the example.***

Transcript	Coder 1	Coder 2	Agree?	TRUE	Predicted
INV: how old (i)s he then ?	out	out	TRUE	out	out
CHI: &-um , fifteen .	out	in	FALSE	in	in
INV: you shouldn't let him play with them .	out	out	TRUE	out	in
CHI: I know , but I xxx without me asking	ambiguous	in	FALSE	in	in
INV: what do you say when it plays with your toys?	out	out	TRUE	out	out

However, both human coders and classification models struggle in certain situations. For example, interrogative questions and statements are not considered part of the storytelling, but the answers to such questions or statements may constitute a story (Table 6). That said, performance is likely to be improved by further work. It may even be the case that simply more labeling will increase the amount of training data available to improve the model's performance without any other alterations to its architecture. Another route for potential improvements in performance relates to the fact that machine learning models often have an inherent model trade-off between precision and recall. For example, assigning a single response (either "in" or "out") for every classification generates a recall of 1.0 but terrible accuracy, precision and F1 scores. Because "in" and "out" statements are not equally balanced in the input, "defaulting" (or having a bias) toward one or the other will change the degree to which the model's performance appears to be better fit to human judgments. In fact, models

can be further tuned to prefer one judgment over the other (Minkov et al., 2006). We have not performed any additional tuning here, but it is possible.

While these results are promising, the model's less than optimal performance is but one reason to be cautious of its potential use as a research tool. Another reason is that many important analyses of narrative behavior rely not only on detecting whether an utterance is part of a narrative, but where that narrative begins and ends. For example, an oft-noted development in childhood is the inverted U-shape of story length: children's initial stories are short, growing longer for several years until they begin to decrease in length as children learn to be more concise (e.g., Esposito et al., 2020). Our model can be of only limited use if all it can learn is to identify whether an utterance is part of a story. To more fully capture human judgments of storytelling, a new start/end model must identify where stories begin and end, as our coders did in Study 1. A future direction for this work focuses on modeling entire stories in a transcript.

General Discussion

Two questions motivated this research. The first is whether it is possible to identify storytelling in ongoing conversation. If this feat can be achieved, it would provide a novel source of data with which researchers and clinicians could glean insights into the development of narrative skill. The second question is whether this identification of conversational storytelling can be automated using modern artificial intelligence tools, in this case GPT-4o-mini. Such automation could potentially accelerate and broaden the study of questions about narrative behavior that can be addressed using transcribed conversations. We believe that these results answer both questions in the affirmative, albeit with some important qualifications and cautions.

With respect to the first question, we found that training on our set of rules was pragmatically effective. Our research assistants understood the rules after only a brief (approximately week-long) training period and applied them in a way that was consistent across coders. Further, the utterances that these rules identified as dialog showed many of the characteristics we would expect if our coding rules accurately separated storytelling utterances from utterances of a non-narrative nature. For example, as seen in prior research (e.g., Losi et al., 2022; Miller et al., 1993; though see Southwood & Russell, 2004), we found that MLU was higher for narrative than non-narrative utterances. Similarly, we found that children "hold the floor" longer during storytelling as they get older, consistent with our understanding of turn-taking behavior across development (e.g., Donnelly & Kidd, 2021). If it is the case that our coding scheme is able to identify stories in dyadic conversations, this approach will open up a new avenue of investigation into the development of narrative skill.

While our results give cause for enthusiasm on this score, they should still be interpreted with some degree of caution for several reasons. First, not all of the

developmental patterns we found are consistent with the predictions of the prior literature. For example, we did not replicate the oft-found result that story length should show an inverted-U pattern (e.g., Esposito et al., 2020). This raises the possibility that some (perhaps many) of the utterances that are highlighted by our story coding system are not actually intended as part of a narrative. Alternatively, of course, it raises the possibility that some (perhaps many) characteristics of storytelling behavior look different in dyadic conversations than in the (largely monologue) storytelling tasks typically used in laboratory or clinical settings (e.g., Abbeduto et al., 1995; Meyer, 1969). It may be impossible to fully disentangle these possibilities with our current dataset, which is entirely based on audio transcriptions; in the absence of knowledge of the visual world shared by the speaker, some utterances become irretrievably ambiguous in their purpose. That said, exploring transcriptions from the same children over longitudinal time may hold some promise for resolving this question. For example, a deeper exploration of the types and tokens that populate children's stories may give us insight into children's narrative intent and may illuminate individual differences in development. As a first step, it must be noted that, our sample of participants is in many ways too small to make confident claims, so simply replicating this research over a wider age range could well be informative.

The need to replicate and extend this research at scale directly motivates our second question: whether our coding scheme can be learned and automatically applied by a LLM. Our results on this score are promising, as the fine-tuned model was able to replicate our coders' judgments with a satisfactory degree of accuracy. While this performance is a necessary prerequisite to the use of a LLM at scale, it is no more than a proof of concept. It remains to be seen whether our approach can actually be scaled up in a practical and productive fashion. A related concern is that our current model only identifies whether utterances are narrative or non-narrative in nature, but many of the most important developments in storytelling behavior involve the length of a story. As such, it is also important to know where stories begin and end, which our model currently has no way to identify. In future work, we will train a model on this task and see if it can replicate human judgments.

With respect to both human and LLM judgments, one qualification to consider is that our definition of a story is quite permissive: any two linked utterances that describe an action or event meets our minimal definition of a story. The permissive nature of this coding scheme is intentional, as we wanted it to be sensitive to very young children's attempts at narrative behavior, which are often simplistic or underspecified (e.g., Berman, 1988; Vion & Colas, 1999). However, it may reduce the signal to noise ratio in our coding scheme, which may subsequently make it harder for the model to distinguish between utterances that are in or out of a story. More refined or elaborate schemes may produce greater degrees of agreement between coders and models, though such a system could incur a cost in terms of an increase in the difficulty of training both humans and machines. While the cost/benefit ratio may favor more

complex coding schemes in some settings, we believe that our more permissive scheme has an advantage in developmental research. That said, we acknowledge that the advantages and disadvantages of various modifications to our story coding system is an empirical question.

Above and beyond parametric variation of our coding scheme and other psychometric assessments, these results open a variety of avenues toward novel research questions. One of the most pressing is to see whether we can improve model performance and breadth of application with more and more varied training materials. Even without further improvements, the coding rules and model developed in this research could immediately be applied to assess storytelling in populations who have not traveled to participate in laboratory research, but whose conversations have been recorded and transcribed. Clinical populations or children with developmental disorders may be of especial interest, given prior results suggesting that narrative ability is associated with a variety of psychopathologies such as dementia, schizophrenia, and attention deficits (e.g., Cowan & Lind, 2024; Faroqi-Shah et al., 2020; Tannock et al. 1993; Usun et al., 2023). Because our coding scheme was developed for use with conversations dominated by children, a pressing question particularly for developmental psychologists is to assess how this coding scheme can be applied to adult conversations, and what differences might emerge between child and adult storytelling in a conversational setting.

Storytelling is a universal human behavior, and one of the basic building blocks of civilization. Nevertheless, as this discussion indicates, there are still many mysteries about the emergence and use of narrative abilities. Not least among those mysteries is the following: why are some of us better at telling stories than others? We believe that conversational interactions play a key role in shaping storytelling skills (e.g., Haden et al., 1997), and hope that the tools we have developed here can contribute to a greater understanding of narrative development. In particular, machine learning may be helpful in evaluating the kinds of information present in conversational settings and assessing which aspects of that information are especially critical for the development of narrative skills. A useful theory of the processes behind narrative development must inevitably give rise to ways to intervene in the development of storytelling abilities. Few of our stories will be as epochal or archetypal as those of Gilgamesh and Enkidu. Even so, the ability to convey one's own stories is an essential human skill, and there are many individuals – from second language learners to academic writers – who could benefit from targeted interventions wisely designed to improve narrative skills.

References

Abbeduto L., Benson G., Short K., & Dolish J. (1995). Effects of sampling context on the expressive language of children and adolescents with mental retardation. *Mental Retardation, 33*, 279–288.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... Zoph, B. (2023). *Gpt-4 technical report.* arXiv preprint arXiv:2303.08774.

