P2PSTORY: Dataset of Children as Storytellers and Listeners in Peer-to-Peer Interactions

Nikhita Singh
MIT Media Lab
Cambridge, USA
nikhita@media.mit.edu

Jin Joo Lee
MIT Media Lab
Cambridge, USA
jinjoo@media.mit.edu

Ishaan Grover
MIT Media Lab
Cambridge, USA
igrover@media.mit.edu

Cynthia Breazeal
MIT Media Lab
Cambridge, USA
cynthiab@media.mit.edu

ABSTRACT

Understanding social-emotional behaviors in storytelling interactions plays a critical role in the development of interactive educational technologies for children. A challenge when designing for such interactions using technology like social robots, virtual agents, and tablets is understanding the social-emotional behaviors pertinent to storytelling—especially when emulating a natural peer-to-peer relation between the child and the technology. We present P2PSTORY, a dataset of young children (5-6 years old) engaging in natural peer-to-peer storytelling interactions with fellow classmates. The dataset consists of rich social behaviors of children without adult supervision, with each participant demonstrating being a storyteller and a listener. The dataset contains 58 video recorded sessions along with a diverse set of behavioral annotations as well as developmental and demographic profiles of each child participant. We describe the main characteristics of the dataset in addition to findings that reveal perceptual differences between adults and children when evaluating the attentiveness of listeners.

ACM Classification Keywords

H.5.2 User Interfaces: Theory and Methods

Author Keywords

Dataset; Children; Storytelling; Listening; Nonverbal Behaviors; Quantitative Methods; Education; Interaction Design

INTRODUCTION

Narrative skills in young children are foundational in shaping language development and predicting future academic success.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CHI 2018, April 21–26, 2018, Montreal, QC, Canada
Copyright © 2018 ACM ISBN 978-1-4503-5620-6/18/04...\$15.00.
<https://doi.org/10.1145/3173574.3174008>

Figure 1: Video-Recordings of Storytelling Interactions. The P2PSTORY dataset includes 58 recorded storytelling interactions of young children. Three time-synchronized cameras captured the frontal-view of each participant along with a bird's-eye view.

Interactive technologies such as tablets, virtual agents, and social robots are actively being developed to support early language learning for children. Given recent studies demonstrating that peer-to-peer learning offers substantial benefits in terms of increasing student confidence, engagement, and communication skills [36], these systems have moved away from acting as tutors to instead engaging children as peers.

Delivering effective narratives in social interactions require successful communication between a speaker and listener through both verbal and nonverbal behaviors [30]. Prior works in interactive storytelling technologies have explored the role of the technology as both story-tellers [34, 37, 10] and story-

listeners [29, 8, 2], utilizing nonverbal cues such as eye blinking, gaze, and nods to enhance attention, engagement, and recall.

To design the social interactions for such systems, human-computer interaction (HCI) researchers pull from psychology literature and observe real-world dynamics to gain insight into the social-emotional behaviors relevant to the interaction context. Although datasets exist in understanding children's behaviors when interacting with adults, no prior work exists in capturing natural interactions of children amongst peers of their own age.

As such, we present P2PSTORY, a dataset of young children (5-6 years old) engaging in natural peer-to-peer storytelling with fellow classmates. In this paper, we begin by reviewing the social-emotional behaviors pertinent to effective speaker-listener interactions and describe how current educational technologies for children use these nonverbal behaviors to enhance the interaction experience and improve learning outcomes. As these systems move toward emulating a peer-based relation, our P2PSTORY dataset can provide greater understanding of how these nonverbal behaviors manifest in children when they are amongst peers. We review the currently available social interaction datasets which focus on either adult-adult or adult-child interactions and highlight the absence of work in capturing child-child interactions. We then describe our method of data collection for P2PSTORY and detail the available raw and annotated data, as well as the demographic and developmental profiles of participants. Using this data, we present preliminary analysis for understanding the differences between adult and child perceptions when evaluating the attentiveness of a listener in a storytelling interaction.

BACKGROUND

Nonverbal Speaker-Listener Behaviors of Children

Young children develop a range of skills in order to become effective communicators. An important part of this development process involves the acquisition of nonverbal behaviors such as gaze, gesture, facial expressions, etc. The ability to express and understand such behaviors is instrumental for effective communication and can enhance attention, engagement and recall in peer-to-peer interactions.

Speakers express nonverbal cues to engage listeners, negotiate turn-taking, garner trust, and provide additional communicative context. Children develop the ability to utilize gaze and gestures for negotiating turn-taking at an early age, even before significant verbal expression [14]. Gaze allows a speaker to communicate syntactic and semantic signals, affect, and interpersonal attitude while speaking [26]. Additionally, studies have shown that increased eye contact improves children's recall of stories [28] and positively influences the perception of the speaker [6]. Speakers also produce facial expressions and gestures in parallel with speech as a means of conveying supporting information and disambiguating information when there is a breakdown of speech signal [7]. Furthermore, modulating speech-based characteristics such as increasing pitch variation and speaking rate can enhance the perception of competency [1, 4, 35].

