

# Interacting with an Animated Conversational Agent in a 3D Graphical Setting

Andrea Corradini  
Southern Denmark University  
Dept of Math & Computer Science  
5230 Odense M, Denmark  
andrea@imada.sdu.dk

Manish Mehta  
Georgia Institute of Technology  
College of Computing  
Atlanta, Georgia 30332, USA  
mehtama1@cc.gatech.edu

Marcela Charfuelan  
Saarland University  
Spoken Language Systems  
66041 Saarbruecken, Germany  
marcela@lsv.uni-saarland.de

## ABSTRACT

This paper presents the main components in the development of a conversational character within the context of an educational computer game. Speech is the main modality used for interacting with the lifelike animated agent but the user can also draw a set of 2D gestural markers to refer to on-screen objects that the agent can talk about and manipulate upon.

Our system attempts to cope with many challenging issues such as the maintenance of an entertaining conversation, the management and the conveyance of multimodal behaviors complete with accompanying human affective and cognitive attributes, and the correct provision of information for educational purposes.

A user study carried out with a group of thirteen kids and directed towards gathering insights about different aspects of the interaction shows a high degree of satisfaction in the use and appearance of our system interface.

## Categories and Subject Descriptors

H.5 [Information Interfaces and Presentation (e.g., HCI)]: Multimedia Information Systems – *animations*; User Interfaces – *natural language, user-centered design*.

## General Terms

Design, Experimentation, Human Factors.

## Keywords

Edutainment, embodied conversational agent, multimodal speech-gesture system, computer game.

## 1. INTRODUCTION

In science-fiction movies, it is not uncommon to see humans who smoothly interact with machines via spoken dialogue. The HAL 9000 computer in the movie 2001: A Space Odyssey is probably the most famous example of such a machine capable of advanced speech processing and perception of the surrounding. However, regarding the current computers' capabilities, Kuck's [17] words "*..With the 20/20 vision of hindsight, we can examine the specific predictions of both the book and the film versions of 2001 and*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission or/and a fee.

*Int'l Workshop on Multimodal Interaction for the Visualization and Exploration of Scientific Data*, at the *International Conference on Multimodal Interfaces (ICMI '05)*, October 3, 2005, Trento, Italy.

Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00.

*observe now that some of them were far too optimistic..*" reveal what the human-computer interaction community has been long knowing: HAL was and still is unquestionably ahead of time.

There are many issues that need to be addressed and solved before computer interfaces can properly become conversational and multimodal. Question-answering systems and frame-based dialogues [2, 13, 28, 32] have been mainly put forward to date. Their main limitation, its fixed context, is simultaneously its greatest strength since it allows building very robust and feasible spoken systems. However, they are a simplification of real human conversational behavior, for they control and restrict the interaction rather than enrich it. On the contrary, non-task oriented dialogue systems allow for a less constrained conversation where the user is free to address any topic in any order within the system's knowledge domain. Information seeking/extraction systems [24, 29] and embodied conversational agents (ECA) [4] belong to this category.

ECAs have attracted a considerable amount of interest and attention over the last years. This emerging interface paradigm is aimed at emphasizing properties of and emulating human-human communication styles that are expected to improve the intuitiveness and effectiveness of user interfaces. Supported by significant progress in 3D graphics and by the increase of computational power, several animated agents have been developed for the most disparate applications ranging from information presentation and online sales to entertainment and tutoring [4, 14, 25]. While individual aspects of animated agents, such as their graphical appearance and quality of synthesized voice, have been improved there are still many limitations concerning the degree of social conversation they can maintain. Particularly when the focus is on computer games, research in conversation and language understanding is rare [10, 22] and mainly concentrates on chatting bots [12, 16, 20] based on augmented versions of the pioneeristic Eliza system [31].

We propose a conversational agent capable of maintaining non-task oriented conversational with 10-18 year old children. The character impersonates the writer H.C. Andersen whereby the domain specific knowledge base includes terms, proper names, geographic reference, and other general information about the character's fairy tales, his real life, his physical presence and his study. Kids are free to address in English, in any order, any topics within the knowledge domains of our agent, using spontaneous speech and mixed-initiative dialogue. They can also enter gestural pen markers to provide additional context to the verbal interaction. Synthetic Interviews [19], with Einstein's talking head is the system that most resembles our work.

