Collaborative Storytelling with an Embodied Conversational Agent

by

Austin J. Wang

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degrees of

Bachelor of Science in Computer Science and Engineering and

Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

May 21, 2003

Author_____
Department of Electrical Engineering and Computer Science
May 21, 2003

Certified by_____
Justine Cassell
Thesis Supervisor

Accepted by_____
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Collaborative Storytelling with an Embodied Conversational Agent

by

Austin J. Wang

Submitted to the

Department of Electrical Engineering and Computer Science

May 21, 2003

In Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

When children tell stories to their peers, they naturally collaborate with each other: co-authoring stories, corroborating when in doubt, and acting as active listeners. Their reliance on each other during, as well as the creative process itself, benefits their literacy development. If an interactive system were to engage a child in collaborative narrative, it would be able to exert greater influence over the child's language processes, without becoming overly intrusive as to obstruct his/her natural behaviors. However, due to the spontaneous nature of improvisational play, the problem becomes a challenging one from both a technical, and a behavioral standpoint. This thesis studies children's collaborative behaviors during storytelling and presents a model of the participants' roles, and how to initiate and participate in collaboration with appropriate speech acts and turn-taking cues. Furthermore, it demonstrates how technologies such as speech recognition, natural language processing with commonsense reasoning, multimodal interfaces, and floor management are critical to realizing a real-time collaborative interaction between children and an embodied conversational agent.

Thesis Supervisor: Justine Cassell
Title: Associate Professor, MIT Media Laboratory

Acknowledgements

This thesis would not have been possible without the following people:

Justine Cassell, my thesis advisor, who introduced me to the fascinating world of ECAs and children's literacy development, has led me to new perspectives towards both academic and personal life, and allowed me to relive kindergarten through the wonderful project of Sam. Thank you for your direction and encouragement;

Henry Lieberman, my mentor, whose expertise on commonsense and firm beliefs in his approach to research, has kept me from going astray many times. I thank him for his motivation and support;

Kimiko Ryokai, James Dai, & Anna Pandolfo, the Sam team, helped me through thick and thin with their support. I am grateful to Kimiko for her parenting, James for his resourcefulness, and Anna for her tolerance;

Hugo Liu, a friend and a constant source of inspiration, introduced me to commonsense, and since then have motivated me to pursue my own research. I thank him for his friendship and his faith in me;

Tom Stocky & Gabe Reinstein, my fellow officemates, managed to make the over-crowded office such a pleasant experience. I thank Tom for his never-diminishing sarcasm, and Gabe for his trusty companionship;

Hannes Vihjalmsson & Ian Gouldstone, taught me the intricacies of Pantomime and OpenInventor. I thank you two for your patience and guidance;

Dona Tversky, was always willing to lend an ear or a helping hand. Thanks for being understanding and accommodating;

Jae Jang, *Amir Hirsch, & Garrett Peavy*, the Sam UROPs, who basically did all the work. Thank you for letting me take all the credit.

I would like to dedicate this thesis to my parents, whose love and blessings have allowed me to pursue my dreams, and to thank my two sisters, who I depend on for support everyday.

Table of Contents

1 Introduction

1.1 Motivation

Recent years have seen the emergence of intelligent educational systems that not only facilitate, but also participate in collaborative learning interactions. This change has partly been driven by potential learning benefits: children collaborate with peers naturally, and often rely on each other for support during learning processes. At the same time, developments in the fields of speech recognition, natural language processing, and computer graphics have opened many possibilities to the interfaces of learning systems. The next challenges lie in understanding how to utilize available technology to engage children directly in collaborative interactions, and in a way as to benefit their cognitive development.

Storytelling has been an area with great promise of educational value. Researchers have confirmed that constructing narratives can bootstrap children's literacy development, and that a large portion of children stories is co-constructed. As a result, many systems were developed to encourage and facilitate children's collaborative storytelling. However, few have been designed to engage children directly as a participant due to the sheer real-time interactivity it requires: children exchange turns spontaneously, are motivated to respond for different reasons (to correct, to elaborate, to question), and produce unpredictable responses.

In order to successfully engage a child in collaborative storytelling, a system's designer must first understand how children negotiate turns. This system of turn negotiation varies between situations, and requires multiple modalities. For example, a

child telling a story will use different verbal and non-verbal cues to respond to an attempt from another child to make a suggestion and an attempt to elaborate on the story.

In addition, a collaborative storytelling system should attempt to monitor the story as it evolves during creation, so that it can make relevant additions. To give the system a better chance of understanding the child's input, one can constrain the context of the story with themes, and make use of structured interactions such as role-play.

An exaggerated effort to manage this spontaneity and unpredictability of children may come in conflict with the original goal of aiding their literacy development. Imposing too much explicit structure to the inherently improvisational process of storytelling would not only render the interaction unnatural, but also reduce its value as an educational tool. Therefore, in this thesis, the solutions to negotiating turns with children, constraining the contexts of stories, and understanding the child's input, were weighed carefully against the potential educational benefit during the design and implementation of a collaborative interaction model for Sam, a storylistening system.

1.2 Research Overview

The goal of this thesis is to implement in an existent Embodied Conversational Agent (ECA), the faculties to benefit the literacy development of a child through collaborative storytelling play. We outline a model of the functional roles of collaborators, and suggest how a storytelling system can engage a child, as well as itself, in these roles in order to scaffold the narrative. This model of roles was derived from a study conducted by Preece (1992); the speech acts and turn-taking behaviors were extracted from data collected by Ryokai et al. (2003). The interaction model was then implemented in Sam, an embodied storylistening system (Ryokai & Cassell, 1999).

Several sub-problems are investigated: how should the system participate in collaboration during storytelling; how can the system understand the child's input, and how can the system respond with natural story continuations. The solutions to these problems brought together technologies from natural user interfaces and artificial intelligence: including features like stochastic segment-based speech recognition, floor management for conversational agents, multi-modal interface with gesture and speech output, and semantic/lexical distancing using the Open Mind commonsense database (Singh, 2002).