Akçay Usun, H., Güneri, G., Şimşek, Ö. F., & Kocayörük, E. (2023). The effect of childhood trauma on psychopathology and well-being: Personal narratives as mediating variables. *Journal of Family Trauma, Child Custody & Child Development*, *20*(4), 410–428. https://doi.org/10.1080/26904586.2022.2164544

Babayiğit, S., Roustone, S., & Wren, Y. (2020). Linguistic comprehension and narrative skills predict reading ability: A 9-year longitudinal study. *British Journal of Educational Psychology.* https://doi.org/10.1111/bjep.12353

Baldassano, C., Hassan, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *The Journal of Neuroscience, 38*(45), 9689-9699. https://doi.org/10.1523/JNEUROSCI.0251-18.2018

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keeping it maximal. *Journal of Memory and Language, 68*, 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beliavsky, N. (2003). The sequential acquisition of pronominal reference in narrative discourse. *Word, 54*, 167-189. https://doi.org/10.1080/01638530309535043

Berman, R.A., (1988). On the ability to relate events in narrative. *Discourse Processes, 11*, 469-497. https://doi.org/10.1080/01638538809544714

Berman, R. A., Slobin, D. I., Aksu-Koç, A. A., Bamberg, M., Dasinger, L., Marchman, V., Neeman, Y., Rodkin, P. C., Sebastián, E., et al. (1994). *Relating events in narrative: A crosslinguistic developmental study.* Erlbaum.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Bliss, L. S., McCabe, A., & Miranda, A. E. (1998). Narrative assessment profile: Discourse analysis for school-age children. *Journal of Communication Disorders, 31*(4), 347-363 https://doi.org/10.1016/S0021-9924(98)00009-4

Bower, G.H., Black, J.B., & Turner, T.J. (1979). Scripts in memory for text. *Cognitive Psychology, 11,* 177-220. https://doi.org/10.1016/0010-0285(79)90009-4

Boyd, B. (2009). *On the origin of stories: Evolution, cognition, and fiction.* Cambridge University Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901. https://doi.org/10.5555/3495724.3495883

Bruner, J. (1975). From comprehension to narrative - A psychological perspective. *Cognition, 3,* 225-287. https://doi.org/10.1016/0010-0277(75)90016-6

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2013). Concreteness ratings for 40,000 generally known English word lemmas. *Behavior Research Methods, 46*, 904-911. https://doi.org/10.3758/s13428-013-0403-5

Buccini, G., Colagé, I., Silipo, F., D'Ambrosio, P. (2019). The concreteness of language: An ancient issue and a new perspective. *Brain Structure and Function, 224*, 1385-1401. https://doi.org/10.1007/s00429-019-01857-z

Burdelski, M. (2019). Young children's multimodal participation in storytelling. *Research on Children and Social Interaction, 3*, 6-35. https://doi.org/10.1558/rcsi.37284

Burdelski, M., & Fukuda, C. (2019). Multimodal membership categorization and storytelling in a guided tour. *Pragmatics and Society, 10*, 337-358. https://doi.org/10.1075/ps.18005.bur

Bürkner P. C. (2017). brms: An R Package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1-28. https://doi.org/10.18637/jss.v080.i01

Cain, K., Nikole, P., & Andrews, L. (2005). Age- and ability-related differences in young readers' use of conjunctions. *Journal of Child Language, 32,* 877-892. https://doi.org/10.1017/S0305000905006884

Channell M., Loveall S., Conners F., Harvey D., Abbeduto L. (2018). Narrative

language sampling in typical development: Implications for clinical trials. *American Journal of Speech and Language Pathology, 27*, 123-135. https://doi.org/10.1044/2017_AJSLP-17-0046

Clark, E. V., & Bernicot, J. (2008). Repetition as ratification: How parents and children place information in common ground. *Journal of child language*, *35*(2), 349-371. https://doi.org/10.1017/S0305000907008306

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory of Language, 50*, 62-81. https://doi.org/10.1016/j.jml.2003.08.004

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *Advances in Psychology, 9,* 287-299. https://doi.org/10.1016/S0166-4115(09)60059-5

Cowan, H. R., & Lind, M. (2024). Narrative identity disturbances in psychopathology: An ecologically valid transdiagnostic framework. *Journal of Psychopathology and Clinical Science, 133*, 503-504. https://doi.org/10.1037/abn0000917

Crawshaw, C. E., Kern, F., Mertens, U., & Rohlfing, K. J. (2020). Children's narrative elaboration after reading a storybook versus viewing a video. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.569891

Creswell, K. G. (2020). Drinking together and drinking alone: A social-contextual framework for examining risk for alcohol use disorder. *Current Directions in Psychological Science, 30*(1), 19-25. https://doi.org/10.1177/0963721420969406

Davidson, A. J., Walton, M. D., Kansal, B., & Cohen, R. (2017). Narrative skills predict peer adjustment across elementary school years. *Social Development, 26*(4), 891-906. https://doi.org/10.1111/sode.12236

Dideriksen, C., Christiansen, M. H., Tylén, K., Dingemanse, M., & Fusaroli, R. (2023). Quantifying the interplay of conversational devices in building mutual understanding. *Journal of Experimental Psychology: General, 152*(3), 864–889. https://doi.org/10.1037/xge0001301

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development, 92,* 609-625. https://doi.org/10.1111/cdev.13511

Elmlinger, S.L., Levy, J.A., & Goldstein, M.H. (2025). Immature vocalizations elicit simplified adult speech across multiple languages. *Current Biology*, 35, 871-881. https://doi.org/10.1016/j.cub.2024.12.031

Eshky, A., Ribeiro, M. S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J., & Wrench, A. (2019). UltraSuite: A repository of ultrasound and acoustic data from child speech therapy sessions. https://doi.org/10.21437/Interspeech.2018-1736

Esposito, G., Venuti, P., Iandolo, G., de Falco, S., Gabriella, G., Wei, C., & Bornstein, M.H. (2020). Microgenesis of typical storytelling. *Early Child Development and Care, 190,* 1991-2001. https://doi.org/10.1080/03004430.2018.1554653

Faroqi-Shah, Y., Treanor, A., Ratner, N.B., Ficek, B., Webster, K., & Tsapkini, K. (2020). Using narratives in differential diagnosis of neurodegenerative syndromes. *Journal of Communication Disorders, 85,* 1-11. https://doi.org/10.1016/j.jcomdis.2020.105995

Fernandez-Baizan, C., Arias, J. L., & Mendez, M. (2021). Spatial orientation assessment in preschool children: Egocentric and allocentric frameworks. *Applied Neuropsychology: Child, 10*(2), 171–193. https://doi.org/10.1080/21622965.2019.1630278

Gardner-Neblett, N., & Iruka, I.U. (2015). Oral narrative skills: Explaining the language-emergent literacy link by race/ethnicity and SES. *Developmental Psychology, 51,* 889. https://doi.org/10.1037/a0038612

Gazella, J. & Stockman, I.J. (2003). Children's story retelling under different modality and task conditions: Implications for standardizing language sampling procedures. *American Journal of Speech Language Pathology, 12,* 61-72. https://doi.org/10.1044/1058-0360(2003/053)

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429258411

Gelmini-Hornsby, G., Ainsworth, S., & O'Malley, C. (2011). Guided reciprocal questions to support children's collaborative storytelling. *International Journal of Computer-Supported Collaborative Learning, 6*, 577-600. https://doi.org/10.1007/s11412-011-9129-5

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grishakova, M., Sorokin, S. (2016). Notes on narrative, cognition, and cultural evolution. *Sign System Studies, 44(4),* 542-561. https://doi.org/10.12697/SSS.2016.44.4.02

Grossen, M. (2009). Interaction analysis and psychology: A dialogical perspective. *Integrative Psychological and Behavioral Science, 44*, 1-22. https://doi.org/10.1007/s12124-009-9108-9

Haden, C.A., Haine, R.A., & Fivush, R. (1997). Developing narrative structure in parent-child reminiscing across the preschool years. *Developmental Psychology, 33*, 295-307. https://doi.org/10.1037/0012-1649.33.2.295

Herman, D. (2003). Stories as a tool for thinking. In D. Herman (Ed.), *Narrative theory and the cognitive sciences* (pp. 163–192). Center for the Study of Language and Information.

Hill, K. A., Cohen, S. S., Olson, I. R., & Newcombe, N. S. (2025). The role of narrative structure in scaffolding children's recall. *Journal of Cognition and Development*, 1-14. https://doi.org/10.1080/15248372.2025.2571554

Imada, T. (2010). Cultural narratives of individualism and collectivism: A content analysis of textbook stories in the United States and Japan. *Journal of Cross-Cultural Psychology, 43(4),* 576-591. https://doi.org/10.1177/0022022110383312

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science, 7*(2), 46–50. https://doi.org/10.1111/1467-8721.ep13175613

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lenth, R. (2025) *emmeans: Estimated Marginal Means, aka Least-Squares Means.* R package version 1.11.1-00001, https://rvlenth.github.io/emmeans/.

Leung, A., Tunkel, A., & Yurovsky, D. (2021). Parents fine-tune their speech to children's vocabulary knowledge. *Psychological Science, 32*(7), 975-984. https://doi.org/10.1177/0956797621993104

Lorenz, K. (1950). The comparative method for studying innate behavior patterns. *Symposia of the Society for Experimental Biology, 4,* 221-254.

Losi, R.V., Tasril, V., Widya, R., & Akbar, M. (2022). Using storytelling to develop English vocabulary on early age children measured by mean length of utterance. *International Journal of English and Applied Linguistics, 2*, 179-187. https://doi.org/10.47467/ijeal.v2i1.1282

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Erlbaum.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs* (3rd ed.). Erlbaum.

