Co-authoring, Corroborating, Criticizing: Collaborative Storytelling between Virtual and Real Children

(preliminary draft)

Austin Wang & Justine Cassell
MIT Media Lab
E15-315
20 Ames St
Cambridge MA
02139
ajwang@mit.edu, justine@media.mit.edu

Abstract
When children tell stories with their peers, they naturally collaborate: co-authoring, corroborating, criticizing, in essence, acting as active listeners. And, their reliance on one another, as well as the creative collaboration itself, benefits their literacy development. An interactive system that engages children in collaborative narrative might be able to have a similarly positive effect on children's development. However, due to the spontaneous nature of improvisational play among children, the problem is a challenging one from both a technical, and a behavioral standpoint. This paper describes a study of children's collaborative behaviors during storytelling and presents a model of the different roles taken by the children, as well as their associated speech acts and turn-taking cues. This model is used as the basis for an implementation of an interactive storytelling peer where keyword spotting, natural language processing with commonsense reasoning, and nonverbal cues to floor management are critical to realizing a real-time collaborative interaction between children and an embodied conversational agent.

1 Introduction

Telling stories is an important learning activity for both children and adults. As a result, a number of computational systems have been developed to encourage and facilitate storytelling. However, despite the fact that the psychological literature demonstrates that, for children, learning gains are magnified if storytelling is carried out collaboratively with peers, few computational systems aimed towards children engage them collaboratively, in part due to the sheer real-time interactivity it requires: children exchange turns spontaneously, criticize, correct, and interrupt one another, and produce unpredictable responses. The current chapter addresses the challenge of engaging children directly in collaborative storytelling interactions, in such a way as to benefit their cognitive and linguistic development.

In the following sections, we look at previous research on the roles of collaborative storytellers, their speech acts, and their turn-taking behaviors. A storyteller's role defines his/her communicative intentions; for example the 'author' role is responsible for improvising a story. Depending on their roles, storytellers use different speech acts and turn-taking behaviors to carry out their responsibilities. In order for a collaborative storytelling system to create stories with

children naturally, it has to assume or assign these storytelling roles by using the speech acts and turn-taking behaviors that are natural to that role.

1.1 The Roles of Peer Collaborative Storytellers

Collaboration during storytelling play can benefit literacy development. Sawyer (1997) proposed that conversational collaboration between peers is one of the most developmentally valuable characteristics of sociodramatic play. Looking at the spontaneous narratives of three children on their way to school, Preece (1992) found their collaborative stories to be more coherent and complex than their non-collaborative ones. Neuman (1991) observed that when children played in a literacy rich environment, they scaffolded each other and resolved conflicts by negotiating the meaning of literacy-related objects or routines. This cognitive conflict resolution has been argued by Piaget (1962) to lead to cognitive restructuring and growth; in fact, Pellegrini (1985) proposed that it is the key factor in play which affects children's literacy development.

Preece (1992) divided children's narratives into two categories, according to participants' specific roles:

- Critic and author – the audience acts as the critic by making suggestions and corrections while the author tells the story;
- Facilitator and collaborators – the facilitator coordinates narrations by assigning character roles, encouraging collaborators to talk about shared experiences or favorite stories, and by suggesting ideas for original imagined stories.

These roles exist in pairs. For instance, it would be unnatural for an author and a facilitator to collaborate. And the roles occur during different kinds of talk: children usually assume the relationship of critics and authors during narration, and the relationship of facilitator and collaborator during *meta-narration* (Sachs, et al., 1984). Both of these types of language have been thought to be essential to producing coherent narratives.

An analysis of the data from a study conducted by Ryokai, Vaucelle, and Cassell (2003) showed a third type of interaction. In the study, pairs of five-year-old girls told stories using a toy house and toy figurines as props. In addition to the behaviors described by Preece, the children also collaborated in an unregulated fashion, where the two children either competed to be the primary author, or became co-authors in the story. Co-authors differs from facilitator/collaborator pairs in that the prior occurs during narration itself, without explicit metanarrative negotiation of who shall narrate what.

- Co-authors – the children share the floor in either an organized fashion (role-play), or an unorganized fashion (simultaneous turns).

Children engage in this interaction when improvising new narratives. Participants constantly exchange turns to add to the story, and unlike the other two interactions, there is no explicit author or coordinator, that responsibility is shared between the participants. In order to produce coherent narrative structures, children use two strategies: they can coordinate explicitly by switching to a facilitator and collaborators interaction, or they can negotiate implicitly during co-author interactions by using a dialogic strategy (Bakhtin, 1981; Wolf & Hicks, 1989). Sawyer (1997) found that improvisational narratives that used dialogisms produced locally coherent plot structures, and were more likely to be well-formed.

Table 1 below illustrates the three pairings of roles, and the corresponding configurations of the participants, along with the scenarios that they are likely to engender.

Table 1 – Roles of collaborative storytellers.

Roles	**Configuration**	**Scenario**
Critics and Authors	One primary author, multiple critics	Retelling familiar anecdotes, or creating new stories
Facilitator and Collaborator	One facilitator, multiple collaborators	Organizing or initiating a story; suggesting and modeling the creation of original fantasies
Co-authors	All co-authors	Improvisational narrative

The following example shows two children engaging in these roles, and switching between them with ease:

Example 1: R and S are narrating; each of them have a figurine as a prop.

(1) R: And when she came down, she saw her mom and daddy. <author>
(2) S: No just her mom. <critic>
(3) R: But then her dad came walking down the stairs, and then he broke his leg and he fell out of the house. <author>
(4) S: Honey honey, what's happened? What's happened? <co-author>
(5) R: I fell out of the house. <co-author>
(6) S: Ooo we better get the ambulance, Cary Cary sweetie come! <co-author>
(7) R: And the little girl said, what should I do, my mom is at the ambulance with my father, and he's going to the hospital. What should I do? <co-author>
...<several sentences later>...
(8) S: She said to her mommy, that is my turn and I'll be the magic mirror. <co-author (dialogic)>
…<several sentences later>…
(9) S: Rachel, but pretend she gets eaten, but she escapes the monster's mouth. <facilitator>

This continuous improvisation by the storytellers and apparent lack of global script or routine to the story creation process demonstrates Sawyer's (2002) theory that collaborative narratives are embedded in the social context. The participants rely on shared social and collaborative knowledge that the listener might not have access to; for example, without looking at the whole transcript, it is hard to categorize the intention of line seven, and deduce the expected response (it's not clear who the question is directed towards). This may cause trouble for linguists who try to decontextualize each frame based on its communicative function, or for the designers of storytelling systems who try to recognize the role of the user on a turn-by-turn basis. Fortunately, the role played by the enunciator may be deduced by knowing the type of speech act and the turn-taking behaviors employed. That is, take the example just given: line number four was coded as co-author, even though S had been a critic up to that point. The communicative function of the speech act suggests that S has the role of either author or co-author. However, after the sentence, S did not exchange gaze with R, and fixed her gaze on her figurine. This is natural turn-taking behavior for a co-author's speech act (role-play), and unnatural for an author's

speech act. Therefore, S has switched from the role of critic, to one of co-author. Similarly, line number nine can be interpreted as either facilitator's speech act (direct) or a co-author's (dialogic role-play). However, S makes eye contact with R and signals that she is expecting an acknowledgement. This turn-taking signal indicates that the speech act was in fact a "direct" speech act, and hence resolves the speakers role to facilitator.