1.3 Thesis Layout

This chapter provides the motivation to the research, as well as an overview of the work. The following chapter outlines the background to the work, in the context of collaboration during storytelling play, floor management in general, in particular during conversational systems, and storylistening systems. The third chapter describes the observations made on data collected by Ryokai et al. (2003) and Preece (1992), and the proposed model of collaborative behaviors. Chapter 4 details the system's design and implementation, including the floor management model, story structure, speech recognition, natural language processing, and commonsense reasoning. The last chapter evaluates the limitations of the system, discusses future work, and ends with conclusions and contributions.

2 Background

This section reviews the past research that is relevant to the design and implementation of collaborative storytelling systems. It begins by presenting the role of collaborative behaviors within children's storytelling play. The second section delves into the theory of floor management and turn-taking, followed by a section on its applications in conversational systems. The last section reviews existing literacy systems, and in particular, a subset of those that uses storytelling to support language development.

2.1 Collaboration during Storytelling Play

Children's collaborating with their peers during storytelling play is a natural phenomenon. Preece (1992) found that children's spontaneous stories involved collaborative telling 12% of the time; Garvey (1990) found that children engaged in focused interaction or mutual engagement during play an average of 66% of the sessions. Children's ability to collaborate also develops with age, especially during the ages of 5 to 7 (Wood, 1995). Researchers found that at the age 3, children start to engage in more "associative" and "cooperative" play (Damon, 1983).

This collaboration during storytelling play can greatly benefit literacy development. Sawyer (1997) proposed that conversational collaboration between peers is one of the most developmentally valuable characteristics of socialdramatic play. In Preece's study (1992), she found that by acting as critics, facilitators, and collaborators to each other during storytelling, children were able to produce more coherent and complex stories than they could individually. In another study conducted by Neuman (1991), researchers observed that when children played in a literacy rich environment, they

would scaffold 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; Pellegrini (1985) proposed that it is the key factor in play which affected children's literacy development. Other educational researchers have also found that the presence of peers during learning activities aid learning and development in the early years and primary education (Rogoff, 1990; Topping, 1992; Whiting & Whiting, 1975; Wood, 1996).

On closer inspection, the benefits of collaborative storytelling may depend on the nature of the interaction. Sawyer (2002) observed that improvisational storytelling without any play or narrative structure generates segments of local coherence, rather than globally coherent plots. He proposed that *meta-play* (Sachs et al., 1984; Trawick-Smith, 2001) and *dialogic strategies* (Bakhtin, 1981; Wolf & Hicks, 1989) could *scaffold* children's narratives in order to produce more complex and coherent stories.

2.2 Floor Management

Floor management is the system used by participants in verbal communication to negotiate the current speaker, and is important to achieving comprehensible and communicative dialogue between multiple parties. Goffman (1967) argues that it is essential for participants to negotiate turns in order to avoid undesirable collisions. Yngve (1970) postulated that this phenomenon "is nearly the most obvious aspect of conversation [p. 568]." Jaffe and Feldstein (1970) also emphasize the saliency of turn taking and the importance of avoiding interruption.

During face-to-face verbal communication, humans negotiate turns using many modalities, such as eye gaze, hand gestures, intonation, body posture, and head

movement (Duncan, 1972; Kendon, 1967; Goodwin, 1981). Researchers have also found that the timing and ordering of these behaviors serve to signal turns between speakers (Rosenfeld, 1978; Duncan 1974).

Additionally, floor management behaviors vary greatly from context to context. Sacks (1974) suggests that conversations, as well as other systems of verbal communication, such as ceremonies, debates, meetings, press conferences, seminars, therapy sessions, interviews, and trials, differ in the behaviors exhibited and the way they are agreed upon. The relative saliency of each behavior is also context-dependent. For example, eye-gaze plays less of a role in turn-taking during conversations between strangers (Beattie, 1980), and during discussions that impose a high cognitive load on the conversants (Rutter, 1978).

There has been some research of floor management behaviors within the context of children's storytelling. Preece (1992) observed three young children engage in spontaneous narratives and noted various verbal behaviors used for designating turns. However, non-verbal behaviors were not noted since the interaction was only recorded on audiotapes.

2.3 Floor Management in Natural Conversational Systems

Conversational systems have been a hot topic of research; much of the work is divided between applications regarding information retrieval, planning, customer service, advice-giving, and education (Beshkow, 1997; Glass, 1995; Bertenstam, 1995; Thorisson, 1996; Allen, 2001). 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 constraints 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.

2.4 Literacy Systems and Storytelling

Many intelligent literacy systems use stories as a medium to support children's language development. Some of them target certain aspects of a child's language and provide contextual feedback (Mostow, 1996; Wiemer-Hastings, 1999), others prompt for more information (Glos & Cassell, 1997). A subset of literacy systems supports children's storytelling by acting as the stage or audience for such an activity (Vaucelle, 2001; Ananny 2002; Ryokai & Cassell, 1999; Ryokai & Cassell, 1999).

2.4.1 Literacy Systems

The LISTEN project created by Mostow (1994) listens to children read stories and uses speech recognition to translate their speech as well as different aspects of their speech, such as prosody. 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 advance stories with fewer mistakes, and less frustration. In Wiemer-Hastings' (2002) project called Select-a-Kibitzer, children type in their written stories, and the systems analyses the text using natural language techniques such as latent semantics analysis, to determine the coherence, purpose, topic, 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 tries to link stories to physical objects. The system consists of a collection of stuffed animals and a computer terminal. 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.

By specifically targeting certain aspects of children's storytelling or reading skills, these systems can improve those aspects very effectively. Nonetheless, it is debatable whether the literacy of the child as improved as a whole. The effectiveness of such a bottom-up approach is dependent on whether we have managed to correctly identify the criteria for better literacy.

2.4.2 Interactive Storylistening Systems

Storylistening systems offer an alternative approach; using children's natural storytelling behavior as a basis, these systems enhance the process: children are able to annotate stories, share stories with others, replay stories, and rearrange story segments. In doing so, these systems highlight certain important facets of storytelling, such as written stories, decontextualization for an audience, temporal arrangement, and allow children to explore these facets on their own, in the absence of preconceptions of what constitutes good storytelling.