Mar, R. A., Li, J., Nguyen, A. T. P., & Ta, C. P. (2021). Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin and Review, 28*, 732–749. https://doi.org/10.3758/s13423-020-01853-1

Marcus, G.F. (1993). Negative evidence in language acquisition. *Cognition*, *46*, 53-85. https://doi.org/10.1016/0010-0277(93)90058-4

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 119*(2), 313-330. https://doi.org/10.3115/974481.974559

Marjanovic-Umek, L., Kranjc, S., Fekonja, U. (2002). *Developmental levels of child's storytelling.* Paper presented at the 12th Annual Meeting of the European Early Childhood Association, Lefkosia, Cyprus.

Mayer, M. (1969). *Frog, Where Are You?* The Dial Press.

McCabe, A., (1997). Developmental and cross-cultural aspects of children's narration. In M. Bamberg (Ed.), *Narrative development* (pp. 137–174). Routledge.

McCabe, A., & Peterson, C. (1991). Getting the story: A longitudinal study of parental styles in eliciting narratives and developing narrative skill. In A. McCabe & C. Peterson (Eds.), *Developing narrative structure* (pp. 217–253). Lawrence Erlbaum Associates, Inc.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Merritt, D. D., & Liles, B. Z. (1989). Narrative analysis: Clinical applications of story generation and story retelling. *Journal of Speech and Hearing Disorders, 54*, 438-447. https://doi.org/10.1044/jshd.5403.438

Miller, P.J., Hoogstra, L., Mintz, J., Fung, H., & Williams, K. (1993). Troubles in the garden and how they get resolved: A young child's transformation of his favorite story. In C.A. Miller (Ed), *Memory and affect in development: Minnesota symposia on child psychology* (pp. 87-114). Lawrence Erlbaum Associates.

Minkov, E., Wang, R. C., Tomasic, A., & Cohen, W. (2006, June). NER systems that

suit user's preferences: adjusting the recall-precision trade-off for entity extraction. In Proceedings of the human language technology conference of the NAACL, companion volume: short papers (pp. 93-96). https://doi.org/10.3115/1614049.1614073

Montanari, S. (2004). The development of narrative competence in the L1 and L2 of Spanish-English bilingual children. *International Journal of Bilingualism, 8*, 449-497. https://doi.org/10.1177/13670069040080040301

Muraki, E. J., Reggin, L. D., Feddema, C. Y., & Pexman, P. M. (2023). The development of abstract word meanings. *Journal of Child Language, 52*, 195-207. https://doi.org/10.1017/S0305000923000429

Nesbit, J. C., & Hadwin, A. F. (2006). Methodological issues in educational psychology. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (pp. 825–847). Lawrence Erlbaum Associates.

Pesco D., & Gagné, A (2017) Scaffolding narrative skills: A meta-analysis of instruction in early childhood settings. *Early Education and Development, 28(7)*, 773-793, https://doi.org/10.1080/10409289.2015.1060800

Piaget, J. (1959). *The Language and Thought of the Child*. Routledge & Kegan Paul.

Pinto, G., Tarchi, C., & Gamannossi, B.A. (2018). Kindergartners' narrative competence across tasks and time. *The Journal of Genetic Psychology, 179*, 143-155. https://doi.org/10.1080/00221325.2018.1453775

Ponari M., Norbury C. F., & Vigliocco G. (2017). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science, 21(2)*, e12549. https://doi.org/10.1111/desc.12549

Ponari M., Norbury C. F., & Vigliocco G. (2020). The role of emotional valence in learning novel abstract concepts. *Developmental Psychology, 56,* 1855-1865. https://doi.org/10.1037/dev0001091

Potratz, J. R., Gildersleeve-Neumann, C., & Redford, M. A. (2022). Measurement properties of mean length of utterance in school-age children. *Language, Speech, and Hearing Services in Schools, 53*(4), 1088-1100. https://doi.org/10.1044/2022_LSHSS-21-00115

Pratt, M. W., Luszcz, M. A., MacKenzie-Keating, S., & Manning, A. (1982). Thinking about stories: The story schema in metacognition. *Journal of Verbal Learning and Verbal Behavior, 21*, 493-505. https://doi.org/10.1016/S0022-5371(82)90727-9

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.1) [Computer software]. *R Foundation for Statistical Computing*. https://www.R-project.org/.

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports, 14(1)*, 6086. https://doi.org/10.1038/s41598-024-56706-x

Reggin L.D., Muraki E.J., & Pexman P.M., (2021). Development of abstract word knowledge. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2021.686478

Rice, M.L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, B. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research, 53*, 333-349. https://doi.org/10.1044/1092-4388(2009/08-0183)

Schick, A., & Melzi, G. (2010). The development of children's oral narratives across contexts. *Early Education and Development*, *21*(3), 293–317. https://doi.org/10.1080/10409281003680578

Scionti, N., Zampini, L., & Marzocchi, G.M. (2023). The relationship between narrative skills and executive functions across childhood: A systematic review and meta-analysis. *Children, 10*, 1391. https://doi.org/10.3390/children10081391

Severing, R., & Verhoeven, L. (2001). Bilingual narrative development in Papiamento and Dutch. In L. Verhoeven and S. Strömqvist (Eds), *Narrative Development in a Multilingual Context.* John Benjamins Publishing Company. https://doi.org/10.1075/sibil.23.10sev

Southwood F, & Russell A.F. (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech Language and Hearing Research, 47*(2), 366-76. https://doi.org/10.1044/1092-4388(2004/029)

Sperry, L. L., & Sperry, D. E. (1996). Early development of narrative skills. *Cognitive Development, 11*(3), 443–465. https://doi.org/10.1016/S0885-2014(96)90013-1

Stenning, K., & Michell, L. (1985). Learning how to tell a good story: The development of content and language in children's telling of one tale. *Discourse Processes, 8,* 261-279. https://doi.org/10.1080/01638538509544614

Tannock R., Purvis K. L., Schachar R. J. (1993). Narrative abilities in children with attention deficit hyperactivity disorder and normal peers. *Journal of Abnormal Child Psychology, 21(1),* 103-117. https://doi.org/10.1007/BF00910492

Tolins, J. & Fox Tree, J.E. (2014). Addressees backchannels steer narrative development. *Journal of Pragmatics, 70*, 152-164. https://doi.org/10.1016/j.pragma.2014.06.002

Tolins, J., & Fox Tree, J. E. (2016). Overhearers use addressee backchannels in dialog comprehension. *Cognitive Science, 40*(6), 1412–1434. https://doi.org/10.1111/cogs.12278

Ukrainetz, T. A., Justice, L. M., Kaderavek, J. N., Eisenberg, S. L., Gillam, R. B., Harm, H. M., (2005). The development of expressive elaboration in fictional narratives. *Journal of Speech, Language, and Hearing Research, 48*, 1363-1377. https://doi.org/10.1044/1092-4388(2005/095)

Vigliocco G., Meteyard L., Andrews M., Kousta S. (2009). Toward a theory of semantic representation. *Language and Cognition, 1*, 219–248. https://doi.org/10.1515/LANGCOG.2009.011

Vion,M. & Colas, A. (1999) Maintaining and reintroducing referents in French: Cognitive constraints and development of narrative skills. *Journal of Experimental Child Psychology, 72*, 32-50. https://doi.org/10.1006/jecp.1998.2475

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Vygotsky, L. S. (1987). Thinking and speech. In R.W. Rieber & A.S. Carton (Eds.), *The collected works of L.S. Vygotsky, Volume 1: Problems of general psychology (pp. 39–285)*. Plenum Press. (Original work published 1934.)

Wagner, C. R., Nettelbladt, U., Sahlén, B., & Nilholm, C. (2000). Conversation versus narration in pre-school children with language impairment. *International Journal of Language & Communication Disorders, 35*(1), 83-93. https://doi.org/10.1080/136828200247269

Ward, N., & Tsukuhara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese of *Journal of Pragmatics, 32*, 1177-1207. https://doi.org/10.1016/S0378-2166(99)00108-3

Weisleder A, Fernald A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science, 24(11)*, 2143-52. https://doi.org/10.1177/0956797613488145

Yabe, M., Oshima, S., Eifuku, S., Taira, M., Kobayashi, K., Yabe, H., & Niwa, S. I.

(2018). Effects of storytelling on the childhood brain: near-infrared spectroscopic comparison with the effects of picture-book reading. *Fukushima Journal of Medical Science*, *64*(3), 125-132. https://doi.org/10.5387/fms.2018-11

Yu, Y., Bonawitz, E. and Shafto, P. (2019), Pedagogical questions in parent–child conversations. *Child Development, 90,* 147-161. https://doi.org/10.1111/cdev.12850

Zanchi, P., & Zampini, L. (2020). A standardized test to assess children's narrative skills. *European Journal of Psychological Assessment, 37,* 15-22. https://doi.org/10.1027/1015-5759/a000603

Data, Code and Materials Availability Statement

The data, code, and materials necessary to reproduce all the analyses presented here, as well as the pre-registration for analyses, are publicly accessible at the following Open Science Framework Repository: https://osf.io/bzgan/. (DOI 10.17605/OSF.IO/BZGAN)

Ethics Statement

All data used in this project were shared from the CHILDES repository (MacWhinney, 1991). This research is consistent with both APA ethical principles, and the ethical principles laid out in the CHILDES TalkBank Code of Ethics.