Similarly, listeners use nonverbal behaviors in the form of backchannels for a number of communicative functions in speaker-listener interactions including (a) indication of attention, interest and understanding (or lack thereof), (b) repair or clarification of message, (c) sentence completion and (d) maintenance and regulation of interaction [17, 12, 9, 25]. Backchanneling behavior can take several forms including gesture-based backchannels (postural shifts, gazing patterns), non-lexical backchannels (short vocalized sounds like “uh-huh” and “hmm”), phrasal backchannels (simple words like “wow” and “got it”), and substantive backchannels (clarification questions). Children's backchannel responses are slow to develop and increase substantially with age. A prior study [18] found the frequency of backchannel behaviors increased threefold between the ages of 7 to 11. Older children develop the ability to respond to more contexts while requiring fewer cueing signals from the speaker [18]. In adult-child interactions, older children were observed to respond using nonverbal behaviors like gaze, smiles, and nods more frequently than verbal ones such as “uh huh” and “yes” [25].

Applications in Educational Interactive Technologies

Story-listening technologies have drawn inspiration from human-like nonverbal behaviors to support children's narratives. StoryMat, an interactive listening system in the form of a quilt, had children tell stories to stuffed animals while the mat recorded the voice and movement of objects for playback later on. When using this interactive system, children were found to partake in more advanced narratives compared to a passive mat [8]. In SAGE [2], a collaborative storytelling platform using a computer and interactive stuffed animal, children shared stories with a wise “sage” and its rabbit assistant. During these interactions, the stuffed rabbit backchanneled using nonverbal behaviors such as body movements and eye blinks to show engagement. In [29], a robotic learning companion attentively listened to children telling stories. The backchanneling robot predicted opportunities for providing listener responses based on the child's speaker cues. A contingently responding robot was perceived by children storytellers as more attentive and interested compared to a non-contingent robot.

Story-telling technologies have primarily focused on developing adaptive, personalized, and expressive behaviors to achieve greater learning outcomes. Embodied conversational agents (ECA) that told stories more linguistically advanced than that of the child were perceived to be peer-like to the participants, whose narratives also became more sophisticated after interaction [34]. Similarly in [37], when having storytelling sessions with a robot that adapted its stories to the child's abilities, children used more diverse words in their later stories compared to children who interacted with an unadaptive robot. Highly expressive and emotive robot storytellers (that use body movements, contextual sounds, or speech with a wide range of intonation) led to better recall of narratives compared to static or vocally-flat robots [21, 10]. These studies demonstrate how nonverbal behaviors when appropriately designed in interactive educational experiences can improve children's language learning as well as effect their perceptions about the social technology.

Available Datasets on Speaker-Listener Interactions

Datasets of dyadic social interactions are a crucial part of understanding and designing for the way people engage with social technologies. Currently available datasets that target the understanding of human social-emotional communication primarily capture interactions between adults or adults & children, with no prior work that capture child-child interactions.

Adult-adult interactions have been extensively studied in the ALICO and MultiLis corpora. ALICO aimed to capture the spoken and gestural dynamics of storyteller-listener dialogue [23]. MultiLis focused on identifying individual differences and similarities of listener responses by having three listeners simultaneously interact with the same speaker [11]. To the best of our knowledge, two datasets currently exist that investigate the social interactions of children but through adult-child pairs. The MMDB dataset includes recordings of semi-structured play between a toddler and adult in an interaction context of autism diagnosis [33]. The EmoReact dataset consists of a large number of short video clips of children expressing emotions in reaction to various objects or media in an interaction context with a television director [27].

Though these prior datasets offer rich insight into the social-emotional dynamics of adult-adult and adult-child interactions, they are limited in generalizing to the dynamics of child-child interactions.

P2PSTORY DATASET

The P2PSTORY dataset provides rich social interaction data of children storytelling amongst peers, both from the perspective of the storyteller and the listener. The dataset consists of five major categories of content necessary to understanding the social-emotional behaviors of children in this context.

Audio and Video: Three time-synchronized cameras captured the frontal-view of storyteller's and listener's faces with an additional bird's-eye view capturing a greater view of the dyad (see Figure 1). Also time-synchronized with the video, we captured audio recordings using a high-quality microphone (the MXL AC404 USB conference microphone). The video recording have a resolution of 720x1280 at 30hz and encoded as Photo JPEG and the audio recordings are 16bit at 44.1 khz.

Behavioral Features: Video recordings were coded for a wide range of behaviors in order to enable better understanding of children's social-emotional communication. The following behaviors were selected as they were either found in prior works or commonly observed in the storytelling interaction: gaze, posture, nods, smiles & frowns, eyebrow movement, and backchannel utterances. In addition, interaction-level features were annotated including listener's attention and whether dyads were on or off task.

Prosodic features: Modulating speech properties has been previously shown to enhance the perception of speakers [1, 4, 35]. In accordance with this finding, the storyteller's use of prosodic cues including pitch, energy, pauses, filled pauses, and long utterances was annotated.