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 flipping the pages in the book.

StoryMat (Ryokai & Cassell, 1999) is a system designed to support young children's fantasy storytelling. Children sit on a large soft mat, and play with various story-eliciting shapes. They are encouraged to narrate their stories with a stuffed animal. StoryMat is embedded with sensors, and records the location and trajectory of the toy along with the child's audio. The story content as well as the actions performed in the story are stored within StoryMat, and can be recalled by placing the toy over an area of the mat used during a previous play session.

DollTalk (Vaucelle, 2001) was created as an attempt 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 stuff animals had accelerometers that monitored 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 stuff animal that was speaking at the time.

TellTale (Ananny, 2002) enabled children to create, share and edit oral language in a way similar to how they will eventually create written language. It was composed of several segments of a worm, each of which had a recording device. When the children are done recording, they can connect the pieces of the worm to hear the story they've created.

Although these interactive storylistening systems succeed in supporting children in their storytelling acts, their passive approach means they never gain much control over the child's learning process. The opposite is true with other literacy systems; their didactic interaction model is effective in influencing children literacy behaviors. At the same time, it leaves little room for them to improvise freely.

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, and that children can produce more coherent stories when they use certain collaborative strategies. Therefore, future literacy systems should not only facilitate collaboration during storytelling, but also participate in it. To this end, the next section describes a model for collaborative storytelling: starting with the functions of collaborators; it then describes how these functions can be carried out using speech acts; and ends with the turn-taking cues required to perform these speech acts.

3 Model of Collaborative Storytelling

Building upon two previous studies (Preece, 1992; Ryokai et al., 2003), a model of collaborative storytelling is proposed. This model consists of three elements: the roles of collaborative storytellers, their speech acts, and their turn-taking behaviors.

All three components of the model are necessary for creating a computational system that engages in collaborative narration. To begin with, the system's behavior is defined by the role it takes on; for example, if the system's current role were an author, its actions will be oriented towards authoring a story. Secondly, the systems for assuming and assigning these roles are complex, and require a combination of the appropriate speech acts, and turn-taking cues.

The roles identify the set of collaborative functions a participant is expected to perform; these functions are carried out with specific speech acts. However, a speech act does not uniquely map back to a role; several roles may share the same speech act. When this is the case, since most speech acts have different turn-taking behaviors, these behaviors can help determine the role of the speaker.

To summarize, only by partnering speech acts and turn-taking cues, can a system assign and recognize roles in a collaborative storytelling interaction. The first section describes the six roles of participants of collaborative storytelling play. The second section presents the speech acts and their communicative functions. The third section explains the turn-taking behaviors during these speech acts. The four section details the various turn-taking cues.

3.1 The Roles of Peer Collaborative Storytellers

There has been considerable research on the roles of adults during children's storytelling (Eisenberg, 1985; Fivush & Fromhoff, 1988; Heath, 1983; McCabe & Peterson, 1991; Miller & Sperry, 1988; Tizard & Hughes, 1984). However, the role of peers during peer-to-peer storytelling has received less attention.

Preece (1992) provides a good starting point in a study where she observed spontaneous narratives of three children on their way to school. She divided their interactions into two categories, where participants had specific roles:

- Critics (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 are bound to each other. For instance, it would be unnatural for an author and a facilitator to collaborate with each other. Therefore, it is important to properly coordinate the storytelling interaction such that the system and the user have compatible roles.

3.1.1 The Interaction between Critics and Author

When children take on the critic/author interaction, one child tends to assume the role of the primary author, and the other children act as critics, which also encompasses passive listeners. This is usually observed when one child is retelling a familiar story, or

has been nominated to create a new story. Primary authors are usually selected via facilitator and collaborator interactions.

3.1.2 The Interaction between Facilitator and Collaborators

During this interaction, the role of the facilitator is to direct and coordinate the story, and the proposed plot is negotiated with the collaborators. Both parties can narrate the story, usually after the plot is agreed upon. However, due to the dominant nature of the facilitator role, it is usually clear how the roles are distributed. For the same reason, children often compete to be the facilitator. Certain interactions show two children taking turns being the facilitator and are both involved in the direction and casting of the story. On the other hand, some children dominate the facilitator role and remain the facilitator throughout the entire story.

There is a stark distinction between the two interactions: 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 study conducted by Ryokai, Vaucelle, and Cassell (2003) showed a third type of interaction. In the study, pairs of five-year-old girls were allowed to tell 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. However, unlike the also competitive facilitator role, the primary author in a co-author interaction does structure the story by coordinating with the other author, but tries to steer the story single-handedly through narration. Again, the same distinction exists between

facilitator/collaborator and co-authors interactions: the prior uses meta-narration, while the later uses narration. We shall refer to this behavior as the relationship of co-authors:

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

3.1.3 The Interaction between Co-authors

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 coordination, 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.

All six roles were observed in the data collected by Ryokai, Vaucelle, and Cassell (2003). 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 occur. 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.

To demonstrate how this can be done, before going into the specifics, 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 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.

Table 1 – Relationships between participants during collaborative storytelling.

Relationship	**Configuration**	**3.1.3.1.1 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

3.2 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.

3.2.1 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.

  R: "OK. And she was sitting down. She got up, and she said, mom, dad, where are you? Mom, dad, where are you?"

S: “Go to your magic mirror maybe.” <Suggestion>

- *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.

  R: “And when she came down she saw her mom and daddy.”

  S: “No. Just her mom.” <Correction>

- *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 supplement information.

  S: “I want yummy, yummy. Me want Harry Potter.”

  R: “Harry Potter. The prince was Harry Potter, right?” <Question>

- *Answer* – the speech act that answers the question by providing the information requested. In all cases observed, the person who the question was directed towards was responsible for answering; however, it is not to say that other participants may not answer if there were more than two of them.

  R: “She saw just what she had been looking for that very same night.”

  S: “The gooey?“

  R: “No. She wasn't looking for that thing.” <Answer>

- *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.

  S: “Go to your magic mirror maybe.”