Authorship and Contributorship Statement

Anthony Tomasic and **Erik Thiessen** conceptualized the project. **Madeline Elston** and **Megan Waller** operationalized the narrative, developed a coding manual, and coded all transcripts. **Anthony Tomasic** and **Dhruv Nambiar** devised the approach for using ChatGPT to identify storytelling. **Dhruv Nambiar** implemented and trained ChatGPT, as well as reporting its results. **Dhruv Nambiar** and **Megan Waller** analyzed data, with **Dhruv Nambiar** responsible for descriptive statistics of mode behavior, and **Megan Waller** responsible for regression analyses. **Megan Waller** created all figures. **Erik Thiessen** and **Megan Waller** wrote the first draft of the manuscript. All authors contributed to revising and finalizing the manuscript.

Acknowledgements

We thank Sarah Fisher with her help developing the first draft of the coding manual. We thank Alyssa Robert for her help finding transcripts and refining the coding manual. We thank Jasmine Cha and Joceyln Cordero for their coding of story beginnings and endings. We thank Molly Niehaus for managing undergraduate student workloads.

Appendix A

The following is the text provided to human coders for identifying storytelling in transcripts:

What is a NARRATIVE?

We define a narrative as a section of speech that contains at least two **interconnected** statements, at least one of which is an **action** or **event**.

What is an ACTION or EVENT?

An **action** describes a line in which a subject is described as performing something, often to achieve an aim. "The little duck goes swimming" would be an action.
An **event** describes how something changed from one state to another. "It began to rain" would be an example of an event.

These actions and events may be fictional or non-fictional, and can take place in the past, present, or future, and can be hypothetical or actively occurring. "I would put on my dress and show it to the class!" would be considered an action, as it contains an explicit action that could possibly be performed (like the future tense, or conditional if-then statements).

Statements such as preferences, desires, descriptions of current states, or imperatives (commands) are not considered narrative actions or events. "I love playing basketball" or "She has two sisters" or "Pass me the salt" alone are not considered narrative statements. However, all but imperatives can be part of a narrative if they are connected to an action statement (see below). Statements like "I need to/have to/can put on a dress" are still statements of fact. While they describe an action, no action is being performed.

- However, children may use frames such as "hafta" incorrectly. When more than three actions or events are repetitively linked in this way, they may be considered a narrative.

Metacommentary about an action or event, where the speaker is giving an opinion about the narrative, separate from telling the story itself (such as "what a silly thing to say") is not an event or action.

What is considered "INTERCONNECTED"?

When an action or event is identified, if that statement is connected to another statement either temporally (e.g., using sequencing such as "first," "then," etc), causally

(e.g., "because"), or thematically (i.e., linking concepts together, such as directly describing character, setting or events), a narrative is present. One event or action is not sufficient to constitute a narrative, it must be interconnected to another statement, event, or action.

- Note for "if,then" format each line constitutes its own action, therefore forming a narrative.
-

Actions or events **not** interconnected with another do not constitute a narrative, as those could be considered states of fact without a narrative connection between them. By this definition, statements of preference or fact that are related to an action or event in these ways should be considered "in story". When a narrative is identified, any lines following and preceding the action or event that are linked in any of three ways listed above should also be marked as in story. Metacommentary and commands should still be out-of-story, unless they are part of dialog.

Multiple Speakers

Narratives can be shared between multiple speakers and don't have to be specific to one person. If two or more lines fit the definition of narrative stated above, it doesn't matter who said the lines, they should all be marked as in story.

As one person is speaking, another person may provide short statements or clarification questions to indicate they are listening or understand the story better. These should be marked as out of story. For example, back-channelling behavior, meaning a conversational partner providing brief commentary to indicate they are listening ("mmhm", "right"), should be labeled as out of story. However, answers to these questions, as long as they are not vague "yes" or "no" answers, should be marked as in-story.

Parents also often prompt children to tell them stories in these transcripts. These prompts should be marked as out of story. Generally, questions and their responses should also be marked out of story even though the question may contain an event. If the answer does not support an ongoing story or contain an action/event, it should be marked as out of story. For example, "Did you do math problems at school today" should be marked as out of story.

Line divisions in the transcript are arbitrary, therefore, some lines might not contain full statements. If the same person is speaking, and the lines surrounding it are narrative, as long as the phrase doesn't directly break out of story (such as an aside about managing the microphone), it can be marked as in. This also includes one word lines at the start of a narrative like "yes" or "no".

Ambiguous

Without knowledge of the scene in front of us, or the intentions of the speaker, sometimes it is unclear whether a line is a narrative or linked to a narrative, and should be marked as ambiguous. Here are some common ambiguous situations:

- The lines relate temporally to the narrative, but do not contain enough information to confirm whether or not they actually relate.
- This narrative is not **explicit**, meaning that the action is not directly stated, but instead is implied. "I do it" alone does not provide enough explicit information to understand the narrative, therefore lines that were underspecificed should be marked as ambiguous.
- The lack of subject or ungrammatical utterances makes it hard to fully guarantee that the line is part of a narrative.
- The thematic relationship between the line and a previous narrative is possible but not directly clear
- Sound effects that are unclear in their relationship to the story

Lines with just "xxx" should be marked as out.

Appendix B

Transcript Formatting and Cleaning

Each transcription file comes in two parts - the metadata for the transcription and the actual transcription itself. Each transcription is a conversation between a single child and one or more adults. The metadata file records the age of the child (in months) at the time of the recording. The transcript file contains a single line for each utterance of a person in the transcript. The utterance length is determined by a set of rules in the CHILDES transcription creation rulebook. In the CHILDES system, an utterance is considered to be a continuous and complete stretch of meaningful speech produced by a single talker. It is typically bounded by silence, or by a change of talker. These utterances can vary in length from a single word to multi-clause statements.

Important for our coding purposes, each utterance line starts with a symbol indicating the speaker. The remainder of the line is the spoken content. The content itself contains the spoken words plus additional notations for sounds, intonations, pauses, and other acoustic features specified in the CHILDES manual. To make the process of coding less laborious, each 'raw' transcript was 'cleaned' by removing all orthographic symbols aside from the plain English text of the conversation. This entailed splitting each utterance into segments delimited by spaces. The angle brackets and parentheses were deleted, and we removed all segments where every character is not a letter (e.g., it contains transcription symbols that indicate it was only partially transcribed), all segments that are completely contained with square brackets, and all segments that start with an ampersand. The segments that remain are rejoined in their

original order with spaces between them. See below for an example of a raw and cleaned transcript.

Example of original transcript file:
CHI: Mama look_it telephone !
MOT: telephone you gonna call somebody ?
MOT: (.) yeah .
MOT: I like the animal xxx .
MOT: who you gonna call ?
CHI: I don't know but this phone don't work .
CHI: can you get it to +/.
MOT: oh [!] there you go .

Example of cleaned transcript file:
CHI: Mama look_it telephone
MOT: telephone you gonna call somebody
MOT: yeah
MOT: I like the animal
MOT: who you gonna call
CHI: I don't know but this phone don't work
CHI: can you get it to
MOT: oh there you go

Appendix C

Table C1. ***Transcripts included in Study 1 by age, gender, and corpora.***

Age	Gender	Corpus
6;00.14	female	HSLLD
6;00.23	female	HSLLD
6;01.00	male	Bliss
6;01.00	female	Peterson McCabe
6;02.00	male	Warren
6;02.00	female	Warren
6;03.19	female	Gelman
6;06.00	female	Peterson McCabe
6;09.06	male	Gelman
6;09.09	male	Fletcher
6;09.09	male	Fletcher
6;10.00	female	Peterson McCabe
6;10.00	female	Fletcher
6;10.13	male	Gelman
6;11.00	male	Fletcher
6;11.08	female	Fletcher
6;11.20	N/A	Feldman
6;11.21	male	Fletcher
6;11.23	male	HSLLD
6;11.9	female	Fletcher
7;00.14	female	Fletcher
7;00.21	male	HSLLD
7;00.28	female	Fletcher
7;00.29	male	Fletcher
7;01.19	male	Fletcher
7;02.04	female	Fletcher
7;02.12	male	Fletcher
7;02.24	male	Fletcher
7;03.00	female	Peterson McCabe
7;03.03	male	Fletcher
7;03.26	female	Fletcher
7;04.22	female	HSLLD
7;04.22	female	HSLLD
7;05.00	female	Peterson McCabe