Speech acts and turn-taking behaviors are evidently very important to assigning and recognizing roles during collaboration. To give an example of the possible chaos if turn-taking cues were not taken into account, we can revisit line seven in example 1. The communicative function of the sentence can be categorized as role-play, which would type the speaker as a co-author, or question, which cast the speaker as an author. Depending how the listener interprets the sentence, s/he could either respond with a suggestion, or continue to role-play.

Given that it is difficult to ensure global coherence during improvisational collaborative narrative (Sawyer, 2002), these roles present a way to ensure local coherence: by defining a shared script and responsibilities, these roles act as *scaffolds* to children's storytelling play (Trawick-Smith, 2001). However, a collaborative computational system would only be able to engage children in, and itself assume, these roles with the understanding of the various speech acts and turn-taking behaviors. The following section introduces the taxonomy of ten speech acts, along with their corresponding turn-taking behaviors.

1.2 Collaborative Speech Acts

Speech acts are used by storytellers to carry out their various functions and responsibilities within their roles, and can be categorized by their communicative functions. This section illustrates the communicative functions of each speech act with examples found in the data collected by Ryokai, Vaucelle, and Cassell (2003), and explains how different speech acts may lead to different turn-taking behaviors.

1.2.1 Speech Acts between Critics and Author

These speech acts are used to give or elicit feedback during the authoring of a new story or retelling of a familiar one.

Suggestion – suggestions are made by the critic to the author, and usually occur when the author is hesitating. Suggestions are not disruptive, in that it is not always necessary to acknowledge or incorporate them. They usually refer to an event or idea that takes place in the future of the story.

Correction – critic's corrections to authors are often unsolicited, and occur when critics dispute a certain aspect of the narration. Corrections are disruptive in that failure to acknowledge or incorporate them will lead to further conflicts.

Question – can be posed by both critics and authors. The questioner is usually unsure about an aspect of the author's story, and is looking for clarification or supplemental information.

Answer – the speech act that answers the question by providing the information requested.

Acknowledge – the author can acknowledge a correction or suggestion either non-verbally, by using certain turn-taking cues, or verbally, by incorporating the feedback into the story.

1.2.2 Speech Acts between Facilitator and Collaborators

Facilitators and collaborators define a script before engaging in the narration. The following speech acts are used to negotiate the plot, characters, and various details in the narrative, and occur before and during the construction of the story.

Direct – the facilitator explicitly coordinates the story or casts play characters. The language used to propose or elaborate play ideas by speaking out of character is called meta-narrative (Sachs, et al., 1984).

Acknowledge – after the facilitator proposes a plot or designates a role, collaborators use this speech act to show acknowledgement.

Elaborate – after the facilitator has proposed the plot, and it has been acknowledged, either the facilitator or the collaborators can elaborate by supplying details to the story.

1.2.3 Speech Acts between Co-authors

Co-authors use these speech acts to narrate, through either role-play or simultaneous turns. Role-play speech acts also encompass some language used to coordinate the story; Sawyer (1997) called this *implicit metacommunication*, and defined it to be children proposing or elaborating play ideas by speaking in character.

Role-play – role-play involves multiple children co-constructing a narrative through their play characters.

Simultaneous turns – this occurs when children are competing for the turn, and may result in both children speaking concurrently. In the following example, R spoke out of turn, and S does not acknowledge R's comment.

This taxonomy is not a complete characterization of children's speech acts during storytelling, but all of the acts that result in turns being exchanged. The following two examples illustrate how different speech acts will involve different turn-taking behaviors. Both excerpts are extracted from the example 1 above, and show children in a critic and author interaction. Even though S plays the critic in both cases, the types of S's speech acts differ from example to example. As a result, S's turn-taking behavior also varies.

Example 2: S makes a correction to R.

R: So, she walked, walked, walked, walked, all the way downstairs. And when she came down she saw her mom and daddy.
S: No. Just her mom.
R: But then her dad came walking down the stairs, and then he broke his leg…

S corrected R's statement about a girl coming downstairs and seeing her mom and dad. S did not preempt her correction with any turn-taking cues, and simply started speaking at the next

sentence boundary. Nonetheless, R is able to understand S's correction and acknowledges by incorporating it into her story right away.

Example 3: S makes a suggestion to R:

> R: Pretend she went right, and she got eaten by the claws devil, but she escapes. Yeah, she did. She walked down the stairs, and she walked just as she was told. She went right into the hospital. And she said to the wizard, where is the…
> S: Mommy's here at the top floor.

For a suggestion speech act, children tend to make suggestions only when the other child solicits it, usually through signs of hesitation. In this example, R began to drawl mid-sentence. Such paralanguage and syntactic cues signal S to offer suggestions.

These two examples demonstrate how different speech acts, even within the same role, can have different turn-taking behaviors, and reinforce the idea of using these behaviors to distinguish between speech acts. By analyzing the data collected by Ryokai, Vaucelle, and Cassell (2003), ten types of collaborative speech acts that resulted in turns being exchanged were identified. They are presented in Table 2, and are categorized according to the roles for which they are used; the turn-taking behaviors for each speech act are also listed.

Table 2 – Taxonomy of children's collaborative speech acts.

Roles	**Speech act**	**Speaker**	**Function**	**Turn-taking behaviors**
Critics and authors	Suggest	Critic	To suggest an event or idea to the story	Eye gaze towards author, author may use paralanguage drawls and socio-centric sequences like "uhh"
	Correct	Critic	To correct what's been said	Eye gaze towards author
	Question	Both	To seek clarification or missing information	Eye gaze towards other, lack of backchannel feedback like head nods, increased body motion, author stops gesturing
	Answer	Both	To clarify or supply missing information	Eye gaze towards other, rising pitch, question syntax, author stops gesturing
	Acknow-ledge	Author	To acknowledge a suggestion or correction	Eye gaze towards critic, backchannel feedback like "mm-hmm", author stops gesturing
Facilitator and collaborator	Direct	Facilitator	To suggest storylines and designate roles	Eye gaze towards collaborator, socio-centric sequences like "OK", both stop gesturing
	Acknow-ledge	Collaborator	To acknowledge a role designation or storyline suggestion	Eye gaze towards facilitator, backchannel feedback like head nods, both stop gesturing

	Elaborate	Both	To narrate following suggested script	Eye gaze towards other, may start gesturing
Co-authors	Role-play	Both	Play the role of characters in the story	Eye gaze towards action, prosody of in-character voice, gesture with prop
	Simultan-eous turns	Both	Compete for turn	

2 Previous Work

In the next sections we look at previous work in conversational systems and storytelling systems.

2.1 Conversational systems

Conversational systems have received a lot of attention, and a subgroup of these systems has incorporated natural turn taking behaviors in order to create a more natural human computer interaction. Donaldson and Cohen (1997) outlined a system that uses constraint satisfaction to facilitate floor management in an advice-giving agent, where the beliefs and desires of the agent motivates it to take turn, and constraints such as the user's pause length, intonation, and volume, restrict it from doing so. Allen (2001) describes an architecture for building conversational systems with human-like behaviors such as turn taking, grounding, and interruptions. Allen points out that such systems must be able to incrementally understanding the ongoing dialogue as well as incrementally generating responses. Floor management behaviors are generated depending on the goals of the agent, the agent's understanding of the dialogue, the state of the world, and the state of the floor. However, the system only described turn taking on the functional level, and did not suggest any actual instances of floor management cues.

Apart from relying analyzing verbal behaviors, researchers have also explored other modalities as means to facilitate turn taking. Darrell et al. (2002) presents an agent that uses eye gaze as an interface to turn detection. If the user is determined to be looking at the agent, it is assumed that the speech is directed towards the agent. They conducted a study where subjects were given a choice between using their eye gaze, flicking a switch, or saying "computer", to signal that they are talking to the agent. The subjects thought the eye gaze method was the most natural.