R: “Ooo. I think I should go to my magic mirror.” <Acknowledgement>

3.2.2 Speech Acts between Facilitator and Collaborators

These 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. They are also responsible for narrating the proposed plots, these speech acts are different from other narrations in that they adhere to a predetermined script.

- *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).

  R: “Pretend she went right, and she got eaten by the claws devil, but she escapes.” <Direct>

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

  See example below.

- *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.

  S: “Rachel, but pretend she gets eaten, but she escapes the monster's mouth.”

  R: “She's eaten.” <Acknowledge>

  R: ”And then, when her mom and dad come home they say, oh. Where's Annie? Where's Annie? Oh no. Where is she? And, then, she says, hmmm.

I think they don't love me any more, because she escaped that alligator's mouth." <Elaborate>

3.2.3 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.

  S: "Honey honey, what's happened? What's happened?" <Role-play>

  R: "I fell out of the house." <Role-play>

  S: "Ooo we better get the ambulance, Cary Cary sweetie come!" <Role-play>

- *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.

  S: "Ah hmmm. He got eaten. She said [SCREAM], the evil monster has been here. Oh, my husband. Oh, there he is. There he is on the top of the roof."

  R: "That's a monster, monster, disguised." <Simultaneous turns>

  S: "Honey, I'm right here. But he didn't hear her, so she jumped up. [LAUGHTER] She's a spy kid, actually."

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 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), I identified ten types of collaborative speech acts that resulted in turns being exchanged. They are presented in the 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**	**3.2.3.1.1**	**Intention/Function**	**Turn-taking behaviors**
Critics and authors	Suggest	Critic	To suggest an event or idea to the story	Eye gaze towards author, author may show 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, higher pitch ending, syntax of question, author stops gesturing
	Acknow-ledge	Author	To acknowledge a suggestion or a correction	Eye gaze towards critic, backchannel feedback like "mm-hmm", author stops gesturing

Facilitator and collaborator	Direct	Facilitator	To suggest storylines and to designate roles	Eye gaze towards collaborator, socio-centric sequences like "OK", both stops gesturing
	Acknow-ledge	Collaborator	To acknowledge a role designation or storyline suggestion	Eye gaze towards facilitator, backchannel feedback like head nods, both stops 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 play act, syntax marking the end of a grammatical clause, prosody of in character voice, gesture with prop
	Simultan-eous turns	Both	Compete for the turn	

3.3 Turn-taking Behaviors during Speech Acts

This section presents the turn-taking behaviors for each speech act; the following section goes into detail about the individual behaviors.

3.3.1 Turn-taking Behaviors between Critics and Author

There are certain turn-taking cues that are common to all critic and author speech acts: at turn intervals, critics and authors shift their eye gaze towards each other, and away from the play act itself; they also stop gesturing with their props, and may rotate their bodies to face each other.

- *Suggestion* – suggestions can be solicited explicitly with request, or implicitly with hesitation. Turn-taking signals include paralanguage drawls, and socio-centric sequences like "mmm", and "uhh". Duncan (1972) referred to these sequences as stereotyped expressions which are

observed before a speaker yields his/her turn, but they do not add any substantive information.

- *Correction* – Critics do not express turn-taking signals prior to corrections, they simply start speaking if they want to make a correction. However, they do maintain eye contact with the author.
- *Question* – The turn-taking behaviors that occur before posing a question include looking at the author, lack of backchannel feedback like head nods, and uneasiness signaled by increased body motion.
- *Answer* – The person answering establishes eye contact with the questioner and takes the turn when s/he sees the completion of the question in terms of syntax (direct and indirect questions) and intonation (higher pitch).
- *Acknowledge* – the author establishes eye contact with the critic, and may supplement verbal acknowledgement with backchannel feedback such as head nods and "mm-hmm".

3.3.2 Turn-taking Behaviors between Facilitator and Collaborators

During facilitator and collaborators interactions, children can communicate with either a narrator's or a play character's voice, and they maintain eye contact unless during elaboration. Children stop gesturing with their props during the direct and acknowledge speech acts, but they often gesture when elaborating. They may also rotate their bodies to face each other during turn changes.

- *Direct* – The facilitator maintains eye contact during, and sometimes ends with socio-centric sequences like "ok" to signify that s/he is expecting an acknowledgement from collaborators.
- *Acknowledge* – Collaborators can acknowledge by providing backchannel feedback, or by agreeing verbally.
- *Elaborate* – Their gaze is directed towards the play act instead of each other, and they may use both a narrator's and a play character's voice.

3.3.3 Turn-taking Behaviors between Co-authors

There are relatively fewer turn-taking cues for these speech acts. Cues during role-playing are mostly syntactical, and maybe include changes in prosody and intonation. However, there are virtually no turn-taking cues during simultaneous turns, which make this speech act hard to recognize and employ. The language is mostly in-character for role-play, and can be both narrator or play-character for simultaneous turns. Co-authors are usually concentrated on the play act, and do not alter their gaze or body positions during turns.

- *Role-play* – Role-play is usually improvisational, and the turn-taking cues vary and are usually subtle. Syntax and prosody can help identify the end of a frame; sometimes children may stop gesturing with their props when they yield their turns.
- *Simultaneous turns* – Since both children are speaking concurrently, they are not respecting each other's turns. Turn-taking cues are seldom observed during simultaneous turns.

3.4 Turn-taking Cues

The following sections illustrate each turn-taking cue in detail. I build upon the turn-taking signals proposed by Duncan (1972), and give examples of these signals found within the data collected by Ryokai, Vaucelle, and Cassell (2003).

3.4.1 Intonation

Intonation is the change of pitch, and is often used at the end of a question, or right before a suggestion, when the person is hesitating. For example:

Child: Can you get the chandelier back on? <Raised pitch>

Child: Where's my room? Hmm. <lowered pitch>

3.4.2 Eye Gaze

Eye gaze is one of the most common ways to yield a turn to someone else. During speech acts such as role-play and story direction, the speaker often shifts gaze away from the prop to the audience as a turn-yielding signal. A child will often accompany acknowledgements with eye contact with the other child.

3.4.3 Syntax

When co-authoring, children only start elaborating when the other has completed a grammatical clause. It can also be useful in role-play; in the next example, a child directs a question at the magic mirror, which happens to be the character played by the other child. This successfully transferred the turn to the other child.