7;05.02	male	HSLLD
7;05.08	female	HSLLD
7;06	male	Nippold
7;06.00	female	Peterson McCabe
7;06.00	male	Nippold
7;06.00	male	Nippold
8;02.20	N/A	Feldman
8;05	male	Gillam
8;07.00	male	Nippold
8;07.00	female	Peterson McCabe
8;10.00	female	Peterson McCabe
8;10.30	female	HSLLD
8;11.24	female	HSLLD
8;9.00	male	Nippold
9;00.00	male	Rescorla
9;00.00	male	Rescorla
9;01.00	male	Nippold
9;01.13	male	HSLLD
9;01.13	male	HSLLD
9;02.21	male	HSLLD
9;03.01	female	HSLLD
9;04.00	female	Peterson McCabe
9;04.22	female	HSLLD
9;05.02	female	HSLLD
9;06.00	female	Nippold
9;06.05	male	HSLLD

Table C2. ***Summary descriptions of each corpora represented in the dataset.***

Corpora	# of Transcripts	Description
HSLLD	17	Home Visits as part of the Home-School Study of Language and Literacy Development. These visits included a number of activities, including book reading, mealtime conversations, and toy play.
Bliss	1	Children interacting with an investigator during toy play. Aged-matched controls for SLI population.
Peterson McCabe	0	Dialog between experimenters and children at a nursery school and elementary school in Ohio. Verbal prompts were embedded in conversation while the child worked on a construction project.
Warren	2	Children interacting with either their mothers or their fathers in their home. White, middle-class, non-professional families. Parents were instructed to play with or talk to their children as naturally as possible.
Gelman	3	Mother-child dyads discussing a researcher-created picture book entitled "Who Can…?" designed to elicit conversations about gender.
Fletcher	16	Female experimenter and child when at the child's nursery school. The conversation included a stick-on-game, free/guided conversation, and a balloon picture story.
Feldman	2	Full-term, healthy siblings of children with brain lesions. Used a picture-story book to elicit narrative descriptions.

Nippold	7	Chess players interviewed by an experimenter about their experiences with chess.
Gillam	1	Children look at a picture book and discuss images with an experimenter
Rescorla	2	Children look at a picture book and discuss images with an experimenter

Appendix D

Table D1. ***All utterances marked as ambiguous.***

Speaker	Child Age (months)	Utterance
CHI	82	whoops .
CHI	82	whoops .
CHI	82	oohhhweeee@i .
CHI	81	yeah , they're building xxx .
CHI	84	xxx dusty bin .
CHI	87	I hadta Nick always xxx .
CHI	87	xxx to put it down there .
CHI	87	xxx my uncle .
INV	86	don't hurt me .
CHI	86	don't hurt me .
CHI	86	xxx big mouth .
CHI	83	&~a:h then there was a xxx .
CHI	83	and he xxx .
CHI	83	the xxx .

Appendix E

Study 1: Model Structures & Results

All analyses were conducted in R (version 4.2.1). In all models, the child's age in months was centered on the sample mean. This centering improves interpretability of the model's intercept and main effects in the presence of interactions, because the reference point for age corresponds to the average age in the dataset instead of 0 months, which has no substantive meaning in this context. All categorical predictors were sum coded (-1, 1) so that the main effect of age captured the average effect across the two groups, and the comparison between the two groups is symmetric. The levels of each categorical variable are specified for each model below. P-values were obtained using the *lmerTest* package (which uses Satterthwaite's degrees of freedom method). For linear mixed-effects models, we report 95% profile-likelihood confidence intervals. For logistic models, we report Wald CIs due to numerical instability of profile-likelihood intervals in these models. Additional for logistic models, model estimates and CIs are converted to odds ratios for easier interpretation, so note that for these models, CIs that include 1 indicate that the data are compatible with both increased and decreased odds (comparable to intervals containing 0 for linear regression models).

All Bayesian models were fit using the brms package in R (Bürkner, 2017). For each model, four Markov chains were run for 4,000 iterations each, including 2,000 warmup iterations, yielding 8,000 post-warmup samples. All models were fit using the same random seed (713) to ensure full reproducibility of results. Linear models assumed a Gaussian likelihood, and logistic models used a Bernoulli likelihood with a logit link function. Because these analyses were largely exploratory, we specified weakly informative priors for fixed and random effects. These priors were chosen to regularize parameter estimates and improve computational stability while avoiding strong substantive assumptions. Priors were scaled to reflect the range and distribution of each outcome variable.

Proportion of Storytelling by Speaker

To detect differences in the amount of storytelling for different speakers, we used the logistic regression function below. We included a random intercept and slope for the Speaker Group by Transcript to account for variability in both baseline likelihood of storytelling in a transcript, as well as variability in the differences in prevalence of storytelling between adult and child speakers in each transcript.

Table E1. ***Frequentist model predicting story coding from the interaction between age and speaker group.***

Fixed Effect	Estimate	Odds Ratio	SE	*z*-value	*p*-value	95% CI (Odds Ratio)
Intercept	–0.963	0.381	0.111	–8.704	$< 2 \times 10^{-16}$	[0.307, 0.473]
Age(centered)	0.009	1.001	0.008	1.080	.280	[0.992, 1.027]
Speaker	–1.100	0.333	0.137	–8.022	8×10^{-16}	[0.254, 0.436]
Age : Speaker	–0.009	0.991	0.011	–0.848	.397	[0.970, 1.012]

Story Coding (0 = out-of-story, 1 = in-story) ~ Child's Age in Months * Speaker Group (–1 = Child, 1 = Other Speakers) + (Speaker Group | Transcript)

Table E2. ***Random effects for frequentist model predicting story coding from the interaction between age and speaker group.***

Group	Effect	Variance	SD	Correlation
Subject	Intercept	0.672	0.820	
	Speaker	1.063	1.031	–0.13

Table E3. ***Bayesian model predicting story coding from the interaction between age and speaker group.***

Effect	Estimate	95% CrI	Odds Ratio	95% CI (Odds Ratio)
Fixed Effects:				
Intercept	–0.96	[-1.19, -0.74]	0.38	[0.30, 0.48]
Age (Centered)	0.01	[-0.01, 0.03]	1.01	[0.99, 1.03]
Speaker	–1.09	[-1.37, -0.82]	0.34	[0.25, 0.44]
Age : Speaker	–0.01	[-0.03, 0.01]	0.99	[0.97, 1.01]
Random Effects:				
SD Intercept	0.86	[0.70, 1.05]		
SD Speaker slope	1.08	[0.88, 1.33]		
Correlation	–0.12	[-0.39, 0.17]		

Priors: β ~ Normal(0,1); Intercept ~ Normal(0,1.5); SD parameters ~ Exponential(1). All models converged successfully (all $\hat{R} = 1.00$).
Formula: Story Coding ~ Age (Centered) * Speaker Group + (Speaker Group | Transcript ID)

The Bayesian model results were consistent in direction and magnitude with frequentist models.

Number of Words Contributed to a Story by Child Age

The length of children's stories as they got older was assessed using the following two linear regression models, including a random intercept for each transcript to account for variability in baseline rates of story length across children. The first model only has the linear term for age, while the second model includes a quadratic term to test for the hypothesized U-Shaped curve for story length. The fit of the two models were then compared using a likelihood ratio test.

Table E4. ***Frequentist linear model predicting word count in story from age.***

Fixed Effect	Estimate	SE	df	*t*-value	*p*-value	95% CI
Intercept	71.098	17.809	47.256	3.992	.0002	[36.288, 106.035]
Age (centered)	3.733	1.414	47.624	2.641	.011	[0.972, 6.508]

Child's Word Count in Story ~ Child's Age in Months + (1 | Transcript)

Table E5. ***Random effects for frequentist linear model predicting word count in story from age.***

Random Effects	Variance	SD
Transcript (Intercept)	18385	135.6
Residual	2430	49.3

Quadratic Model:

Table E6. ***Frequentist quadratic model predicting word count in story from age.***

Fixed Effect	Estimate	SE	df	*t*-value	*p*-value	95% CI
Intercept	68.770	27.711	46.71	2.482	.017	[15.108, 122.476]
Age (centered)	3.646	1.641	47.232	2.222	.031	[0.469, 6.828]
Age^2	0.015	0.133	46.744	0.112	.911	[-0.243, 0.273]

Child's Word Count in Story ~ Child's Age in Months + (Child's Age in Months)^2 + (1 | Transcript)

Table E7. ***Random effects for frequentist quadratic model predicting word count in story from age.***

Random Effects	Variance	SD
Transcript (Intercept)	18784	135.05
Residual	2429	49.29

Table E8. ***Comparison of linear and quadratic models predicting word count in story from age.***

	AIC	BIC
Linear Model	7737.1	7755.3
Quadratic Model	7739.1	7761.9

Table E9. ***Bayesian linear model predicting story length from age.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	67.80	[36.01, 100.51]
Age (Centered)	2.51	[0.15, 4.72]
Random Effects:		
SD Intercept (TranscriptID)	133.47	[108.87, 162.81]
Residual Variation		
Sigma	49.38	[46.72, 52.28]

Formula: Story Length ~ Age (Centered) + (1|TranscriptID)
Priors: Age ~ Normal(0,2); Intercept ~ Normal(40,50); SD parameters ~ Exponential(0.02).