Socio-Demographic and Developmental Features: Children's development literature indicates that acquisition of speaker

Figure 2: **ASQ Score Distribution.** Histogram of participants' ASQ scores, which measures a child's social-emotional development.

cues and listener responses continues through to adolescence and often varies from child-to-child [18, 13, 24]. To capture potential causes for variations across children, demographic and socio-emotional development parameters were collected for each participant. Socio-demographic parameters captured include gender, age, birth date, household income, ethnicity, mother's highest education, siblings' ages, and results from the Ages and Stages Questionnaire (ASQ). ASQ is a standardized measure of a child's social-emotional development [3]. The output of the survey is a score where lower values indicate that a child's development is as expected while higher values are indicative of potential need for assessment by a professional. Cumulative ASQ scores are provided as well as raw answers from the questionnaire.

Child Perceptions: To better understand how children perceived the effectiveness of their interaction partner, participants were asked to rate their partner on measures relating to attention and understanding.

The remainder of this section describes the study procedures of our data collection, the annotation scheme, data organization, and some initial behavioral statistics.

Method

Participants

Children of typical development were recruited from a Boston public elementary school whose curriculum already included an emphasis on storytelling. A total of 18 students from a single kindergarten (K2) classroom participated in the study. The average age was 5.22 years-old ($SD = 0.44$) and 61% were male. Overall, 10 participants identified as White, 3 as Black or African American, 2 as Hispanic or Latino, 1 as Asian, 1 as Mixed, and 1 not specified.

The distribution of ASQ scores ($\mu=30.0$ and $SD=22.0$) for our population is provided in Figure 2. As shown, most children were of typical development with a few exceptions.

Storytelling Task

Over a span of five weeks, each participant completed at least three rounds of storytelling with different partners and text-less

Figure 3: **Example Storybook.** Each text-less scene introduces a new character or event. Children developed their own narratives using these illustrations.

Variations	R1	R2	R3	R4
Instructions	heavy	light	light	none
Partner Type	friend	ability lo/hi	mixed-gender	misc
# Story Scenes	1	2	3	3

Table 1: **Round Differences.** Each round varied in the amount of instruction, type of partner, and number of story scenes. For example, the Round 2 had a light amount of instruction, students with high storytelling ability were paired with students with low ability (based on teacher’s suggestions), and two story scenes were used. Round 3 paired participants with a partner of opposite gender, if they had not previously. Round 4 was for redos.

storybooks. The storybooks were a series of colored pictures with illustrated characters and scenes that children could use to craft their own narratives (see Figure 3 for an example storybook). In the first round, students were paired with friends and told a story using only one scene from the storybook. Subsequent rounds varied in the level of instructions, partner type, and in the number of additional story scenes (see Table 1 for details on round variations).

In a *dyad session*, the pair of students took turns narrating a story to their partner with each turn generating a *storytelling episode* (see Figure 4). In total, our data collection consists of 58 storytelling episodes, where the average length of a child’s story was 1 minute and 17 seconds totaling to about 75 minutes of content.

Figure 4: **Storytelling Rounds, Dyads, and Turns.** Each participant had multiple rounds of storytelling with different partners. In a dyad session, a pair of students took turns (T1, T2) narrating their story to their partner. In sum, the data collection consists of 3 rounds (with a supplementary 4th round for redo opportunities) totaling 29 dyad sessions, which equates to 58 individual storytelling episodes.

Study Procedure
A dyad session averaged 15-minutes in duration and consisted of four stages: Instruction, Story Construction, Story Sharing, and Recall & Questionnaire.

Instruction: During the classroom’s free-choice time, pairs of students were brought to the school’s 1-on-1 session office. Students were informed that they are practicing telling stories and good listening skills. The proctor reviewed what constitutes being a good listener based on an existing classroom concept of ‘body still’, ‘voices quiet’, ‘eyes watching’, and ‘ears listening’ (see classroom poster Figure 5). The proctor then described, using an example storybook, how to make up a narrative using the illustrated characters and setting. In addition, the students watched a pre-recorded video of an adult telling an example story. Each student then randomly drew a token from a hat that indicated their storybook selection for the current dyad session.

Story Construction: Randomly paired with one of the participants, proctors went into separate locations (also randomly selected) to help participants generate a story about their book. In these 1-on-1 sessions, proctors focused on the story content (e.g., coming up with character names, generating a narrative with a beginning, middle, end etc.), asking questions to elicit more detail and further the story line. Proctors were provided with list of questions as possible prompters. Depending on a child’s storytelling ability, either none, few, or all the questions were used in a session (see Table 2 for list of questions). The goal of this stage was to help equalize differences in the population’s storytelling ability and to avoid an individual’s difficulty with the task itself hampering the social dynamics. By providing varying degrees of guidance, each participant entered into the next stage ready with a practiced story in-hand.