Child: And she said to the magic mirror, how will I ever get them back again?

3.4.4 Socio-centric Sequences

These are stereotyped expressions that often follow a substantive statement (Duncan, 1972). Examples are “and then”, “hmm”, “uh huh”. These behaviors are found during suggestion speech acts, when the speaker uses them to show hesitation; they are often accompanied by the paralanguage “trailing off” effect of the speaker’s voice.

3.4.5 Paralanguage

Paralanguage signals are variations in the sound of the speech. Examples include prosody, drawl, and loudness. This phenomenon is extremely common in the context of storytelling: when playing fantasy roles, when issuing commands, and so on. In terms of turn-taking, the paralanguage drawl often signals for a suggestion, and the meta-narratives during the speech act of story direction usually use a more authoritative prosody.

3.4.6 Backchannel Feedback

The audience uses backchannel feedback to acknowledge what the author said; the behaviors include utterances such as “mm-hmm”, and gestures such as head-nods. It is useful during acknowledgements. An audience may signal that they wish to ask a question by not displaying backchannel feedback.

3.4.7 Body Motion and Gestures

This includes the movement of body parts, as well as physical artifacts such as toys. Gestures behaviors are extremely frequent when children are using props; the starting and stopping of these gestures can clearly mark the being and end of one’s turn.

Given that these turn-taking behaviors are based on Duncan's observations on adults, children's uses for them are different in several ways. Children seem to use body motion and gestures much more frequently; this may be due to the play element of storytelling, where children often act out their stories with props, they also seem more comfortable than adults are to gesture with their entire bodies.

Eye gaze is also used frequently; they switch between looking at the play act and each other, and helps us determine the roles they are playing. Intonation and prosody is another integral part of storytelling; children use these to switch between in-character and out-of-character narrations.

Children use these turn-taking cues and speech acts to coordinate their storytelling, and increase the interactivity of the process. The roles provide a local script such that the responsibilities and functions of the parties are well defined. For these reasons, I have chosen to follow the same approach when designing a system that participates in collaborative storytelling.

4 Implementation

According to the floor management model described above, a subset of speech acts and their corresponding turn-taking behaviors were chosen and implemented into an existing storylistening system, *Sam* (Cassell et al., 2000). Extensive changes were made to Sam, in order to broaden its storytelling abilities to accommodate a more collaborative interaction. Sam's prerecorded stories were modified to support non-linear narratives; each story has multiple paths and endings. The turn-taking strategies were retuned for a more spontaneous and dynamic interaction. The speech recognition was added so that children can collaborate with speech acts and verbal turn-taking behaviors that are natural to them. At the same time, considerably effort was taken to constrain the interaction such that it would be manageable by an autonomous agent.

The first section introduces Sam, and the physical interface of the system. The second section describes the collaborative roles that Sam tries to assume, and explains why only a subset of these were implemented.

The next three sections present the solutions to the three sub-problems targeted by this thesis: participating in collaborative storytelling by pairing speech acts with turn-taking behaviors; translating their input using speech recognition; and responding cohesively using natural language processing with commonsense reasoning. The two sections after that present the interaction aspects of Sam: the play interface, and the design of the stories. The last section gives a summary of the current state of the implementation.

4.1 Sam: A Storylistening System

Sam was designed to engage children in the act of storytelling and listening with the aim of aiding their literacy development. Sam is composed of two sections, a virtual 3D character that is a cartoon rendition of a 6 year old, and a wooden toy house. The character is displayed on a large (40 inch) plasma display positioned behind the house, and results in a body height and size that is realistic for a 6 year old. (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 two 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 will be referred to as if she were female from this point on.

Figure 1 – Sam greeting

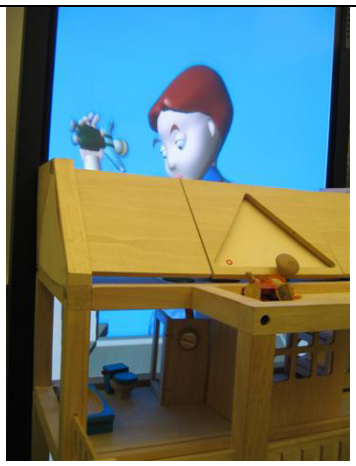

Figure 2 – Sam gesturing with figurine

4.2 Collaborative Interactions and Roles

Sam is able to collaborate by assuming three of the six 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 *correct* 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.

In addition to generating speech acts and turn-taking behaviors, it is also important for Sam to recognize them, such that the collaboration is balanced. Without access to many of the modalities where these cues occur, such as eye-gaze, gesture, and

so on, Sam has to rely on the verbal channel, which is why the *correct* speech act was chosen to represent the author/critic interaction. It is possible to identify a *correct* speech act by its communicative function; however, it is much easier to detect the turn-taking behaviors used. When children make corrections, they simply start speaking; therefore, Sam only needs to monitor the audio-in channel, and if she detects speech from the child during her turn as an author, she will interpret the input as a correction.

The other two speech acts are generated. The *direct* speech act was chosen because of its dominant nature: the chance that a child will respond to a dominant speech act is higher than that of a passive speech act. The *role-play* speech act was opted over *simultaneous turns* for the co-author role because the turn-taking behaviors are more manageable. Since there are no well-defined turn-taking cues for *simultaneous turns*, Sam would have a hard time giving the turn to the child.

4.3 Collaborative Storytelling: Speech Acts & Turn-taking

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.

4.3.1 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 syntactically 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 continues 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".

4.3.2 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 turn-taking behaviors that have increasing severities. 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.

4.3.3 Multi-modal Interface

To perform the various turn-taking cues described above, Sam requires a multi-modal interface: she 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.

4.4 Translating Children's Input: Speech Recognition

This part of the project is still under development, as follows:

The speech recognition engine being developed is 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). Acoustic data of children's stories is being collected and the final acoustic model should contain roughly 120 minutes worth of utterances. A sophisticated noise model will be incorporated into the grammar, and includes 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 would like to extract as much semantic information from the child's input as possible. Having a complete transcript will aid the natural language processor to do so. Therefore, during the child's turn, the speech recognizer operates in dictation mode, with a grammar of several thousand phrases.