Table E10. ***Bayesian quadratic model predicting story length from age.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	61.44	[21.68, 102.53]
Age (Centered)	2.34	[-0.05, 4.62]
Age^2 (Centered)	0.03	[-0.12, 0.18]
Random Effects:		
SD Intercept (TranscriptID)	134.20	[108.59, 164.49]
Residual Variation		
Sigma	49.38	[46.73, 52.18]

Formula: Story Length ~ Age (Centered) + Age^2 + (1|TranscriptID)
Priors: Age ~ Normal(0,2); Age^2 ~ Normal(0, 0.1), Intercept ~ Normal(40,50);
SD parameters ~ Exponential(0.02).

The Bayesian model results were largely consistent in direction and magnitude with frequentist models, except the credibility interval of the linear effect in the quadratic model did include 0. This, however, does not change our conclusions, as we were mainly interested in the parabolic predictor.

Length of In-Story Conversational Turns by Speaker

The lengths of in-story conversational turns were assessed using the following linear regression model. Originally, the model structure included a random slope for the Speaker group by Transcript, however this returned a singular fit. Therefore, the slope was removed.

Table E11. ***Frequentist model predicting turn length from the interaction between age and speaker group.***

Fixed Effect	Estimate	SE	df	*t*-value	*p*-value	95% CI
Intercept	32.15	4.612	58.143	6.971	3×10^{-9}	[23.126, 41.188]
Age (centered)	1.213	0.364	59.820	3.328	.002	[0.499, 1.926]
Speaker	–3.40	0.866	2799.75	–3.926	.00008	[-5.109, -1.709]
Age : Speaker	–0.185	0.077	2809.11	–2.384	.017	[-0.337, -0.033]

Number of Words in a Turn ~ Child's Age in Months * Speaker Group (-1 = Child, 1 = Other Speakers) + (1 | Transcript)

Table E12. ***Random effects for frequentist model predicting turn length from the interaction between age and speaker group.***

Random Effects	Variance	SD
Transcript (Intercept)	1218	34.90
Residual	1285	35.84

To explore this interaction, we used the *emmeans* package to estimate the simple slopes of age for both children and other speakers.

Table E13. ***Simple slopes of age on turn length for each speaker group.***

	Trend	SE	df	95% CI
Speaker Group = Other	1.03	0.380	71.9	[0.27, 1.79]
Speaker Group = Child	1.40	0.365	60.6	[0.668, 2.13]

Table E14. ***Bayesian model predicting turn length from the interaction between age and speaker group.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	19.25	[15.12, 23.47]
Age (Centered)	0.81	[0.39, 1.24]
Speaker	–3.69	[–6.90, –0.37]
Age : Speaker	–0.62	[–1.05, –0.19]
Random Effects:		
SD Intercept (TranscriptID)	20.5	[16.19, 25.59]
SD Speaker Slope	20.05	[15.63, 25.33]
Correlation	–0.98	[–1.00, –0.95]
Residual Variation:		
Sigma	34.74	[33.86, 35.66]

Formula: Turn Length ~ Child Age (Centered) * Speaker + (Speaker | Transcript)
Priors: β ~ normal(0,2), intercept ~ normal(30,50), sigma~exponential(0.02).

The direction and magnitude of the effects in the Bayesian model were consistent with the frequentist models.

Mean Length of Children's Utterances

To assess how the length of the children's utterances changed with age, we fit the following linear regression model with a random intercept and slope for story coding by transcript.

Table E15. ***Frequentist model predicting child utterance length from the interaction between age and story coding.***

Fixed Effect	Estimate	SE	df	*t*-value	*p*-value	95% CI
Intercept	5.003	0.151	58.06	33.20	$< 2 \times 10^{-16}$	[4.708, 5.299]
Age (centered)	0.058	0.012	59.66	4.831	.00001	[0.034, 0.081]
Story Coding	1.934	0.103	54.53	18.699	$< 2 \times 10^{-16}$	[1.732, 2.139]
Age : Story Coding	0.024	0.008	57.04	2.970	0.004	[0.008, 0.041]

Length of Child Utterance ~ Age of Child in Months * Story Coding (-1 = out-of-story, 1 = in-story) + (Story Coding | Transcript)

Table E16. ***Random effects for frequentist model predicting child utterance length from the interaction between age and story coding.***

Group	Effect	Variance	SD	Correlation
Transcript	Intercept	1.173	1.083	
	Story Coding	0.440	0.664	0.60
Residual		17.906	4.232	

To explore this interaction, we used the *emmeans* package to estimate the simple slope of age both in and out-of-story:

Table E17. ***Simple slopes of age on child utterance length by story coding.***

	Trend	SE	df	95% CI
Story Coding = in	0.082	0.018	58.4	[0.047, 0.117]
Story Coding = out	0.0329	0.011	53.4	[0.011, 0.055]

Table E18: ***Bayesian model predicting utterance length from the interaction between age and story coding.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	5.00	[4.69, 5.31]
Age (Centered)	0.06	[0.03, 0.08]
Story Coding	1.93	[1.73, 2.14]
Age : Story Coding	0.02	[0.01, 0.04]
Random Effects:		
SD Intercept (TranscriptID)	1.11	[0.90, 1.38]
SD Story Coding Slope	20.05	[0.52, 0.88]
Correlation	0.57	[0.30, 0.78]
Residual Variation:		
Sigma	4.23	[4.17, 4.30]

Formula: Length of Utterance ~ Age (Centered) * Story Coding (in/out) + (Story Coding|TranscriptID)
Priors: Age ~ normal(0,0.2), Story Coding ~ normal(0,2), Age:StoryCoding ~ normal(0,0.2), intercept ~ normal(6,5), sigma~exponential(1).

The direction and magnitude of the Bayesian estimations are consistent with the frequentist analysis.

Unique Conjunctions per Utterance by Age and Story-Coding

The number of unique conjunctions in a child's utterance was predicted using the following linear regression model, which includes a random intercept and slope for story coding by transcript. These random effects account for different rates of unique conjunctions for each child, as well as variability rates of conjunctions per in-story and out-of-story utterances per child.

Table E19. ***Frequentist model predicting unique conjunctions per utterance from the interaction between age and story coding.***

Fixed Effect	Estimate	SE	df	*t*-value	*p*-value	95% CI
Intercept	0.596	0.0278	58.426	21.447	$< 2 \times 10^{-16}$	[0.542, 0.651]
Age (centered)	0.009	0.002	59.823	4.067	.0001	[0.005, 0.013]
Story Coding	0.306	0.019	55.496	16.245	$< 2 \times 10^{-16}$	[0.269, 0.342]
Age : Story Coding	0.004	0.002	58.03	2.334	.023	[0.0005, 0.0065]

Number of Unique Conjunctions in Utterance ~ Child's Age in Months * Story Coding (–1 = out-of-story, 1 = in-story) + (Story Coding | Transcript)

Table E20. ***Random effects for frequentist model predicting unique conjunctions per utterance from the interaction between age and story coding.***

Group	Effect	Variance	SD	Correlation
Transcript	Intercept	0.039	0.198	
	Story Coding	0.013	0.117	0.64
Residual		0.701	0.837	

To explore this interaction, we used the *emmeans* package to estimate the simple slope of age both in and out-of-story.

Table E21. ***Simple slopes of age on unique conjunctions per utterance by story coding.***

	Trend	SE	df	95% CI
Story Coding = in	0.012	0.003	58.4	[0.006, 0.019]
Story Coding = out	0.005	0.002	52.4	[0.001, 0.009]

Table E22. ***Bayesian model predicting unique conjunctions per utterance from the interaction between age and story coding.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	0.597	[0.541, 0.652]
Age (Centered)	0.009	[0.004, 0.013]
Story Coding	0.306	[0.268, 0.344]
Age : Story Coding	0.004	[0.0004, 0.0066]
Random Effects:		
SD Intercept (TranscriptID)	0.204	[0.163, 0.253]
SD Story Coding Slope	0.12	[0.088, 0.157]
Correlation	0.609	[0.314, 0.828]
Residual Variation:		
Sigma	0.838	[0.824, 0.851]

Formula: Number of Unique Conjunctions ~ Child's Age (Centered) * Story Coding + (Story Coding | Transcript ID)
Priors: Age ~ normal(0,0.2), Story Coding ~ normal(0,2), Age:StoryCoding ~ normal(0,0.2), intercept ~ normal(1,1), sigma~exponential(1).

The direction and magnitude of effects of Bayesian estimation were consistent with the frequentist model.

Total Conjunctions per Utterance by Age and Story-Coding

The total count of conjunctions in a child's utterance was predicted using the following linear regression model, with a random intercept and slope for story coding by transcript.