Cassell et al. developed an embodied conversational agent that was capable of negotiating turns with a human conversant. Rea (Cassell et al., 1999) parsed the user's speech for turn-taking signals, and responded depending on the signal, and who had the speaking turn at the time.

2.2 Literacy Systems and Storytelling

Some intelligent systems use stories as a medium to support children's language development. The goal may be to target certain aspects of a child's language and provide contextual feedback (Mostow, 1994; Wiemer-Hastings, 1999), or to encourage idea development (Glos & Cassell, 1997). A subset of learning systems act as the stage or audience for children's storytelling (Nijholt, 2003; Marsella, 2000; Vaucelle, 2001; Ananny 2002; Ryokai & Cassell, 1999; Ryokai & Cassell, 1999).

The **LISTEN** project (Mostow, 1994) listens to children read stories and uses speech recognition to follow their speech. The information is used to generate constructive feedback to the children's oral reading skills. Mostow found that children who used project LISTEN read more advanced stories with fewer mistakes and less frustration. In Wiemer-Hastings' (1999) project Select-a-Kibitzer, children type in their written stories, and the system analyses the text using natural language techniques such as latent semantics analysis, to determine the coherence, purpose, topic, and overall quality of the text. The system then provides feedback through multiple animated characters, each representing one of those variables of measurement.

Glos and Cassell (1997) created **Rosebud**, a system that consists of a desktop interface that recognizes children's stuffed animals through infrared sensors, and invites the children to write stories about their toys. Children type their stories into the computer, which then analyses certain features of the story, and provides relevant feedback and encouragement. If the story is short, the system will prompt for longer stories; if there is not enough temporal information in the story, the system will prompt the child for more.

Virtual Storyteller (Nijholt, 2003) is a story creation platform where computer agents act as the characters, directors, and narrators of the story. Although the current version is not interactive, future versions may allow users to direct the plot, and thereby experiment with narrative structures and consistency.

Marsella (2000) implemented an agent-based pedagogical drama where children watch director and actor agents cooperatively create drama with story structure and other dramatic qualities, they can interact with the system by altering the intentions of some of the actor agents. The actors perform various pieces of dialogue that were deconstructed from whole stories, in an attempt to satisfy director and cinematographer agents who had certain artistic and dramatic requirements.

Story Listening Systems, or SLS (Bers & Cassell, 1998) offer an alternative approach: the system plays the role of an attentive listener to children's stories. In the process of listening and giving feedback, these systems may highlight important aspects of the process, such as plot consistency and structure, decontextualization for an audience, temporal arrangement, and cohesion. An example is **StoryMat** (Ryokai & Cassell, 1999; Cassell & Ryokai, 2001), a system designed to support young children's fantasy storytelling. The implementation consisted of a soft cloth quilt with appliquéd figures. When children told stories with one of the small stuffed animals provided with the quilt, their grip triggered recording of the child's narrating voice and the coordinates of the stuffed animal. When new input was later encountered at the same place on the mat, a movie file of the previous input was automatically triggered and played back via a projector above the mat, and speakers next to it. The current child could then tell her next story. Sometimes she might come up with a continuation of the story she just heard. Or she might continue telling her own story, incorporating some story elements from the story she just heard. In this sense, StoryMat is a kind of imaginary playmate, but who also mediates collaborative storytelling between a child and her peer group. Our evaluation of StoryMat concentrated on three kinds of emergent literacy activities. Results demonstrated that StoryMat encouraged more symbolic transformations in the children's stories for both children who played on the mat alone and with a co-present peer ($F(3,20)=9.7$, $p<.01$). Children playing on StoryMat also incorporated story elements from the stories offered by StoryMat in a way similar to how they did from real life peers ($F(3,20)=3.49$, $p < .05$). Finally, children playing alone on StoryMat more often took the more narratively advanced role of narrator (72%) than of character (28%),

while the control group playing alone acted in character role (95%) much more frequently than in narrator role (5%).

Animal Blocks (Ryokai & Cassell, 1999) was created as an attempt to scaffold children's literacy acquisition by helping them make connections between oral and written stories. A book acts as the stage for the storytelling play, while several animal toys act as props. During storytelling, the child is free to place objects at specific locations and record audio associated with that figurine. A virtual representation of that toy is then projected onto a physical book. Children are encouraged to enter words that supplement their oral story. They can also peruse past stories by other children by flipping the pages in the book.

DollTalk (Vaucelle, 2001) was created to help young children take different perspectives during storytelling play. The child tells his/her story to an animated computer character, using two stuffed animals as props, and their story would be recorded by the system. The stuffed animals contain accelerometers that monitor the movement of those toys; the system assumes that if a toy is being shaken, then the child is narrating a story segment associated with that toy. When the child is done, the recorded audio is played back with two different pitches to signify the stuffed animal that was speaking at the time.

TellTale (Ananny, 2002) illustrates by its form an important concept of writing: units of discourse must hang together somehow, and then be connected to other units, and there must be a beginning and an end. TellTale is a caterpillar-like toy with five modular body pieces and a head. Children can press a button on each of the five body pieces to record 20 seconds of audio. The child can then press a button on that body piece to play back their own voices. The body pieces detach from one another and children can arrange and rearrange them in any order. At any point the child can attach the toy's head to the body to hear the entire audio story played in sequence. Any body piece can be re-recorded, or re-arranged. In our evaluation of TellTale we constructed a control condition where only one piece recorded audio, and it contained 100 seconds of audio (the same amount as the entire original TellTale). 14 children playing alone with the unified unit (UTT) or segmented unit (STT) TellTales were invited to record stories (in a room with no adult present). Stories told with STT were longer (mean of 72 words; 41 seconds) than those told with UTT (42.1 words; 34.2 seconds). Stories told with STT had fewer false starts than those told with UTT indicating that the segmented body pieces may allow children to plan their utterances off-line. Stories told with STT also had contained more conjunctive phrases (and, then, however, when, while, after, later, so, therefore, one day) per word (.1 conjunctions/word) than those told with UTT (.06 conjunctions/word). And when conjunctive phrases did occur in STT, they tended to occur at body piece boundaries, indicating that children treated body pieces as story units, linking them with connectives. In both UTT and STT conditions children tended to tell stories with classic beginnings (e.g., "once upon a time") but only in the STT condition did children also consistently finish their stories with classic endings (e.g., "the end"). Stories told with UTT tended to end in either false starts or long pauses indicating that children may have been having difficulty planning the next utterance. TellTale's segmented interface, then, seems to help children tell stories that are longer, more cohesive (containing fewer disfluencies and more conjunctions) and with more traditional beginnings and ends. The skills children practiced while playing with the segmented version of TellTale (planning, chunking, revising) are very similar to those that are required for written literacy.

Sam (Cassell et al, 2000; Ryokai, Vaucelle & Cassell, 2002) is a virtual child who invites children to participate in collaborative storytelling play with real toys. The Sam system has two

components: an embodied conversational agent – a life-sized child named Sam – and a toy castle with several plastic figurines. Sam is projected on a screen behind the castle, and can both tell stories, using a recorded child's voice, and listen to the real child's stories, responding with appropriate feedback and short comments (See Figure 1.)