Story Sharing: In the *story space* (as shown in Figure 6), each participant of the dyad took turns telling their story. In view of both participants, a coin was flipped to randomly determine who went first. At a small table, the first storyteller sat on the *storyteller chair*, and the other participant sat on the *listener chair*. The storybook was placed on a tabletop book stand, easily visible to both participants. The proctor then described their particular roles:

Figure 5: **Classroom poster** on how to be a good listener.

-
- [Q1] What do you want the character's name to be?
 [Q2] Who do you want the main character to be?
 [Q3] Tell me a bit about this character.
 [Q4] Where does your story happen?, Where is this?
 [Q5] What is happening in this picture?
 [Q6] What else do you see in this picture?
 [Q7] Then what happened?
 [Q8] What did [a character] do then?
 [Q9] How does [a character] feel when that happened?
 [Q10] Why do you think [a character] did that?
-

Table 2: **Proctor Prompts.** List of questions used to elicit more details and further a child's story.

Storyteller role: *Your important job as the Storyteller is to be like a teacher. You want to make sure your classmate is paying attention and understands your story.*

Listener role: *Your important job as the Listener is to use your good listening skills and to pay attention because later I'm going to ask you some questions about his/her story.*

Following this, the proctor left the room to help promote natural peer-interactions by removing adult supervision. After the first storytelling episode was complete, the proctor re-entered to instruct the participants to switch seats and briefly reiterate the exchanged roles. The goal of emphasizing the two roles was to establish a common understanding across participants, as this peer-interaction setting may have been unfamiliar to them.

Recall & Questionnaire: At the end of the dyad session, proctors individually asked the participants to recall their partner's story to the best of their ability. Furthermore, we asked them to rate how well their partner was as a storyteller and as a listener. Using a smiley-face based 5-point Likert scale [31] (see Figure 7), we asked the following questions:

1. When you were the storyteller, how was your partner at paying attention to your story?
2. When you were the storyteller, how was your partner at understanding your story?
3. When you were the listener, how was your partner at making sure you were paying attention to the story?
4. When you were the listener, how was your partner at making sure you understood the story?

Behavior Annotation

From the video recordings, coders used a video-annotation software called ELAN [5] to mark the start and stop times for all the behaviors listed in Table 3 except for the prosodic cues. For each storytelling episode, the behaviors of both listener and storyteller were annotated by multiple coders. The coders were instructed to annotate the video of only one participant (either the listener or the storyteller) and only one behavior

(a)

(b)

Figure 6: **Story Space Setup.** (A) Setup included three different camera angles, a high-quality microphone, listener & storyteller chairs, and a story book with compounding story elements per page. (B) We labeled each chair to further emphasize to a child his/her role as either the storyteller or listener.

Category	Labels (<i>default behavior</i>)	S	L	Inter-rater κ
Gaze	<i>picture</i> , partner, away	X	X	0.89
Posture	<i>upright</i> , toward, away, head down, rocking		X	0.40
Nod	<i>none</i> , nod		X	0.89
Eyebrows	<i>neutral</i> , raise, furrow		X	0.51
Mouth	<i>neutral</i> , smile, frown, laugh, other		X	0.34
Utterance	<i>none</i> , "ok", "oh", "so", "then", "yeah", "uh huh", "ok then", "and then", "and they"		X	0.81
Voicing	<i>silence</i> , storyteller's voice, listener's voice, both		joint	0.83
Task	<i>on task</i> , off task		joint	0.05
Attentive State	listening, not listening, speaking-turn		X	0.45
Prosodic Cues	pitch, energy, pause, filled pause, long utterance, clause ending, said, other, not sure	X		0.32

Table 3: **List of All Annotated Behaviors.** The selected set of nonverbal behaviors were either found in prior works or commonly observed in the storytelling interactions. Each annotation category has a set of mutually exclusive labels and was coded for storytellers (S) and/or listeners (L) or jointly evaluated (joint). The Fleiss' kappa values for inter-rater reliability are interrupted as follows: 0.0 to 0.2 indicating slight agreement, 0.21 to 0.40 indicating fair agreement, 0.41 to 0.60 indicating moderate agreement, 0.61 to 0.80 indicating substantial agreement, and 0.81 to 1.0 indicating almost perfect or perfect agreement.

Figure 7: **Smileyometer.** A pictorial rating scale developed for children.

category at a time. Annotators were assigned projects in an order that avoided them having to watch a particular session repeatedly. For the listeners, coders annotated their perceived attentive state as well as their gaze direction, posture shifts, nods, eyebrow movements, smiles & frowns, and backchannel utterances. Storytellers were annotated for their gaze- and prosodic-based speaker cues. For the attentive state annotation, a *listening* label meant that the participant was paying attention to the storyteller's story. To generate this label, annotators were instructed to: *Act as if you are the storyteller. You want the listener to pay attention to your story. Watch the listener and continuously label the moments when you believe he/she is being attentive as well as moments when you want to say to the listener, "Hey, pay attention!"* It is important to note that our state labels include when a listener takes a *speaking-turn* as a mutually exclusive event. This enables us to filter out behaviors related to conversational structure like turn-yielding cues, which have been previously demonstrated to be different from backchannel-inviting cues [15].