4.5 Responding Naturally: Natural Language Processing

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 & Hsan, 1976) to the child's input. Although the importance of coherence is constantly emphasized over that of cohesion, during the authoring of non-linear, or hypertextual, narratives (Foltz, 1996), Sawyer (1997) observed that children's improvisational narrative were rarely globally coherent. However, 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.

The natural language processor has two main functions: to interpret the child's speech as transcribed by the speech recognizer, and suggest the most appropriate story segment within Sam's repertoire. The first function is accomplished by extracting keywords from the speech recognizer's output with the help of a part-of-speech tagger. The second function involves comparing the semantic/lexical distances between the

extracted keywords and the pre-defined keywords that categorize the various segments in Sam's repertoire, using a commonsense knowledgebase.

4.5.1 Keyword Extraction with Part-of-speech Tagging

Keywords are defined as the verbs, nouns, and adjectives in each sentence, and are extracted by a Brill based part-of-speech tagger (Brill, 1995). The POS tags are part of the Penn Treebank tagset (Marus et al., 1993), and the tagger admits all forms of nouns, verbs, and adjectives as keywords if the word is not included in the list of stop-words. These three parts of speech were chosen because of their relatively high semantic value. The list of stop-words is meant to block frequent false positives such as "until", "soon", and keywords that have little semantic value like "is", "went", "something" and so on. The tagger itself is rule-based, and is trained via Brill's transformational-based learning approach; the original lexicon and rules are incorporated into this Java version.

4.5.2 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, and makes it the perfect 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.

4.5.2.1 Lexical distance

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.

4.5.2.2 Semantic distance

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.

4.5.2.3 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 Adding keyword: plant Adding keyword: flowers Adding keyword: happens Adding keyword: planting Adding keyword: right Adding keyword: realizes Adding keyword: school Adding keyword: bye Adding keyword: leaves Adding keyword: teacher	NODE: ghost EDGE: PRED: hasEffect TARGET: fear SENTENCE: the effect of seeing a ghost is feeling fear DIRECTION: fw WEIGHT: 0.5 ********** NODE: gift EDGE: 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

*4.5.2.4 Coherence and semantic d*istance

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.

4.5.2.5 Cohesion and lexical distance

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.

4.5.2.6 Story segment scoring

The metric for scoring the story segments combines these two distances 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.

4.5.2.7 Story segment selection

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).

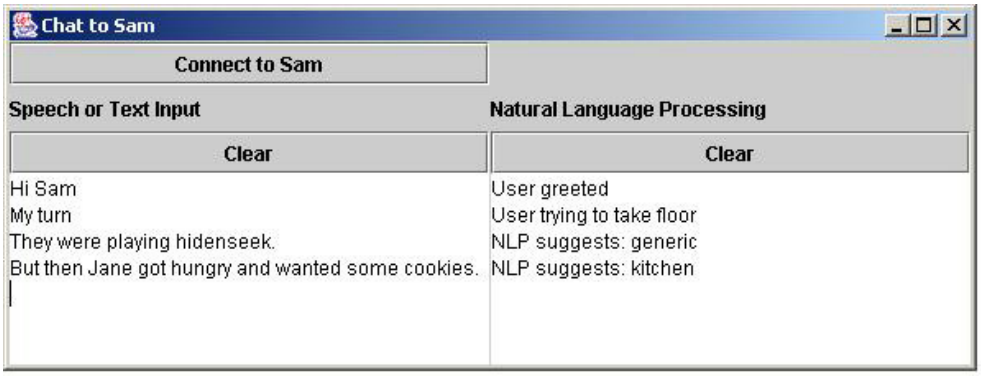

Figure 3 – Screen shot of natural language processor interface.

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 NLP still recommended the generic segment; however, after the second sentence, the scores were spread out enough that the kitchen segment was recommended.

4.6 Playing with Sam

This section describes the various storytelling modes that Sam is able to engage a child in, and how a child would go about playing with Sam. Sam originally engaged children 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.

4.7 Story Design

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's a summary of a typical collaborative story:

1. Sam starts by introducing the characters and the scene of the story. E.g., "Once upon a time, there was a boy named Brad and his best friend Jane."
2. Sam then develops the story by describing some initial events. E.g., "Jane was staying over one time, and they wanted to play hide and seek."
3. Sam then attempts to engage the child as a co-author, using role-play invitations like, "And Jane said…"
    a. If the child responds, then Sam listens and analyzes the content of the child's speech.

b. If the child does not respond for a long period, Sam tries to become the facilitator, and offers a synopsis so that the child may elaborate: "Let's pretend Jane runs into the kitchen first and tries to hide there. She couldn't find a good place so she runs into the Brad's bedroom. Ok?"
    i. If the child elaborates on the story, then Sam listens and analyzes the speech.
    ii. If there is no response for a while, Sam elaborates and ends the story.

If the child responded when Sam offered a turn, or if s/he corrected Sam by interrupting, and therefore acted as a critic, Sam will analyze the speech input and respond with the most cohesive and locally coherent story segment.

4.8 Current State of Implementation

To summarize, Sam is able to perform all the input, output, and processing functions described above, except for speech recognition. However, since the speech recognition is responsible for detecting greetings, navigational cues, as well as understanding the child's collaborative segments, Sam is unable to interact with a child without the use of Wizard of Oz (WOZ). During WOZ operation, an operator controls Sam remotely as if she were a puppet, by listening to the child's voice, and explicitly commandeering Sam's verbal and non-verbal behaviors.

In order to allow all of Sam's modules to function, an adult user can 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.

5 Limitations and Future Work

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. 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.

On the other hand, we have a robust model of the collaborative behaviors between children from the two studies, albeit in very specific contexts: two or three peers, who are familiar with each other or at least acquaintances. The particular sample was chosen because the school environment presents a typical scenario for a system such as Sam. Encouragingly, the subjects did range from several ethnic groups and had different

language backgrounds, providing some support for the generality of the behaviors observed. Nonetheless, it is possible that the collaboration model differs in other contexts. For example, the number of participants may have an effect on the roles they take on.