The house is a two-story playhouse, with a virtual counterpart that is displayed in front of Sam, creating an illusion that the physical house extends into Sam's space. In addition, there are wooden figurines, which are tagged with RFID badges. A small compartment in the attic, which we shall designate "the portal", is accessed via a small swinging door. The locations of the figurines within the house can be tracked by Swatch RFID tag readers embedded in the rooms and the portal. The portal door is latched with electric contacts, such that Sam can sense whether the door is open or shut. Sam has a database of short stories that involve one or two characters, played out by the figurines. All of the stories demonstrate third person narrative voice, reported speech ("character voice"), appropriate introduction of new characters, and appropriate use of cohesive devices.

Figure 1 – Sam greeting

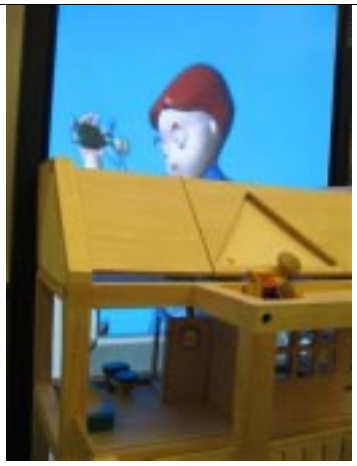

Figure 2 – Sam gesturing with figurine

An empirical study examined the effect of Sam on collaborative peer-like and metalinguistic behavior, and use of decontextualized language. This study compared (a) single children to pairs of children, and (b) interacting with Sam vs. interacting with the castle but without Sam, and examined the use of (i) decontextualized language, (ii) metalinguistic reflection, and (iii) prosocial constructive collaboration. 31 children aged five years were videotaped interacting with the castle, in pairs and alone, with and without the Sam virtual peer. The children's stories were analyzed for the use of emergent literacy, operationalized here as the occurrence of explicit spatial expressions (e.g. "then the boy went to the *kitchen*"), temporal expressions (e.g. "he went downstairs *when he heard the noise*"), and quoted speech (e.g. "then she said, "Oh no!"" or "he said that he wasn't hungry"). Results demonstrated that the presence of Sam significantly increased the frequency with which children used quoted speech ($F(3, 24)=10.58$, $p<.01$), temporal expressions ($F(3, 24) = 30.52$, $p<.01$), and spatial expressions ($F(3, 24) = 68.04$, $p<.01$). No effect was found for number of children (the one child vs. the dyad condition). This suggests that Sam is equally successful in evoking literate behaviors when working with a single child as when working with two children. Children's use of these devices increased over the

course of a single interaction with Sam. Results demonstrated that with each subsequent story, children in the "one child with Sam" condition used more decontextualized language and metalinguistic expressions: for spatial expression, r=.35, p<.05, for quoted speech (r=.27, p<.06), and for temporal expressions (r=*.363, p<.041*). No improvement in use of these devices was found for 'two children without Sam'.

In terms of prosocial collaborative peer-like behavior, children were willing – and eager – to interact with Sam as they might with another child. Children also incorporated elements from Sam's stories, continued Sam's stories, and asked Sam to continue their stories. In fact, in terms of metalinguistic behavior, children even coached Sam in storytelling. Most striking, however, was the children's non-verbal turn-taking behavior, as revealed by their eye gaze. Sam forces turn taking behaviors with children, and because taking storytelling turns is the only way children can interact with Sam, the children in "one child with Sam" group demonstrated turn taking behaviors with Sam beautifully. A Two-Independent-Samples Test revealed significance of p <.01 for each of the following steps: (1) children kept their gaze on Sam when Sam was telling a story and (2) when Sam gave back the toy. When the child began to tell her story, (3) she shifted her gaze to the castle, (4) gazing back at Sam when she was finished and was giving back the turn to Sam. In the child-child condition, on the other hand, children rarely gazed at one another, and no significant turn-taking behaviors were observed.

These *Story Listening Systems*, then, are capable of increasing the use of pro-literacy devices in children by acting as attentive listeners. Their passive approach, however, means they never gain much control over what and how the child is actually learning. The opposite is true with tutoring systems whose didactic interaction model is effective in influencing children's learning, but leaves little room for them to improvise or produce personally meaningful content

There exists a balance between the two paradigms: a literacy system that provides an open-ended stage for storytelling, and yet has direct control over their literacy behaviors. The literature suggests that collaboration is a natural phenomenon during improvisational narrative (Preece, 1992; Garvey, 1990; Wood, 1995; Damon, 1983), and that children can produce more coherent stories when they use certain collaborative strategies (Sawyer, 1997; Newman, 1991). Our work on Sam suggests that children are willing and capable of considering Sam as a peer, and engaging in nonverbal turn-taking eye gaze with Sam. For this reason, we decided to implement turn-by-turn collaborative, critical and co-authoring behaviors in Sam.

3 Implementation

According to the cues to floor management model described in Table 2, a subset of speech acts and their corresponding turn-taking behaviors were chosen and implemented into the existing Story Listening System, *Sam*. Sam was extensively modified: new non-linear stories were added (each with multiple paths and endings), along with collaborative turn-taking strategies, and considerable effort was taken to constrain the themes of the interaction such that it would be manageable by an autonomous agent.

3.1 Sam's Collaborative Storytelling

Sam was modified so that she could assume three of the six collaborative roles: author, facilitator, co-author; when she assumes these roles, she attempts to encourage the child to take on the three corresponding roles (critic, collaborator, co-author) by partnering a speech act within that role with appropriate turn-taking behaviors. Since speech acts vary in their turn-

taking cues and therefore exert different requirements on the output interface, three speech acts, one from each role, were selected to maximize the variety of collaborative interactions, but at the same time, minimize the strain on Sam's interface. These are listed as follows:

- Sam, Author; Child, Critic – Sam *anticipates* the *correction* speech act;
- Sam, Facilitator; Child, Collaborator – Sam *performs* the *direct* speech act;
- Sam, Child, Co-authors – Sam attempts to *engage* the child in role-play.

Careful coordination of both turn-taking behaviors and speech acts are essential when participating in collaborative storytelling. The first part of this section describes the speech acts and cues used when Sam gives the turn to the child, and the second part does so for situations where Sam is taking the turn from the child. The last segment presents Sam's multi-modal interface, and how it is able to convey these cues.

3.1.1 Yielding Turns

During a critic and author interaction, Sam acts as the author, and any interruption from the child is interpreted as a *correct* speech act. When Sam is telling stories collaboratively, and it's her turn, she gives the turn to the child if she detects an audio level higher than a certain threshold. The turn-yielding signal involves stopping her hand gestures, shifting eye-gaze from the figurine to the child, and leaning forward slightly towards the child for two seconds.

Sam can also engage the child in a facilitator and collaborator interaction by using a *direct* speech act. She does so with meta-narrative language, and can designate the turn explicitly using either a question or a socio-centric sequence. For example:

> Sam: Let's pretend Jane runs into the kitchen first and tries to hide there. But she couldn't find a good place so she runs into the Brad's bedroom. Ok?

Throughout the *direct* speech act, she maintains eye contact, and does not gesture with her hands.

When Sam takes the role of co-author, she attempts to engage the child via the *role-play* speech act by giving them opportunity to join in as another one of the characters. Here's an excerpt from one of Sam's stories:

> Sam: One day, Jason came to the hospital to see Sara, he has never been to the hospital before, so he's feeling scared. Sara asks him: "oh Jason, what happened to you?" And he said…

In this example, Sam assigns the child to the character Jason by beginning a phrase by Jason. During the turn exchange, Sam does not raise her head to look at the child, or continue her current hand gesture. During all three cases, Sam provides back-channel feedback during the child's turn: Sam nods his head, or says "uh huh".