While it could be argued that the annotators were primed to have greater saliency for certain listener behaviors, it is important to note that these behaviors were never connected to listener's attention. The instructions to markup specific behaviors (e.g. nods, smiles, gazes) were focused on the *when* and the timing of behaviors. Our upcoming analysis then tries to connect the annotator's perception regarding listener's attention to these marked behaviors.

Based on the *Task* annotation, we further exclude moments from our analyses when both of the children participants are completely off-task from the storytelling activity. Off-task labels identified moments where children participants momentarily detracted from the storytelling task. For example, they were distracted from something in the environment (e.g., a funny poster). The off-task label was used to filter out these moments to discount any emitted nonverbal behaviors during this time (e.g., smiles that are in response to the poster and not story).

We developed a custom program to help coders easily annotate when and what type of prosodic cue was detected in the storyteller's speech. The program played back the audio recording of a storytelling episode, and coders were asked to simulate in real-time being a listener and mark the moments when they want to backchannel by simply tapping on the space bar. After this simulation, coders then reviewed the audio snippets surrounding these moments to reflect on what prompted their backchannel and categorize their reasoning from the following options:

- pitch (intonation in voice, change in tone)
- energy (volume of voice, softness/loudness)
- pause (pause in speech, long silence)
- filled pause (e.g., 'um', 'uh', 'so', 'and')
- long utterance or wordy (a long contiguous speech segment)
- clause ending (end of a sentence/statement)
- said (related to the speech content)
- other
- not sure

This stimulus-based coding is a method for annotators to identify when they wanted to backchannel (backchannel moment) as well as categorize the reasoning behind what caused their wanting to backchannel (speaker cue event). The null-space that is not marked has an implied default label of 'none.' Three coders underwent this simulation.

This simulation method is an extension of prior work [19], which gauges strength of cues based on its ability to elicit

social responses from multiple listeners that all vicariously experience the same speaker. In our work, we add the additional step of categorizing the reasoning (the type of speaker cues: pitch, energy, etc.) behind why the listener chose to backchannel at a given moment. Their annotation acts as a helpful baseline of how an adult listener would have responded at this socially interesting moment (i.e., post a perceived and detectable social cue) to then compare against how the child listener in the real interaction actually responded.

The inter-rater reliability score per behavior category is shown in Table 3. Overall, we achieved a moderate level of agreement (Fleiss' $\kappa = 0.55$, $SD = 0.28$) [16]. Certain behaviors like gaze direction and head nods are easily observable and therefore achieve higher levels of agreement. When annotations involve a greater degree of subjective evaluation like whether a change in prosody is interpreted similarly across coders or whether a child is on/off task, the level of agreement between the coders decreases. Furthermore, the greater the number of label options, the greater the opportunities for disagreement. For example, the nod category, with only two possible labels, has a score indicating almost perfect agreement, while the low score for the mouth annotations could be attributed to its larger set of possible labels.

Data

Organization

The data¹ is organized by comma-separated values (CSV) files for each storytelling episode which is uniquely identified by its Round, Dyad, and Turn. We provide the raw data of annotations (ELAN program, exports, etc.) with timestamps and labels for all the behaviors listed in Table 3. Additionally, we provide scripts capable of parsing this raw data of timestamps and durations to create a single cohesive file capturing all the exhibited behaviors of a storytelling episode. These scripts take as input the varying granularity of the time series data (for example, 200 ms or 500 ms). For the purposes of this paper, we used a time interval of 500 ms.

Statistics

Listener behaviors such as gazing at partner, leaning forward, eyebrow raises, smiling, and nods have been previously identified as backchanneling behaviors related to collaborative feedback [29, 25]. Statistics on the relative rate and proportion of these particular nonverbal behaviors are provided in Table 4. Partner gazes, smiles, and forward leans are commonly observed behaviors of children listeners while short utterances and nods are rare and infrequent. This is consistent with the finding that young children's utilize nonverbal cues more frequently than verbal cues such as "ok" and "uh huh" [25].

Speaker behaviors such as prosodic cues and gaze have been shown to help in enabling more collaborative and engaging communication [28, 1, 4, 35]. Table 5 shows statistics of these gaze-based and prosodic-based nonverbal cues. Pitch, gaze, and pauses are frequently used by a majority of the population, while long/wordy utterances are less frequently observed. It is

¹To request data access, visit: <https://www.media.mit.edu/projects/p2pstory/overview/>

Behavior	Rate		%Proportion		%Episodes
	Mean	SD	Mean	SD	
Gaze Partner	4.1	3.2	8.3	7.1	94.8
Smile	2.8	2.2	20.2	23.4	81.0
Lean Toward	1.9	1.9	21.6	24.9	79.3
Brow Raise	1.4	1.8	5.0	7.7	60.3
Utterance	0.6	1.5	-	-	22.4
Nod	0.2	0.7	0.6	1.6	17.2

Table 4: **Listener Behavior Statistics:** Descriptive statistics of positive listener backchannels. Rate describes the frequency of each behavior per minute. Proportion describes the percent of time each behavior was held within an interaction. Percent episodes indicates the proportion over all interactions where the behavior was present at least once.