Table E23. ***Frequentist model predicting total conjunctions per utterance from the interaction between age and story coding.***

Fixed Effect	Estimate	SE	df	t-value	p-value	95% CI
Intercept	0.665	0.032	58.215	20.885	$< 2 \times 10^{-16}$	[0.603, 0.728]
Age (centered)	0.010	0.002	59.68	4.220	.00008	[0.006, 0.016]
Story Coding	0.359	0.022	54.716	15.701	$< 2 \times 10^{-16}$	[0.315, 0.405]
Age : Story Coding	0.005	0.002	57.032	2.683	.010	[0.001, 0.008]

Total Conjunctions in Utterance ~ Age of Child in Months * Story Coding (–1 = out-of-story, 1 = in-story) + (Story Coding | Transcript)

Table E24. ***Random effects for frequentist model predicting total conjunctions per utterance from the interaction between age and story coding.***

Group	Effect	Variance	SD	Correlation
Transcript	Intercept	0.050	0.224	
	Story Coding	0.020	0.141	0.74
Residual		1.089	1.043	

To explore this interaction, we used the *emmeans* package to estimate the simple slope of age both in and out-of-story:

Table E25. ***Simple slopes of age on total conjunctions per utterance by story coding.***

	Trend	SE	df	95% CI
Story Coding = in	0.016	0.004	58.2	[0.008, 0.023]
Story Coding = out	0.006	0.002	50.3	[0.001, 0.010]

Table E26. ***Bayesian model predicting total conjunctions per utterance from the interaction between age and story coding.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	0.665	[0.60, 0.728]
Age (Centered)	0.011	[0.005, 0.016]
Story Coding	0.36	[0.031, 0.407]
Age : Story Coding	0.005	[0.001, 0.009]
Random Effects:		
SD Intercept (TranscriptID)	0.229	[0.183, 0.284]
SD Story Coding Slope	0.145	[0.105, 0.193]
Correlation	0.707	[0.435, 0.900]
Residual Variation:		
Sigma	1.043	[1.028, 1.06]

Formula: Total Conjunctions ~ Child's Age (Centered) * Story Coding + (Story Coding | Transcript ID)
Priors: Age ~ normal(0,0.2), Story Coding ~ normal(0,2), Age:StoryCoding ~ normal(0,0.2), intercept ~ normal(1,1), sigma~exponential(1).

The direction and magnitude of effects of Bayesian estimation were consistent with the frequentist model.

Rate of Conjunctions per Word for Children's stories

To evaluate how the rate of conjunctions per word in a child's story changed with age, we fit the following linear regression model, with a random intercept for transcript.

Table E27. ***Frequentist model predicting rate of conjunctions per word from age.***

Fixed Effect	Estimate	SE	df	*t*-value	*p*-value	95% CI
Intercept	0.132	0.005	59.78	26.769	2×10^{-16}	[0.121, 0.140]
Age (centered)	0.0006	0.0004	66.51	1.508	.136	[-0.0001, 0.0014]

(Total Conjunctions in a Story / Total number of words said by child in the story) ~ Child Age in Months + (1|Transcript)

Table E28. ***Random effects for frequentist model predicting rate of conjunctions per word from age.***

Random Effects	Variance	SD
Transcript (Intercept)	0.0008	0.029
Residual	0.005	0.073

Table E29. ***Bayesian model predicting rate of conjunctions per word from age.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	0.130	[0.121, 0.140]
Age (Centered)	0.0006	[-0.0002, 0.0014]
Random Effects:		
SD Intercept (TranscriptID)	0.030	[0.022, 0.039]
Residual Variation:		
Sigma	0.074	[0.070, 0.078]

Priors: Age ~ normal(0,0.01), Intercept ~ normal(0.1,0.1), sigma~exponential(10).
Formula: (Total Conjunctions in Story / Total Number of Child Words) ~ Child Age + (1|Transcript)

The direction and magnitude of effects of Bayesian estimation were consistent with the frequentist model.

Mean Utterance Concreteness Score

To assess how the mean concreteness of children's utterances changed with age, we fit the following linear regression model, with random intercept and slope for story coding by transcript.

Table E30. ***Frequentist model predicting mean utterance concreteness from the interaction between age and story coding.***

Fixed Effect	Estimate	SE	df	*t*-value	*p*-value	95% CI
Intercept	2.236	0.021	52.77	104.678	$< 2 \times 10^{-16}$	[2.195, 2.278]
Age (centered)	–0.002	0.002	53.95	–1.248	.218	[–0.005, 0.001]
Story Coding	0.155	0.022	55.93	7.116	2×10^{-9}	[0.112, 0.197]
Age : Story Coding	0.001	0.002	57.073	0.589	.558	[–0.002, 0.004]

Average Concreteness Score of Words in the Utterance ~ Age of Child in Months * Story Coding (-1 = out-of-story, 1 = in-story) + (Story Coding | Transcript)

Table E31. ***Random effects for frequentist model predicting mean utterance concreteness from the interaction between age and story coding.***

Group	Effect	Variance	SD	Correlation
Transcript	Intercept	0.020	0.141	
	Story Coding	0.021	0.144	–0.81
Residual		0.755	0.869	

Table E32. ***Bayesian model predicting mean utterance concreteness from the interaction between age and story coding.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	2.238	[2.194, 2.280]
Age (Centered)	–0.002	[–0.005, 0.001]
Story Coding	0.154	[0.110, 0.197]
Age : Story Coding	0.001	[–0.002, 0.004]
Random Effects:		
SD Intercept (TranscriptID)	0.144	[0.109, 0.186]
SD Story Coding Slope	0.148	[0.110, 0.189]
Correlation	–0.774	[–0.929, –0.545]
Residual Variation:		
Sigma	0.869	[0.856, 0.883]

Priors: Age ~ normal(0,0.2), Story Coding ~ normal(0,1), Age : Story Coding ~ normal(0,0.2), Intercept ~ normal(3,1), sigma~exponential(1).
Formula: Average Concreteness Score of Utterance ~ Child Age * Story Coding + (Story Coding|Transcript)

The direction and magnitude of effects of bayesian estimation were consistent with the frequentist model.

Average SD of Concreteness within an Utterance

To assess how the standard deviation of concreteness scores with the children's utterances changed with age, we fit the following linear regression model with a random intercept and slope for story coding by transcript:

Table E33. ***Frequentist model predicting SD of utterance concreteness from the interaction between age and story coding.***

Fixed Effect	Estimate	SE	df	*t*-value	*p*-value	95% CI
Intercept	2.290	0.042	58.664	54.751	$< 2 \times 10^{-16}$	[2.208, 2.372]
Age (centered)	0.015	0.004	60.444	4.518	.00003	[0.008, 0.021]
Story Coding	0.560	0.033	56.053	17.121	$< 2 \times 10^{-16}$	[0.496, 0.624]
Age : Story Coding	0.003	0.003	58.712	1.323	.191	[–0.002, 0.008]

Standard deviation of concreteness score of words in the utterance ~ Age of Child in Months * Story Coding (-1 = out-of-story, 1 = in-story) + (Story Coding | Transcript)

Table E34. ***Random effects for frequentist model predicting SD of utterance concreteness from the interaction between age and story coding.***

Group	Effect	Variance	SD	Correlation
Transcript	Intercept	0.092	0.303	
	Story Coding	0.051	0.226	0.03
Residual		1.093	1.046	

Table E35. ***Bayesian model predicting SD of utterance concreteness from the interaction between age and story coding.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	2.291	[2.202, 2.375]
Age (Centered)	0.015	[0.008, 0.022]
Story Coding	0.558	[0.490, 0.624]
Age : Story Coding	0.003	[-0.002, 0.008]
Random Effects:		
SD Intercept (TranscriptID)	0.313	[0.255, 0.386]
SD Story Coding Slope	0.233	[0.183, 0.292]
Correlation	0.037	[-0.251, 0.311]
Residual Variation:		
Sigma	1.046	[1.029, 1.065]

Priors: Age ~ normal(0,0.2), Story Coding ~ normal(0,1), Age : Story Coding ~ normal(0,0.2), Intercept ~ normal(3,1), sigma~exponential(1).
Formula: SD Score of Utterance ~ Child Age * Story Coding + (Story Coding|Transcript)

The direction and magnitude of effects of bayesian estimation were consistent with the frequentist model.