3.1.2 Taking Turns

When the child is finished with the *correct* speech act, the cues to relinquish the turn include syntax and the shifting of eye-gaze towards to the other person (Goodwin, 1981). Since Sam does not recognize either of these cues, Sam simply goes back to authoring when a two second silence is detected. She acknowledges the correction by displaying back-channel feedback (Duncan, 1972), and by narrating a story segment that incorporates the correction.

For the other two speech acts, children refrain from interrupting each other mid-turn, and only interject when they have received proper turn-yielding signals. However, as they become more impatient, their behaviors become more aggressive. For example, in the following example in which the child S was narrating to Sam and another listener, the listening child became increasingly uneasy, and began to shift her body posture frequently, while gesturing with her hands, until finally the adult present recognized her desire to tell a story and regulated the turns.

Example 4:

S: They got her in the ambulance said, nope, nope, nope. We're not going to get her again. Then, the little wizard came and said, oh. They're not going to get her? So, he disguised her. And she was like ohhhh. Then the ambulance came. Oh. Another sick person. They put her up, and then the disguise came off. She was fixed again. And, from now on, she knows not to jump down the castle, instead, she always takes the stairs. The end. Your turn, Sam.

A: First we're going to let Rachel go. And, then, OK. Sam wants to go.

Sam models impatience in much the same way. During the child's turn, Sam gets increasingly impatient, and will attempt to take the turn using increasingly demanding turn-taking behaviors. After a long period of the child speaking, Sam will lean forward and plea: "can I go now?" If the child does not relinquish their turn, Sam continues to listen, until after another minute or so, Sam will interrupt by leaning forward, gesturing, and saying: "OK, my turn!" and will attempt to continue the story. Duncan (1974) found that the listener's claiming the speaking turn was preceded by the display of a back-channel signal, either vocal or visual.

3.1.3 Multi-modal Interface

To perform the various turn-taking cues described above, Sam uses eye-gaze, body and head posture, hand gestures, and speech to negotiate turns. In addition to exchanging turns, the interface is also responsible for acting out stories and for giving backchannel feedback during the child's stories. The life-sized 3D humanoid model is animated by the Pantomime toolkit (Chang, 1998), which enables numerous degrees of freedom over motor control.

All of Sam's graphical and audio output is predefined. Each output command consists of a script defining the timings of speech and gesture actions. A female adult sound actor records all stories and utterances; the audio is then raised in pitch and slowed down so that it resembles that of a 6-year-old child. The gestures are based on observations of narrations by real children in the same context, and are meant to add to the realism of the experience, and reinforce events within the stories.

The resultant physical behavior is an emulation of an agreeable and attentive 6-year-old child. For example, during the user's turn to tell a story, Sam tracks the location of the figurines with her eyes, nods her head, and voices backchannel feedback.

3.2 Responding Naturally

Responding naturally constitutes different things for different types of speech acts. To acknowledge a correction, Sam should incorporate the correction into the story. When participating in role-play, Sam should continue the story that makes sense given the events narrated by the child. There are seemingly unlimited variations to how the system should

respond, and since Sam's speech is prerecorded for realism's sake, responding naturally is an extremely challenging problem.

The themes of the stories help to restrict the context. Furthermore, when Sam engages the child in role-play, or directs the story, the story content is designed to increase the chances of the child's responses falling under certain categories. For example, in one of Sam's stories, Sam describes how a boy and a girl were playing hide and seek, and as Sam is narrating about the girl trying to find a hiding place, Sam gives the turn to the child. Given the priming of the story, the child is more likely to describe how the girl finds her hiding place. Sam's possible responses to the child include a response for each general location within the house, such as the kitchen, bedroom, or bathroom. A generic response is also available in case the child decides to deviate and none of Sam's other responses is appropriate.

Sam responds with the story continuation that is most cohesive and locally coherent (Halliday & Hasan, 1976) to the child's input. Although the importance of coherence is constantly emphasized over that of cohesion, Sawyer (1997) observed that children's improvisational narrative were rarely globally coherent. On the other hand, he also observed that improvisation resulted in "pockets of coherence", and that the stories maintained consistent characters and themes throughout, suggesting that it may be more natural for the system to respond with cohesive responses that were locally coherent, as opposed to globally coherent responses. In the current system, coherence is ensured by comparing the semantic/lexical distances between the words used in the child's story and the pre-defined keywords that categorize the various segments in Sam's repertoire, using a commonsense knowledgebase.

3.2.1 Semantic/lexical Distancing with Commonsense Reasoning

Semantic distancing is one way of calculating the local coherence of two segments, whereas lexical distancing is a good but incomplete way of measuring relative cohesion between two narrative segments. The Open Mind Common Sense Knowledge Base (Singh, 2002) contains both semantic and lexical relations between words, which makes it a good candidate for the knowledge base in such an application. This section first describes the Open Mind database, how it was adapted for this application, and then explains how semantic and lexical distances between words can be calculated using the database.

Two words are lexically linked by either having similar identities of reference, or being semantically close or related (Halliday & Hasan 1976; Hoey, 1991). For example, 'job' and 'employment' are lexically linked because they are synonyms for occupation; 'prince' and 'princess' on the other hand are both members of the same group (the royal family), and are therefore lexical linked. Other formal lexical relations include: hypernymy (isA), hyponymy (isKindOf), common subsumer (equivalentOf), meronymy (partOf), holonymy (hasA), and antonymy (complementOf). Lexical distance is the number of lexical links between two words. Since there can be multiple lexical chains connecting two words, there are many ways of calculating the lexical distance: using lexical chains found in the discourse history, or only using context-relevant lexical relations. The definition depends on the application, and for this particular implementation of Sam, the lexical relations used are hypernymy and meronymy, and all lexical links are counted. These relations were chosen because they were the ones available from the Open Mind database.

The semantic distance between two words is a similar idea to lexical distance, except the types of relations are different. There are no formal definitions for semantic relations, but several commonly used ones are found in the Open Mind database: hasLocation, hasProperty, hasAbility, hasStep, hasWant, and so on. For this implementation, a subset of these was selected in order to speed up the calculation, and was chosen based on its probable relevance in children's stories. These include: hasLocation, hasStep, hasEffect, and hasWant.

3.2.2 Open Mind Common Sense Knowledge Base

The Open Mind project is an attempt to gather commonsense knowledge from the public, and is composed of over a million pieces of commonsense, compiled from the English sentences entered by the public via the Open Mind website. The commonsense knowledge is represented in a network of concept nodes, such as "brother", or "swimming". Connections between nodes in the network represent semantic or lexical connectedness. For example, the node "father" is lexically connected to the node "man" via the relation "isA"; while the node "back yard" is semantically connected to "grow plants" via the relation "hasUse".

To optimize the Open Mind database as a lexical/semantic web of concepts pertinent to children's stories, a context-specific network was extracted from the original database by only retaining concept nodes within five predicate distances from keywords (nouns, verbs, adjectives) mentioned in children's stories collected in the study by Ryokai et al. (2003). Table 3 shows a sample of the extracted keywords and the resultant database:

Table 3 – Building the context-specific commonsense database.