Behavior	Rate		%Episodes
	Mean	SD	
Pitch	4.1	1.7	100.0
Gaze Partner	4.0	3.4	93.1
Pause	3.6	2.1	94.8
Energy	1.1	1.2	65.5
Filled Pause	1.0	1.1	62.0
Long Utterance (Wordy)	0.3	0.5	29.3

Table 5: **Storyteller Cue Statistics:** Rate describes the frequency of each behavior per minute. Percent episodes indicates the proportion over all interactions where the cue was present at least once.

important to note the significant variance in all these behaviors across the 58 interactions, pointing to possible individual differences in children's development and expression of listener responses and speaker cues.

ANALYSES

To understand which listener behaviors are communicative of attention to an observer, we consider two sets of perspectives. One, the child storyteller's perception of their listening partner. Second, the perception of an adult third-person observer of the same listener. As detailed in our data collection procedure, storytellers rated their partner's level of attention to their story on a Likert scale of 1-5. Additionally, adults coders annotated the attentive state of listeners as listening or not listening throughout each of the 58 interactions. On average, children paid attention for 73% of the interaction length, with a standard deviation of 20%. Our analysis below aims to predict both the annotated and perceived attention metrics.

While it is possible that prior interpersonal relationships may impact the behaviors student's use in an interaction (as storytellers or listeners), the primary goals of this analysis are to provide an overview of the data and to offer broad strokes on differences in adult and child perception by treating the data as a generalized whole.

Results of Linear Regression

Behavior	Metric	Coefficient	Std Error	T Stat	P Value
Gaze Partner	Rate	-1.66	1.26	-1.32	0.19
	Proportion	0.79	0.92	0.86	0.39
Lean Toward	Rate	-1.00	1.03	-0.98	0.33
	Proportion	0.14	0.12	1.17	0.25
Brow Raise	Rate	3.53	1.91	1.84	0.07
	Proportion	-2.19	0.78	-2.80	0.007*
Smile	Rate	1.09	1.13	0.97	0.33
	Proportion	-0.14	0.18	-0.78	0.43
Nod	Rate	-21.62	8.94	-2.42	0.02*
	Proportion	16.05	6.37	2.52	0.01*
Utterance	Rate	1.31	1.96	0.67	0.50

Table 6: **Results of Linear Regression on Adult's Perception of Attention:** The linear regression model predicts the proportion of time the child is paying attention (or listening) in the interaction as annotated by the coder. Rate and proportion metrics of each listener response behavior were used as predictors.

Results of Ordinal Regression

Behavior	Metric	Coefficient	Std Error	T Stat	P Value
Gaze Partner	Rate	-31.32	14.76	-2.12	0.03*
	Proportion	20.44	10.60	1.92	0.05*
Lean Toward	Rate	31.27	14.77	-2.12	0.03*
	Proportion	-3.10	1.92	-1.61	0.10
Brow Raise	Rate	13.83	12.08	1.14	0.25
	Proportion	-3.58	11.05	-0.32	0.74
Smile	Rate	-17.12	12.21	-1.40	0.16
	Proportion	4.46	2.04	2.18	0.02*
Nod	Rate	365.07	215.21	1.69	0.08
	Proportion	-268.79	170.40	-1.57	0.11
Utterance	Rate	-2.93	21.31	-0.13	0.89

Table 7: **Results of Ordinal Regression on Child's Perception of Attention:** The ordinal regression model predicts the storyteller's rating of the listener's attention in an interaction as defined on a Likert scale of 1-5. Rate and proportion metrics of each listener behavior were used as predictors.

Adult Perception of Attention

A linear regression was performed to determine the effects of listener's behavior on the proportion of time the listener was rated as attentive as annotated by an adult. Results are shown in Table 6. The proportion of time spent nodding held a positive relationship ($\beta = 16.05, p = 0.01$) with attention. The rate of nodding, however, had a strong negative relationship ($\beta = -21.62, p = 0.02$). Additionally, an increase in the rate of brow raises seemed indicative of increased attention ($\beta = -2.19, p = 0.07$) while a greater proportion of time spent on brow raises held a negative relationship ($\beta = 3.53, p = 0.007$).

Child Perception of Attention

To complement this finding, we performed an ordinal regression model to evaluate the effects of listener's behavior on the

storyteller's perception of listener attention. In Table 7, we see that proportion of time spent gazing ($\beta = 20.44, p = 0.05$) or smiling at the speaker ($\beta = 4.46, p = 0.02$) had a positive relationship with attention ratings. Rate of gazing at the partner however, had a negative relationship ($\beta = -31.32, p = 0.03$).