A possible extension to the interaction model of Sam is the addition of social intelligence through user profiling. During the study, we observed that some children favored being the facilitator, while others collaborated more often. It may be possible for Sam to detect this preference, in order to offer the preferred role to the child. A more in-depth study on children's collaborative behavior under different scenarios will help address these uncertainties.

5.2 Technical Limitations

Almost all of the technologies involved in the implementation of Sam are undergoing intensive development. In addition to the challenges posed by the spontaneity and unpredictability of a child, Sam's multi-modal interface, speech recognition and output, natural language processing, and commonsense reasoning, all provide a healthy dose of technical limitations.

An important problem 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, however, advances in language processing, computer vision and haptics may enable a more balanced input modalities in the future.

The second largest hindrance due to inept technology is the limited coverage offered by the prerecorded stories. Due to the poor quality of speech synthesis, all of Sam's speech output is recorded before hand, precluding a flexible and adaptive speech output. This is the reason why the design of Sam is forced to goes to such lengths to restrain the context of the storytelling, and as a result, detracts from the social and educational benefits of improvisational storytelling play.

As opposed to improving the way we constrain the context in Sam, we should move towards the generative speech output paradigm. Partnered with speech synthesis, semantic and syntactic models for speech acts, a commonsense corpus like Open Mind can potentially generate natural coherent, cohesive, and collaborative responses in real-time.

In terms of input, the deficiency in the existing implementation is the speech recognizer. The recognizer is still being developed and has yet to be tested, but is expected to have an error rate of no lower than 30% at full dictation mode, where the grammar is around 1000 words. This level of error can have drastic effects on Sam's ability to understand the user's story and respond coherently. However, incremental refinements can be made to the acoustic and language model by collecting speech data of children telling stories with Sam.

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 facets: 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 knowledge-based approach to part-of-speech tagging and coherent and cohesion estimation is powerful, but at the same time suffers from an inherent weakness. Neither of the knowledge bases were specifically designed for children’s applications, which means that although their knowledge can be both broad and deep, they are broad and deep in the wrong areas.

The current implementation of the keywords extraction algorithm is not robust enough to handle children’s spoken improvisational narrative. The POS tagger was never trained on a children’s corpus, and the right type of corpus for this application is rare (unlike a corpus of children’s written stories).

Although the Open Mind project was open to the public, the majority of its contributors were adults, who talked about topics pertinent to adults. The resultant knowledge is fairly skewed from our desired area, and pruning the database will only speed up the processing time, but not improve the relevance of the knowledge.

The obvious solution would be to improve the knowledge-base; and as Sam gradually accumulates transcripts of interactions with children, these foundations would automatically be improved. 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.

In terms of recognizing other speech acts, Sam's narrow grammar, which consists of greetings, farewells and turn-related speech acts, improves the recognition rate greatly. The tradeoff between good recognition and broad coverage means that Sam will not able to recognize any equivalent speech acts that are outside of her grammar: a near miss is treated by the speech recognizer as a false. This means that unless children greet Sam exactly as she expected them, Sam will not interpret it as a greeting. One way would be to improve the speech recognition overall, but there will always be a tradeoff between coverage and accuracy. Another approach would be to offload the communication responsibilities to other channels, and rely more on gestures to convey greetings and turn-taking cues.

With such limited types of speech acts, the interaction can become too singular. Children have been observed to chat with Sam, and have even asked 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 may be worthwhile to investigate how children maintain relationships with storytelling partners over long periods.

6 Contributions and Conclusions

Collaboration during literacy acts has been shown to improve children's literacy development. In this thesis, 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 system was implemented in Sam, our prototype collaborative storytelling system.

The technical tools required to engage children in collaborative interactions with virtual characters are still in primitive stages. 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, I feel that there is both technical these limitations in sensing and output can be overcome by carefully managing the context of the interaction, and by using appropriate speech acts and turn-taking behaviors. Furthermore, the system of collaborative behaviors can enable educational systems to cooperate with children during storytelling or other literacy tasks.

7 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.

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

Beattie, G. (1980). “The role of language production processes in the organization of behavior in face-to-face interaction”, *Language Production*, Vol.1, 69-107.

Bendford S., et al. (2000). “Designing Storytelling Technologies to Encourage Collaboration Between Young Children”, *In ACM CHI ’00 Conference Proceedings*, The Hague, The Netherlands.

Bertenstam, J., and Beskow, J. (1995). “The Waxholm system – a progress report”, *Proceedings of Spoken Dialogue Systems*, Vigsø, Denmark.

Beskow, J. (1997). “Animation of Talking Agents”, *In Proc. AVSP*, p149-152, Rhodes, Greece.

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.

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.

Foltz, P.W. (1996). "Comprehension, coherence and strategies in hypertext and linear text", *Hypertext and Cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Glass, J. (1995). "Multilingual spoken-language understanding in the MIT Voyager system", *Speech Communication*, vol. 17, 1-18.

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

Goffman, E. (1967). *Interaction ritual: Essays on face-to-face behavior*. Garden City, New York: Doubleday.

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 Workship on applications of Natural Language to Information Systems*. 178-190, Vancouver, British Columbia.

Greene, S. J. (1999). "Buidling 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.

Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue*. New York: Academic Press.

Kendon, A. (1972). "Some relationships between body motion and speech: an analysis of an example", *Studies in dyadic communication*. New York: Pergamon 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.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). "Building a Large Annotated Corpus of English: The Penn Treebank", *in Computational Linguistics*, *19*, 2, 313-330.

Martlew, M. (1983). "The Psychology of written language: developmental and educational perspectives", *Wiley series in developmental psychology*, New York: Wiley.

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.

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

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.

Peterson, C. & McCabe A. (1983). *Developmental psycholinguistics: Three ways of looking at a child's narrative*. New York: Plenum

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.

Robertson, J. and Wiemer-Hastings, P. (2002). "Feedback on children's stories via multiple interface agents", *In proceedings of International Conference on Intelligent Tutoring Systems*, Biaritz.