Adding keyword: child	NODE: ghost
Adding keyword: plant	EDGE:
Adding keyword: flowers	PRED: hasEffect
Adding keyword: happens	TARGET: fear
Adding keyword: planting	SENTENCE: the effect of seeing a ghost is feeling fear
Adding keyword: right	DIRECTION: fw
Adding keyword: realizes	WEIGHT: 0.5
Adding keyword: school	**********
Adding keyword: bye	NODE: gift
Adding keyword: leaves	EDGE:
Adding keyword: teacher	PRED: hasLocation3
	TARGET: box
	SENTENCE: something you find in a box is a gift
	DIRECTION: fw
	WEIGHT: 0.5
	EDGE:
	PRED: hasLocation3
	TARGET: party
	SENTENCE: something you find at a party is a gift

	DIRECTION: fw WEIGHT: 0.5 EDGE: PRED: hasLocation5 TARGET: store SENTENCE: you are likely to find a gift in the store DIRECTION: fw WEIGHT: 0.5 EDGE: PRED: hasLocation5 TARGET: birthday party SENTENCE: you are likely to find a gift in a birthday party DIRECTION: fw WEIGHT: 0.5

Semantic distance is a good assessment of relevance (Brooks, 1998), and has been applied widely in applications such as information retrieval, document summarization, and even hypertext construction (Green, 1997, 1999). Recent research has shown that relevance plays a large role in the coherence of text (Lehman & Schraw, 2002); it could therefore be an effective heuristic to the coherence of two separate story segments. Semantic relations in Open Mind include: "hasRequirement", "hasConsequence", "hasLocation", and so on.

Cohesion between story segments can also be estimated by lexical distance. Halliday and Hasan (1976) divided cohesive relations into four main groups:

1. Reference, including antecedent-anaphor relations, the definite article *the*, and demonstrative pronouns;
2. Substitution, including such various pronoun-like forms as *one*, *do*, *so*, etc., and several kinds of ellipsis;
3. Conjunction, involving words like *and*, *but*, *yet*, etc.;
4. Lexical cohesion, which has to do with repeated occurrences of the same of related lexical items.

The final relation is well-represented in the Open Mind database, with lexical relations such as "isA", "hasPart", "hasColocate", and so on. While the other three relationships are syntactical; therefore, they cannot be addressed by the lexical approach.

The metric for scoring the story segments combines semantic and lexical distance in the following way:

$$\text{Score of story segment } x = \sum_{0}^{cs} k \times \left(1 + \frac{f}{d(x)}\right) \times \left(1 + \frac{n(x)}{f}\right)$$

Equation 1 - Metric for ranking Sam's story continuations.

Where c is the number of keywords from the child's input; s is the number of concept descriptors for the current story segment; k is the segment's current score (initially equal to 0.5); d is the average number of semantic/lexical relations separating the two words; n is the number of different semantic/lexical paths connecting the two words; and f is a fudging factor (set to 5).

The metric is designed to rank the story segments aggregately over the child's entire turn, in order to support the idea of local coherence. The metric favors a story segment that already has a high score, which means that keywords have less and less effect as the leading segments emerge. By only calculating the semantic/lexical distance for the child's last input, as opposed to say, the entire discourse history, the most coherent and cohesive segment within the local context is selected. When the child finishes speaking and gives the turn back to Sam, the story segment with the best score is performed. If the scores are all below a certain threshold (equal to 2), or if there is no clear winner (within 1 of each other), the generic story segment is narrated. Figure 2 shows the child's input, and table 4 shows the output of the NLP. The story is about hide-and-seek, and Sam has four possible responses, one in the context of the kitchen, one set in the bathroom, one in the bedroom, and one outside (the generic response).

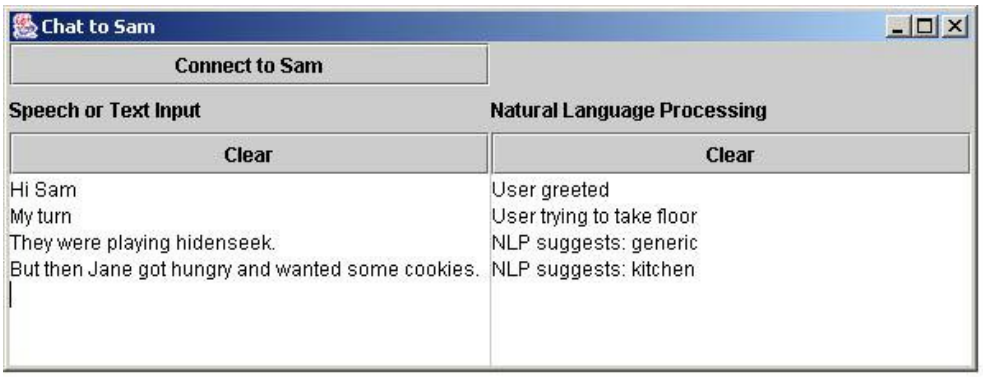

Figure 3 – Screen shot of natural language processor interface.

Table 4 – Trace of natural language processor.

Sentence: they were playing hidenseek.
tagged: they/PRP were/VBD playing/VBG hidenseek/NN ./.
keywords: [hidenseek, were, playing]
FINDING PATHS BETWEEN playing AND kitchen
Scoring segment: 0, of topic: kitchen, with keyword: playing. Score=1.5
FINDING PATHS BETWEEN playing AND bedroom
Scoring segment: 1, of topic: bedroom, with keyword: playing. Score=1.5
FINDING PATHS BETWEEN playing AND bathroom
Scoring segment: 2, of topic: bathroom, with keyword: playing. Score=2.0
FINDING PATHS BETWEEN playing AND outside
Scoring segment: 3, of topic: outside, with keyword: playing. Score=0.5
Current best segment is: 3 outside
Sentence: but then jane got hungry and wanted some cookies
tagged: but/CC then/RB jane/PRP got/VBD hungry/JJ and/CC wanted/VBD some/DT cookies/NNS ./.

```
keywords: [cookies, hungry, got, wanted]
FINDING PATHS BETWEEN hungry AND kitchen
Scoring segment: 0, of topic: kitchen, with keyword: hungry. Score=3.0
FINDING PATHS BETWEEN hungry AND bedroom
Scoring segment: 1, of topic: bedroom, with keyword: hungry. Score=1.5
FINDING PATHS BETWEEN hungry AND bathroom
Scoring segment: 2, of topic: bathroom, with keyword: hungry. Score=2.0
FINDING PATHS BETWEEN hungry AND outside
Scoring segment: 3, of topic: outside, with keyword: hungry. Score=0.5
Current best segment is: 0 kitchen
```

The trace shows the POS tagging the sentence, extracting the keywords, and rescoring the segments based on semantic/lexical distance calculations. Only changes in the segments scores are shown. You can see that even though segment 2 scored the highest after the first sentence, the generic segment was still recommended; however, after the second sentence, the scores were spread out enough that the kitchen segment was recommended.

4 Playing with Sam: An Example

In the previous version of Sam, children could engage in two types of interactions:

- Storytelling – Sam narrates a complete story from beginning to end while the child acts as a passive listener;
- Storylistening – Sam listens to the child's story and provides back-channel feedback through speech, eye-gaze, and head nods;

She is now capable of a third:

- Collaborative storytelling – Sam and the child take turns to contribute to the same storyline and collaboratively construct a coherent story.

The child has full control over the type of interaction by placing different numbers of figurines in the portal. If none of the figurines is detected in the portal, Sam switches to storylistening mode; if both figurines are in the portal, Sam switches to storytelling mode. If the child decides to hold on to one figurine, and place the other in the portal, then Sam engages the child in collaborative storytelling.

In order to allow Sam and the child to collaboratively tell stories, without demanding of the system perfect speech recognition, each story is designed to engage the child while strictly defining the context. The themes used include: a visit to a toy factory, and a dinosaur museum. Sam begins every story by introducing the characters, and setting the theme of the story. There are designated points in the stories where Sam attempts to pass the turn to the child. Depending on the child's responsiveness, and how the speech recognition module interprets their response, Sam will switch between different roles.