DISCUSSION

In order to better understand the nature of child-child interactions, we began by exploring the rate and proportion of listener behaviors and speaker cues. In evaluating these statistics, we found gaze, smile, and forward leans were used by most child listeners, while nods and short utterances occurred infrequently. We also found that most children storytellers used pitch, gaze, and pauses as cues. The variance found in the presence of these behaviors across interactions suggests poten-

tial individual differences in the children population, meriting further investigation.

To build on this, we evaluated the attentiveness of a listener from two perspectives. First, a subjective metric based on adult annotations of attentive state. And secondly, a subjective rating provided by the child storyteller about the listener's level of attention. Our noteworthy finding is demonstrating that behaviors relevant to the subjective perception of attention differ between a child and an adult. Brow raises and nods are salient listening behaviors related to adult's perception of attention and conversely, gazes, leans, and smiles are salient to a child's perception. Given this finding, it is relevant to consider using nonverbal language understood by young children when designing a natural peer-to-peer relationship between a child and an interactive technology.

CONCLUSION

P2PSTORY is a dataset of peer-to-peer storytelling interactions between children of 5-6 years old demonstrating rich social-emotional behaviors as both storytellers and listeners. The objective of this work was to capture children engaged in natural social interactions, not hindered from adult supervision or structured from constrained tasks. Children participated in multiple interactions as either a storyteller or listener, allowing for the exploration of individual variations in behaviors across interactions of either roles. Additionally, these interactions took place with familiar peers as opposed to novel adults. In these three ways, the dataset is a unique contribution to the study of human-human interactions for the design and evaluation of interactive story-listening and story-telling technologies. Furthermore, we provide initial evidence highlighting the importance of accounting for the relevance of social-emotional communicative behaviors as a function of the observer, which significantly differs in perception between an adult and a child.

One of the fundamental basis of human-computer interaction research stems from Reeves' and Nass' work demonstrating how the human mind will respond to technology as social actors—capable of evoking the same social responses as they would with a human partner [32]. Through data-driven methods to better understand human social-emotional interactions, we can more appropriately design the human-like behaviors of interactive technologies for targeted social interactions such as storytelling.

FUTURE WORK

P2PSTORY provides many opportunities to analyze and model the back-and-forth dynamics of dyadic interactions of children as well as to investigate individual similarities and differences of behavior when comparing across participants and across roles. Furthermore, the demographic and developmental profiles allow for in-depth analysis on how factors such as gender, social-emotional development, and household environment can contribute to observed variations. Additionally, the multi-modal nature of the dataset enables the development of recognition models of children's facial expressions, gestures and postures, prosodic signals, and speech.

We believe that this dataset can be a powerful resource for informing both the model and design of technologies that interact with children. While the size of the dataset might be considered small compared with others, we claim that the richness and variance of features across different interactions can be used to build machine learning models, but the model's complexity will be constrained in having fewer parameters. Prior works exist in demonstrating how to build interaction models using small sample sizes [22, 20]. However, the core value of the dataset is to inform the design of child-computer and child-robot interfaces.

ACKNOWLEDGMENTS

We thank professors David DeSteno and Paul Harris for their feedback in this research. We also thank Katharine Brush, Raina Hall, Emily Holding, Yiyu Li, and Matthew Sears for annotating videos and helping conduct user studies. This work was supported by the National Science Foundation (CCF-113898), Dolores Zohrab Liebmann Fellowship, and William Asbjornsen Albert Memorial Fellowship.

REFERENCES

1. David W Addington. 1971. The effect of vocal variations on ratings of source credibility. (1971).
2. Marina Umaschi Bers, Edith Ackermann, Justine Cassell, Beth Donegan, Joseph Gonzalez-Heydrich, David Ray DeMaso, Carol Strohecker, Sarah Lualdi, Dennis Bromley, and Judith Karlin. 1998. Interactive storytelling environments: coping with cardiac illness at Boston's Children's Hospital. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press/Addison-Wesley Publishing Co., 603–610.
3. Diane Bricker, Jane Squires, Linda Mounts, L Potter, Robert Nickel, Elizabeth Twombly, and Jane Farrell. 1999. Ages and stages questionnaire. *Paul H. Brookes: Baltimore* (1999).
4. Bruce L Brown, William J Strong, and Alvin C Rencher. 1973. Perceptions of personality from speech: Effects of manipulations of acoustical parameters. *The Journal of the Acoustical Society of America* 54, 1 (1973), 29–35.
5. Hennie Brugman and Albert Russel. 2004. Annotating Multi-Media/Multi-Modal Resources with ELAN. In *Language Resources and Language Evaluation (LREC)*.
6. Judee K Burgoon, Thomas Birk, and Michael Pfau. 1990. Nonverbal behaviors, persuasion, and credibility. *Human Communication Research* 17, 1 (1990), 140–169.
7. Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 413–420.
8. Justine Cassell and Kimiko Ryokai. 2001. Making space for voice: Technologies to support children's fantasy and