Rate of First-Person Utterances

To assess how often children used first person pronouns by age, we fit the following logistic regression model on their utterances, with random intercept and slope for story coding by transcript:

Table E36. ***Frequentist model predicting rate of first-person utterances from the interaction between age and story coding.***

Fixed Effect	Estimate	Odds Ratio	SE	*z*-value	*p*-value	95% CI (Odds Ratio)
Intercept	-1.056	0.348	0.075	–14.064	$< 2 \times 10^{-16}$	[0.300, 0.403]
Age (centered)	0.007	1.007	0.006	1.220	.222	[0.996, 1.019]
Story Coding	0.428	1.534	0.076	–5.662	2×10^{-8}	[1.323, 1.779]
Age : Story Coding	-0.014	0.986	0.006	–2.370	.018	[0.974, 0.997]

Whether the utterance contained first person (0 = no first person, 1 = first person) ~ Age of Child in Months * Story Coding (-1 = out-of-story, 1 = in-story) + (Story Coding | Transcript)

Table E37. ***Random effects for frequentist model predicting rate of first-person utterances from the interaction between age and story coding.***

Group	Effect	Variance	SD	Correlation
Transcript	Intercept	0.266	0.516	
	Story Coding	0.271	0.520	0.54

To explore this interaction, we used the *emmeans* package to estimate the simple slope of age both in and out-of-story:

Table E38. ***Simple slopes of age on rate of first-person utterances by story coding.***

	Trend	SE	95% CI
Story Coding = in	–0.007	0.01	[–0.027, 0.013]
Story Coding = out	0.021	0.006	[0.009, 0.033]

Table E39. ***Bayesian model predicting rate of first-person utterances from the interaction between age and story coding.***

Effect	Estimate	95% CrI	Odds Ratio	95% CI (Odds Ratio)
Fixed Effects:				
Intercept	–1.06	[–1.22, –0.90]	0.35	[0.30, 0.41]
Age (Centered)	0.007	[–0.005, 0.019]	1.01	[1.00, 1.02]
Speaker	0.43	[0.27, 0.59]	1.54	[1.31, 1.80]
Age : Speaker	–0.014	[–0.027, –0.002]	0.99	[0.973, 0.998]
Random Effects:				
SD Intercept	0.544	[0.430, 0.693]		
SD Speaker slope	0.547	[0.430, 0.693]		
Correlation	0.495	[0.194, 0.733]		

Priors: β ~ Normal(0,1); Intercept ~ Normal(0,1.5); SD parameters ~ Exponential(1). All models converged successfully (all $\hat{R} = 1.00$).
Formula: First Person ~ Age (Centered) * Story Coding + (Story Coding | Transcript ID)

The direction and magnitude of effects of Bayesian estimation were consistent with the frequentist model.

Caregiver Repetition of Child's Prior Utterances

Table E40. ***Frequentist model predicting caregiver repetition from the interaction between age and story coding.***

Fixed Effect	Estimate	SE	df	*t*-value	*p*-value	95% CI
Intercept	0.178	0.011	52.93	16.76	$<2 \times 10^{-16}$	[0.157, 0.199]
Age (centered)	–0.0005	0.008	54.59	–0.54	0.594	[–0.002, 0.001]
Previous Line Story Coding	0.031	0.005	43.06	5.89	5×10^{-7}	[0.020, 0.041]
Age : Previous Line Story Coding	–0.00001	0.0004	48.18	–0.031	0.976	[–0.0008, 0.0008]

Proportion of caregiver words repeating from prior 3 child utterances ~ Age of Child in Months * Previous Child Line Story Coding (-1= out-of-story, 1 = in-story) + (Previous Child Line Story Coding | Transcript)

Table E41. ***Random effects for frequentist model predicting caregiver repetition from the interaction between age and story coding.***

Group	Effect	Variance	SD	Correlation
Transcript	Intercept	0.006	0.078	
	Previous Line Story Coding	0.0008	0.029	0.83
Residual		0.070	0.265	

Table E42. ***Bayesian model predicting caregiver repetition from the interaction between age and story coding.***

Parameter	Estimate	95% CrI
Fixed Effects:		
Intercept	0.178	[0.157, 0.200]
Age (Centered)	–0.0004	[–0.002, 0.001]
Story Coding	0.030	[0.020, 0.041]
Age : Story Coding	–0.00001	[–0.0008, 0.0008]
Random Effects:		
SD Intercept (TranscriptID)	0.079	[0.064, 0.098]
SD Story Coding Slope	0.029	[0.018, 0.041]
Correlation	0.78	[0.491, 0.967]
Residual Variation:		
Sigma	0.265	[0.261, 0.269]

Priors: Age ~ normal(0,0.01), Story Coding ~ normal(0,0.2), Age : Story Coding ~ normal(0,0.01), Intercept ~ normal(0.3,0.2), sigma~exponential(10).
Formula: SD Score of Utterance ~ Child Age * Story Coding + (Story Coding|Transcript)

The direction and magnitude of effects of Bayesian estimation were consistent with the frequentist model.

Rate of Caregiver Questions following Child Utterances

Table E43. ***Frequentist model predicting caregiver question rate from the interaction between age and story coding.***

Fixed Effect	Estimate	Odds Ratio	SE	*z*-value	*p*-value	95% CI (Odds Ratio)
Intercept	0.351	1.420	0.127	2.754	0.006	[1.106, 1.823]
Age (centered)	–0.020	0.980	0.010	–2.033	0.042	[0.961, 0.999]
Previous Line Story Coding	–0.033	0.967	0.055	–0.605	0.545	[0.868, 1.078]
Age : Previous Line Story Coding	0.007	1.007	0.004	1.614	0.107	[0.998, 1.016]

Whether caregiver turn contains a question (0 = No Question, 1 = Question) ~ Age of Child * Previous Child Line Story Coding (–1 = out-of-story, 1 = in-story) + (Previous Child Line Story Coding | Transcript)

Table E44. ***Random effects for frequentist model predicting caregiver question rate from the interaction between age and story coding.***

Group	Effect	Variance	SD	Correlation
Transcript	Intercept	0.882	0.939	
	Previous Line Story Coding	0.092	0.304	0.26

Table E45. ***Bayesian model predicting caregiver question rate from the interaction between age and story coding.***

Effect	Estimate	95% CrI	Odds Ratio	Odds Ratio CrI
Fixed Effects:				
Intercept	0.350	[0.096, 0.607]	1.42	[1.10, 1.84]
Age (Centered)	-0.021	[-0.041, 0.0004]	0.98	[0.96, 1.00]
Speaker	-0.035	[-0.147, 0.077]	0.97	[0.86,1.08]
Age : Speaker	0.007	[-0.002, 0.0165]	1.007	[0.998, 1.017]
Random Effects:				
SD Intercept	0.982	[0.802,1.21]		
SD Speaker slope	0.323	[0.217, 0.45]		
Correlation	0.228	[-0.155, 0.563]		

Priors: β ~ Normal(0,1); Intercept ~ Normal(0,1.5); SD parameters ~ Exponential(1). All models converged successfully (all $\hat{R} = 1.00$).
Formula: Question ~ Age (Centered) * Story Coding + (Story Coding | Transcript ID)

The direction and magnitude of the effects are largely consistent between the bayes and frequentist models. However, the credibility interval for the effect of age on question frequency does include 1 in this model, indicating further uncertainty in the robustness of this effect.

Proportion of Story Beginnings after Caregiver Questions

To calculate how the rate of caregiver questions prompting story beginnings changed with age, we separated the data into all the child's initial utterances of a conversational turn. We then fit a logistic regression model using the following equation. We initially included a random intercept and slope for the relationship between the presence of questions in a turn within each transcript, however, this model did not converge. We therefore dropped the slope, leaving the intercept to account for both variability in the baseline rate of story beginnings per transcript.

Table E46. ***Frequentist model predicting story beginnings from the interaction between age and caregiver questions.***

Fixed Effect	Estimate	Odds Ratio	SE	*z*-value	*p*-value	95% CI (Odds Ratio)
Intercept	–2.602	0.074	0.093	–27.882	$<2 \times 10^{-16}$	[0.062, 0.088]
Age (centered)	0.015	1.014	0.007	2.020	.043	[1.0004, 1.0291]
Question	0.375	1.456	0.064	5.835	5×10^{-9}	[1.283, 1.651]
Age : Question	–0.004	0.996	0.005	–0.810	0.417	[0.985, 1.006]

Whether the following child utterance starts a story (0 = No Story Beginning, 1 = Story Beginning) ~ Child's Age in Months * Whether Prior turn Contains Question (-1 = no Question, 1 = Question) + (1 | Transcript)

Table E47. ***Random effects for frequentist model predicting story beginnings from the interaction between age and caregiver questions.***

Random Effects	Variance	SD
Transcript (Intercept)	0.248	0.497

Table E48. ***Bayesian model predicting story beginnings from the interaction between age and caregiver questions.***

Effect	Estimate	95% CrI	Odds Ratio	Odds Ratio CrI
Fixed Effects:				
Intercept	−2.81	[−3.05, −2.58]	0.06	[0.05, 0.08]
Age (Centered)	0.011	[−0.007, 0.029]	1.01	[0.99, 1.03]
Question	0.38	[0.188, 0.596]	1.46	[1.21, 1.82]
Age : Question	−0.009	[−0.025, 0.006]	0.99	[0.98, 1.01]
Random Effects:				
SD Intercept	0.514	[0.191, 0.823]		
SD Speaker slope	0.283	[0.020, 0.575]		
Correlation	−0.46	[−0.972, 0.658]		

Priors: β ~ Normal(0,1); Intercept ~ Normal(0,1.5); SD parameters ~ Exponential(1). All models converged successfully (all $\hat{R} = 1.00$).
Formula: Story Beginning ~ Age (Centered) * Prior Turn Contains Question + (Prior Utterance Turn Question | Transcript ID)

The 95% credible interval for the age effect included 1, indicating greater uncertainty in this estimate relative to the frequentist model. However, the effect of questions predicting more story beginnings was similar in direction and magnitude across approaches. The interaction term was likewise consistent in direction and small in magnitude in both models.

License