Here is an example of an actual interaction between a child and the implemented collaborative Sam[1]:

Sam: Let's tell a story together. Let's pretend, once there was a little boy called Jack, and his best friend Mary. Jack and Mary were playing at home one day and there was nobody else around. They parents were out working, and they had the entire house to themselves. They got bored watching TV and they wanted to play a game. So Jack asks Mary: 'let's play a game. What do you want to play? How about we play hide and seek?" Mary's excited because the house is really big and there are lots of places to hide. She can hide upstairs in the bedroom or bathroom, or downstairs in the kitchen or bedroom. She says: "sure, let me go hide and you can start counting." So Jack faces the wall and starts to count, "one, two, three". Mary shouts: "no peeking!" and runs off. She, mmm, then she…

C: four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, ready or not, here I come!

Sam: Let's pretend Mary runs into the kitchen first and tries to hide there. She couldn't find a good place so she runs into the Jack's bedroom. Ok?

C: Ok. She runs into Jacks bedroom and hides behind his bed. But then she notices that it's not tall in there, because she's tall, for Jack to see her. So she runs into the bathroom and she tries to hide in the shower.

Sam: Jack looks everywhere for Mary, in the bedrooms, in the kitchen, but he couldn't find Mary. He goes into the bathroom, and sees Mary hiding there. He creeps up to her and when he's right behind her shouts: "ahhh!". Haha, Mary was got sooo scared she almost fainted, and they laughed and laughed together. The end.

In this example, Sam starts off by setting the scene of the story through the role of an author, but hesitates about one of the character's actions. The child responds with a continuation, however, Sam was unable to find a good match, so Sam takes the role of a facilitator and suggests a script. The child continues by following the script, and Sam is able to extend the story coherently as a co-author.

5 Limitations and Future Work

It remains to test the effect of collaborative Sam on children's emergent literacy behaviors. Even before that evaluation, however, it is clear that a certain number of limitations exist in the current implementation.

5.1 Theoretical Limitations

One piece of the interaction model is missing before Sam can naturally complete the collaborative exchanges with a child. Currently, Sam is able to assume and assign collaborative roles: it implicitly understands that by assuming the role of the facilitator, the child would be encouraged to become the collaborator. However, given that the child has assumed the designated role, the model for detecting and predicting the subsequent speech acts is still weak.

[1] Due to the poor performance of the current speech recognition system, Sam's responses were launched on a control panel controlled by a researcher through a Wizard of Oz setup.

For example, is a *question* speech act always followed by an *answer* speech act? If Sam proposes a plot as the facilitator, should she expect the child to *acknowledge*, or *elaborate*?

Cohesion and local coherence are used to mediate all three of the speech acts that Sam responses to, however, this approach may not be extensible to other speech acts. For example, when responding to a *question* speech act, the most natural response is to answer the question. To be able to do so convincingly requires a different set of natural language abilities, and the same is true for other speech acts such as *suggest*, or *simultaneous turns*. Further investigation into the language processing requirements of the other speech acts will be required before an autonomous system can collaborate using them.

Finally, with limited stories and speech acts, the interaction can become repetitive. Children have been observed to chat with Sam, and even ask her questions. Since Sam does not recognize questions or other conversational speech acts, she can only respond with being silent, or by telling a story. Children may find Sam less convincing as a story partner as time goes on. It is worthwhile to investigate how children maintain relationships with storytelling partners over long periods.

5.2 Technical Limitations

Sam is currently able to perform all the input, output, and processing functions described above, except for speech recognition. We are currently undertaking an approach to children's speech recognition based on a stochastic segment-based recognizer called SUMMIT (Phillips & Goddeau, 1994), which was trained specifically for the JUPITER weather domain (Glass, Hazen, & Hetherington, 1999). Its language model is being retrained on transcripts from Ryokai, Vaucelle, and Cassell's study (2003), and will then be tested against approximately 120 minutes of acoustic data of children's stories from a subsequent study. A noise model will be incorporated into the grammar to deal with ambient noise, and unintelligible phrases.

When Sam has the turn, the speech recognizer will have a restricted grammar of 60 phrases, containing speech acts such as greetings and farewells as well as verbal turn-taking behaviors. A restricted grammar has a much lower error rate than full dictation, but offers sparser coverage. This is a worthy tradeoff given that during her turn, Sam is only concerned with turn-taking attempts from the child, and the only speech act not explicitly solicited by Sam during her turn (*correct*), occurs without any turn-taking cues.

However, during the child's turn, Sam will need to extract as much semantic information from the child's input as possible. Therefore, during the child's turn, the speech recognizer will operate in dictation mode, with a grammar of several thousand phrases.

In the meantime, while we continue development of speech recognition for this task, Sam interact with children using a Wizard of Oz (WoZ) set-up. During WoZ operation, an operator controls Sam remotely as if she were a puppet, by listening to the child's voice, and choosing one of Sam's verbal and non-verbal responses. Nevertheless, an adult user can currently interact with Sam via a commercial speech recognition package (IBM ViaVoice). The commercial speech recognizer takes the place of the research speech recognizer, and transcribes the user's speech for the natural language processor. With this setup, the system is able to understand greetings and other navigational cues, and tell collaborative stories with the user.

Another important technical problem in dealing with input is the disproportionate input and output capabilities of Sam, causing an imbalance between Sam's collaborative abilities and the

user's expectation. Sam is able to generate two collaborative speech acts confidently, with control over speech communicative function, gesture, eye-gaze, and body posture. However, she is only able to recognize the *correct* speech act by the occurrence of an interruption. This imbalance may result in the user being confused or disappointed during turn-yielding. We have to hope that advances in language processing, computer vision and haptics may enable more balanced input modalities in the future.

In terms of output, due to the poor quality of speech synthesis, all of Sam's speech output is currently recorded beforehand by a voice actor, precluding a flexible and adaptive response. This is the reason why the design of Sam goes to such lengths to constrain the context of the storytelling, which as a result, detracts from the social and educational benefits of improvisational storytelling play.

Both coherence and cohesion scoring can be improved with better use of the commonsense database and other natural language processing techniques. Coherence is categorized by many factors: temporal linearity, causality, narrative structures such as goals and attempts. All these facets are embedded in the Open Mind commonsense database, however the current NLP does not distinguish between different relations. To ensure temporal linearity, we can use predicate relations such as "hasRequirement"; for causality, we can use the relations "hasConsequence", and "hasEffect"; and to generate goals, we could use the "hasWant" relation.

The natural language processing performance can be further improved by using knowledge specifically related to children's stories; and Sam has access to a growing corpus, as she accumulates more interactions with children. The initial Open Mind database had reasonable knowledge about the common locations of everyday objects, concepts relating to family and a typical home. However, the knowledge is gathered from adults, and can be sparse for concepts that are more child-specific. For example, it reacts well when asked to find the relevance between the concept *shower* and the concept *bathroom*. However, it had trouble associating different kinds of common toys, such as teddies and robots!

6 Contributions and Conclusions

Collaboration during literacy acts has been shown to improve children's literacy development. In this manuscript, we outlined a model of children's functional roles during collaborative narrative, suggested how a system can participate in such an interaction through the execution of specific speech acts and turn-taking cues, and described how such a model was implemented in Sam, our platform for collaborative storytelling with children.

The technical tools required to engage children in collaborative interactions with virtual characters are still at a fairly primitive stage. For example, there have been few reported successes with recognizing children's free speech; natural language processing tools have mostly been designed to deal with well-formed language. Artificial speech synthesis of children's voices is incomparable to the right thing.

Nonetheless, we feel that these limitations can be overcome by carefully managing the context of the interaction, and by using appropriate speech acts and turn-taking behaviors. The outcome is an important one: enabling educational systems to cooperate with children during storytelling or other learning tasks.