- storytelling. *Personal and Ubiquitous Computing* 5, 3 (2001), 169–190.
9. Herbert H Clark, Susan E Brennan, and others. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition* 13, 1991 (1991), 127–149.
 10. Daniela Conti, Alessandro Di Nuovo, Carla Cirasa, and Santo Di Nuovo. 2017. A comparison of kindergarten storytelling by human and humanoid robot with different social behavior. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 97–98.
 11. Iwan de Kok and Dirk Heylen. 2011. The MultiLis corpus—Dealing with individual differences in nonverbal listening behavior. *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues* (2011), 362–375.
 12. Alan R Dennis and Susan T Kinney. 1998. Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality. *Information Systems Research* 9, 3 (1998), 256–274.
 13. Allen T Dittmann. 1972. Developmental factors in conversational behavior. *Journal of Communication* 22, 4 (1972), 404–423.
 14. Anna Filipi. 2009. *Toddler and parent interaction: The organisation of gaze, pointing and vocalisation*. Vol. 192. John Benjamins Publishing.
 15. Agustín Gravano and Julia Hirschberg. 2009. Backchannel-inviting cues in task-oriented dialogue. *Interspeech* (2009).
 16. K A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology* 8, 1 (2012), 23–34.
 17. Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics* 35, 7 (2003), 1113–1142.
 18. Lucille J Hess and Judith R Johnston. 1988. Acquisition of back channel listener responses to adequate messages. *Discourse Processes* 11, 3 (1988), 319–335.
 19. Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2010. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 1265–1272.
 20. Natasha Jaques, Daniel McDuff, Yoo Lim Kim, and Rosalind W Picard. 2016. Understanding and Predicting Bonding in Conversations Using Thin Slices of Facial Expressions and Body Language.. In *Intelligent Virtual Agents (IVA)*.
 21. Jacqueline M. Kory Westlund, Sooyeon Jeong, Hae W. Park, Samuel Ronfard, Aradhana Adhikari, Paul L. Harris, David DeSteno, and Cynthia L. Breazeal. 2017. Flat vs. Expressive Storytelling: Young Children's Learning and Retention of a Social Robot's Narrative. *Frontiers in Human Neuroscience* 11 (2017), 295.
 22. Jin Joo Lee, W Bradley Knox, Jolie B Wormwood, Cynthia Breazeal, and David DeSteno. 2013. Computationally Modeling Interpersonal Trust. *Frontiers in Psychology* 4 (2013), 1–14.
 23. Zofia Malisz, Marcin Włodarczak, Hendrik Buschmeier, Joanna Skubisz, Stefan Kopp, and Petra Wagner. 2016. The ALICO corpus: analysing the active listener. *Language Resources and Evaluation* 50, 2 (2016), 411–442.
 24. Carole L Menig-Peterson. 1975. The modification of communicative behavior in preschool-aged children as a function of the listener's perspective. *Child Development* (1975), 1015–1018.
 25. Lynn C Miller, Raymond E Lechner, and Deborah Rugs. 1985. Development of conversational responsiveness: Preschoolers' use of responsive listener cues and relevant comments. *Developmental Psychology* 21, 3 (1985), 473.
 26. Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A Storytelling Robot: Modeling and Evaluation of Human-Like Gaze Behavior. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 518–523.
 27. Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. 2016. EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 137–144.
 28. James P Otteson and Carol Rodning Otteson. 1980. Effect of teacher's gaze on children's story recall. *Perceptual and Motor Skills* 50, 1 (1980), 35–42.
 29. Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. 2017. Telling stories to robots: The effect of backchanneling on a child's storytelling. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 100–108.
 30. Carole Peterson, Beulah Jesso, and Allyssa McCabe. 1999. Encouraging narratives in preschoolers: An intervention study. *Journal of Child Language* 26, 1 (1999), 49–67.
 31. Janet Read, Stuart MacFarlane, and Chris Casey. 2002. Validating the Fun Toolkit: an Instrument for Measuring Children's Opinions of Technology. In *Interaction Design and Children*. 119–128.
 32. Byron Reeves and Clifford Nass. 1996. *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge University Press.

33. James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanh Kim, and others. 2013. Decoding children's social behavior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3414–3421.
34. Kimiko Ryokai, Cati Vaucelle, and Justine Cassell. 2003. Virtual peers as partners in storytelling and literacy learning. *Journal of Computer Assisted Learning* 19, 2 (2003), 195–208.
35. Bruce L Smith, Bruce L Brown, William J Strong, and Alvin C Rencher. 1975. Effects of speech rate on personality perception. *Language and Speech* 18, 2 (1975), 145–152.
36. Peter Tymms, Christine Merrell, Allen Thurston, John Andor, Keith Topping, and David Miller. 2011. Improving attainment across a whole district: school reform through peer tutoring in a randomized controlled trial. *School Effectiveness and School Improvement* 22, 3 (2011), 265–289.
37. Jacqueline Kory Westlund and Cynthia Breazeal. 2015. The interplay of robot language level with children's language learning during storytelling. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. ACM, 65–66.