Rogoff, T. (1990). *Apprenticeship in Thinking: Cognitive development in social context*. Oxford University Press, Oxford.

Rutter, D., Stephenson, G., Ayling, K., and White, P. (1978). "The timing of looks in dyadic conversation". *British Journal of Social and Clinical Psychology, 17*, 17-21.

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.

Sacks, H. Schegloff, E. A., & Jefferson, G. (1974). "A simplest systematics for the organization of turn-taking for conversation", *Language*, *50*, 696-735.

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.

Teale, W. H. and E. Sulzby (1986). "Emergent literacy as a perspective for examining how young children become writer and readers", *Emergent Literacy: writing and reading.* Norwood, NJ: Ablex.

Thorisson, K. R. (1996). *Communicative Humanoids: A computational Model of Psychosocial Dialogue Skills*. PhD Thesis, MIT Media Laboratory.

Topping, K. (1992). "Cooperative learning and peer tutoring: An overview", *The Psychologist*, 5(4), 151-157.

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.

Whiting, B. & Whiting, J. (1975). *Children of six cultures*. Cambridge, MA: Harvard University 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.

Wood, D. & O'Malley, C. (1996). "Collaborative learning between peers: An overview", *Educational Psychology in Practice*, 11(4), 4-9.

Yngve, V. G. (1970). “On getting a word in edgewise”, *Papers from the sixth regional meeting, Chicago Linguistic Society*. Chicago: University of Chicago Department of Linguistics.

8 Appendix

Three stories from Sam's repertoire are attached. The possible story continuations after each turn-yielding attempt are also given.

Story 1 (Sam has boy figurine)

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…

(give floor to child)

If no response, 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?

If no response for a second time, Sam:
Ok, I'll tell the story. So Mary runs to the kitchen and looks around, she sees a blue closet and she thinks ooo, Jack will never find me there. She tries to get in but she won't fit inside. Jack has finished counting and is starting to look for her. Mary's scared so she quickly runs into Jack's room and hides in his bed.

If the child does respond, depending on what the child says, Sam will give one of the following responses:

Kitchen:
Jack looks everywhere for Mary, in the bedrooms, in the bathroom, but he couldn't find Mary. He goes into the kitchen, 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

Bathroom, Bedroom – same but substitute correct location

Generic response:

Then all of a sudden, 'BANG BANG BANG!!!'. There was a loud bang from outside the house. Mary ran to see what it was, and it was Jack, standing there with his drum and drum stick. 'haha, gotcha' he says. He tricked Mary into coming out of the house, what a naughty naughty boy. The end.

Story 2 (Sam has boy figurine)

OK. I'm going to start. Once upon a time, there was a little boy named Fred and a girl named Jane. They were friends and they met up every Sunday morning to go play in the fields.
One Sunday, Fred woke up early and realized his pet frog was gone. His pet frog's name was Jared and he loved it very much. He checked the jar Jared normally sleeps in, but he wasn't there. He then went downstairs and looked everywhere, inside his boots, in the toilet, in the sink, but he couldn't find Jared.
Then he heard a knock on the front door, and it was Jane. 'Hi Jane. Jared has gone missing, I don't know where he is. Do you think he's inside the house or out in the field?'

And Jane said…, Jane said…

(give floor to child)

If no response, Sam:
Let's pretend Jane starts to go look for Jared with Fred, and they walk through the fields and then the forest, and through the swamp but they still couldn't find him. Ok?

If no response for a second time, Sam:
Ok, I'll tell the story. So Jane says: 'ok, lets look for him in the field', and Fred says, 'ok' So they walk through the field together shouting: 'Jared! Jared!' They saw a deer grazing the field, and they asked the deer whether he had seen Jared, he said nope. So they kept walking and they came to a forest. They saw an owl on the tree, and asked him whether he had seen Jared. The owl said no, but they should go try the swamp. She saw lots of frogs there.

If the child does respond, depending on what the child says, Sam will give one of the following responses:

Outside:
Finally, they came to the swamp, and they saw Jared there with lots of other frogs. The other frogs were Jared's family. Fred was so happy, they spent the rest of the afternoon playing in the mud with Jared and his cousins. The end.

Inside:
Fred and Jane are so happy when the found Jared in the house. Fred told Jared never to hide again, and they went outside and went swimming in the pool. The end.

Generic response:
Fred and Jane were so happy when the found Jared. Jared took them to a pond where they could go swimming. They swam for the rest of the afternoon. They Jane got tired and

said she needed to go home. Fred said good bye, and took Jared back home. He tucked Jared into bed and told him never to escape again. The end.

Story 3 (Sam has girl figurine)

Want to tell a story together? Ok, let's pretend once there was a doctor who worked in the hospital. Her name was Sara, and every kid loved her because she was kind and clever. They would go to her when they don't feel well, or when they hurt themselves, and she would always help them get better.

There was a young boy who also lived in the town, and his name was Jason. Jason was a troublemaker, he loved to scare people by hiding around corners and jumping out at the last moment and shouting "Boooo!" and playing jokes on other people like hiding their things from them.

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…

Give turn!

If not:
Let me see. Let's pretend Jason was home and he was doing something naughty and he got hurt?

If not again:
OK. Let me go. Jason was playing at home and he was being very naughty. He was hiding in this box so he could jump out and scare his mom when she walked by. But his mom was selling that box that day and the movers came to carry it out of the house. Jason jumped out while they were on the stairs and fell down the stairs, and he broke his leg!

Injury:
So Sara said: "Oh dear, we'll need to have an operation on that. Be more careful next time!" Then Sara gave Jason some medicine for the pain, and put him in the emergency room. The operation was a success and his parents were so pleased. But Jason was a good boy after that, he never tried to play jokes on people again. The end.

Sickness:
"oh dear, you don't look well. let me get you some medicine for that."So Sara got him some medicine and gave him an injection, but it didn't hurt since she was really careful. The end.

Generic response:
The nurses took Jason to bed, and fed him hospital food. But he hated hospital food so he yelled and yelled. Doctor Sara said he could go home to his mommy and daddy, and he was happy because he was better and he could eat his mommy's food again. The end.