7 Acknowledgements

Many thanks to Henry Lieberman for his advice and common sense, and to the other members of the Gesture and Narrative Language Group at the MIT Media Lab for their input.

8 References

Allen, J., Ferguson, G., Stent, A. (2001). “An architecture for more realistic conversational systems”, *Intelligent User Interfaces*, 1-8.

Ananny, M. (2002). “Supporting Children's Collaborative Authoring: Practicing Written Literacy while Composing Oral Text", *In Proceedings of Computer-Supported Collaborative Learning Conference*, Boulder, Colorado.

Bers, M. and J. Cassell (1998). “Interactive Storytelling Systems for Children: Using Technology to Explore Language and Identity.” *Journal of Interactive Learning Research* **9**(2): 183-215.

Bakhtin, M. M. (1981). “Discourse in the novel”, *The dialogic imagination*. 259-422. Austin, TX; University of Texas Press.

Brooks, T. A. (1998). “The semantic distance model of relevance assessment”, *In Proceedings of the 61st Annual Meeting of ASIS*, 33-44. Pittsburgh, PA.

Cassell, J., Ananny, M., Basu, A., Bickmore, T., Chong, P., Mellis, D., Ryokai, K., Vilhjalmsson, H., Smith, J., Yan, H. (2000). "Shared Reality: Physical Collaboration with a Virtual Peer", *In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 259-260, Amsterdam, NL.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H. (1999). "Embodiment in Conversational Interfaces: Rea", *ACM CHI 99 Conference Proceedings*, Pittsburgh, PA.

Cassell, J. and K. Ryokai (2001). “Making Space for Voice Technologies to Support Children's Fantasy and Storytelling.” Personal Technologies **5**(3): 203-224.

Chang, J. (1998). “Action Scheduling in Humanoid Conversational Agents”, M.S. Thesis in Electrical Engineering and Computer Science. Cambridge, MA: MIT.

Damon, W. (1983). *Social and Personality Development*. New York: W. W. Norton & Company.

Darrell, T., (2002). “Evaluating look-to-talk: A gaze-aware interface in a collaborative environment”, *Proceedings of CHI 2002*. Minneapolis, MN.

Donaldson, T., Cohen, R. (1972). “Turn-taking in discourse and its application to the design of intelligent agents”, *Working Notes of the {AAAI}-96 Workshop on Agent Modeling”*, 17-23. Portland, OR.

Duncan, S. (1972). “Some signals and rules for taking speaking turns in conversations”, *Journal of Personality and Social Psychology*, Vol. 23, No. 2, 283-292.

Duncan, S. (1974). "On the structure of speaker-auditor interaction during speaking turns", *Language in Society*, vol. 3, 161-180.

Garvey, C. (1990). *Play*. Cambridge, MA: Harvard University Press.

Glass, J., Hazen, T. J., Hetherington, L., (1999). “Acoustic modeling improvements in a segment-based speech recognizer”, *Automatic Speech Recognition and Understanding Workshop*. Keystone, Colorado.

Glos, J., Cassell, J. (1997). “Rosebud: Technological toys for storytelling”, *In Proc. Of CHI ’97 Extended Abstracts*, 359-360..

Goodwin, C. (1981). “Achieving mutual orientation at turn beginning”, *Conversational Organization: Interaction between speakers and hearers*, Chap. 2, 55-89. New York: Academic Press.

Greene, S. J. (1997). “Building hypertext links in newspaper articles using semantic similarity”, *In Proceedings of the Third Workshop on applications of Natural Language to Information Systems*. 178-190, Vancouver, British Columbia.

Greene, S. J. (1999). “Building hypertext links by computing semantic similarity”, *IEEE Transactions on Knowledge and Data Engineering, 11(5)*, 713-731.

Halliday, M. A. K., Hasan, R. (1976). *Cohesion in English*. London: Longman Group.

Hoey, M. (1991). *Patterns of Lexis in Text.* Oxford University Press.

Lehman, S., Schraw, G. (2002). “Effects of coherence and relevance on shallow and deep text processing”, *Journal of Educational Psychology*, *94*, 4, 738-758.

Marsella , S. C.; Johnson, W. L.; and LaBore, C. 2000. Interactive pedagogical drama . In Proceedings of the Fourth International ConferenceonAutonomous Agents, 301#308. New York: ACM Press. http://citeseer.nj.nec.com/marsella00interactive.html

Marsella, S. C., Johnson, W. L., LaBore, C., (2000). “Interactive pedagogical drama”, *In Proceedings of the Fourth International Conference on Autonomous Agents*. New York: ACM Press.

Mostow, J. (1994). “A Prototype Reading Coach that Listens”, *National Conference on Artificial Intelligence*, 785-792.

Neuman, Roskos (1991). “Peers as literacy informants: A description of young children’s literacy conversations in play”, *Early Childhood Research Quarterly*, 6, 233-248.

Nijholt, A., Theune, M., Faas, D., Heylen, D., (2003). “The virtual storyteller: Story creation by intelligent agents”, *In: Proceedings TIDSE 03: Technologies for Interactive Digital Storytelling and Entertainment*, 204-215.

Pellegrini, A.D. (1985). “The relations between symbolic play and literate behavior: A review and critique of the empirical literature”, *Review of Educational Research, 55*, 107-121.

Piaget, J. (1962). *Play, dreams, and limitation.* New York, Norton.

Preece, A. (1992). “Collaborators and Critics: The nature and effects of peer interaction on children’s conversational narratives”, *Journal of Narrative and Life History, 2, 3*, 277-292.

Ryokai, K., Cassell, J. (1999). “Computer Support for Children’s Collaborative Fantasy Play and Storytelling”, *In Proceedings of CSCL ’99.*

Ryokai, K., Cassell, J. (1999). "StoryMat: A Play Space with Narrative Memory", *In Proceedings of IUI '99*, ACM.

Ryokai, K., Vaucelle, C., Cassell, J. (2003) "Virtual Peers as Partners in Storytelling and Literacy Learning", *Journal of Computer Assisted Learning 19(2)*: 195-208.

Sachs, J., Goldman, J., Chaille, C. (1984). "Planning in pretend play: Using language to coordinate narrative development", *The development of oral and written language in social contexts*. 119-128. Norwood, NJ: Ablex.

Sawyer, R. K. (1997). *Pretend play as improvisation: Conversation in the preschool classroom.* Norwood, NJ: Lawrence Erlbaum Associates.

Sawyer, R. K. (2002). "Improvisation and Narrative", *Narrative Inquiry, 12*(2), 319-349.

Singh, P. (2002). "The public acquisition of common sense knowledge", *In Proceedings of AAAI Spring* Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access. Palo Alto, CA, AAAI.

Trawick-Smith, J. (2001). *The play frame and the "fictional dream": The bidirectional relationship between metaplay and story writing.* Annual Meeting of the American Educational Research Association, Seattle, WA.

Vaucelle, C. (2002). Dolltalk: "A computational toy to enhance children's creativity", *In Proceedings of CHI 2002, 20-25,* ACM Press.

Wiemer-Hastings, P. (1999). "Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions", *Special Issue of Interactive Learning Environments*.

Wolf, D., Hicks, D. (1989). "The voices within narratives: The development of intertextuality in young children's stories", *Discourse Processes, 12,* 329-351.

Wood, D., Wood, H., Ainsworth, S. & O'Malley, C., (1995). "On becoming a tutor: Toward an ontogenetic model", *Cognition and Instruction, 13*(4), 565-581.