To set the scene, in the next section we present an overview of the whole system architecture and briefly focus on a description of its main components. We further set forth the description of a field study we have run to assess system ease of use, interactivity, entertaining and educational value. Eventually, we report on the main limitations of our system and conclusions.

## 2. INTERACTING WITH THE AGENT

### 2.1 Architecture Overview

Figure 1 shows a sketch of our system. The software architecture is made up of a set of distributed agents that communicate with each other via TCP/IP by means of a central facilitator whose task is to route messages among different modules. When the user utters a sentence the speech recognizer generates a time-stamped n-best hypothesis list for the natural language understanding (NLU) to process. Similarly, if the user enters gestural pen markers a gesture recognizer module produces a probabilistic hypothesis list to be forwarded to the gesture interpreter. The output from both the NLU and the gesture interpreter is then processed by the input fusion that attempts to extract a unique representation of the input utterance for the character module (CM) to process. The CM is responsible for dialogue planning, emotional state updates and maintenance of the conversational history. Eventually, the response generation (RG) module coordinates the production of synchronized text-to-speech and graphical animations output.

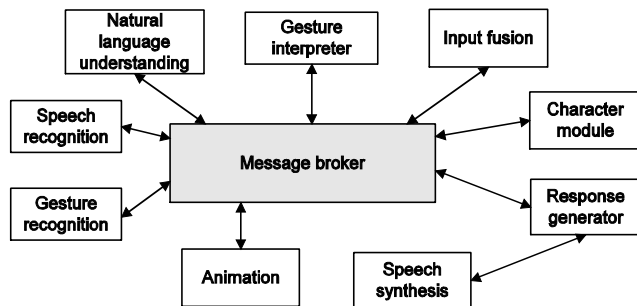


Figure 1. Sketch of the system architecture.

We use an ontology as common knowledge representation shared among modules to have a domain independent system architecture so that moving to another character would only require a modification of the ontology-based knowledge representation.

### 2.2 Children Speech Processing

Due to the nature of human language, a spoken dialogue system has to cope with utterance variation, noisy input speech data, and a variety of disfluencies. The designers of speech processing systems usually train their speech recognizers with acoustic data of adult human speakers that account to the largest possible extent for these issues. Unfortunately, differences in vocal tract size along with accentuated intraspeaker and interspeaker variability [18] cause the acoustic and linguistic characteristics of children's spoken language to be considerably different from those of adults. Recognition of children's non-native speech adds further problems to the existing issues and results in poor speech recognition results. Down the processing pipeline, any poor performance of the speech recognizer becomes hard to recover because the traditional syntax-driven parser and rigid grammars will not work

well. Any attempt to perform complete syntactic analysis to account for all words in an utterance will break down.

In our approach, we make use of an acoustic model created using many hours of data of kids spoken language collected during several Wizard of Oz sessions with our system. Similarly, we tailored the language model of the speech recognizer to children speech using the available off-line data. Following speech recognition, an NLU performs an analysis of the strings in the hypothesis list generated by the speech recognizer.

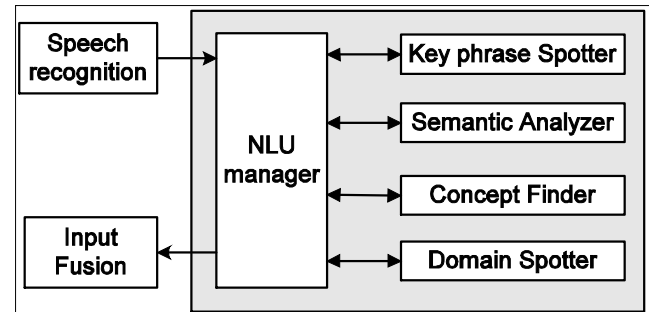


Figure 2. The Natural Language Processing module.

### 2.3 Natural Language Understanding

The NLU module (Figure 2) consists of four main components: a key phrase spotter, a semantic analyzer, a concept finder, and a domain spotter. Any user utterance from the speech recognizer is forwarded to the NLU where a key phrase spotter detects multi word expressions from a stored set of words labeled with semantic and syntactic tags. This first stage of processing usually is helpful to adjust minor errors due to misrecognized utterances by the speech recognizer. Key phrases that are domain-related are extracted, and a wider acceptance of utterances is achieved. The processed utterance is sent on to the semantic analyzer. Here, dates, age, and numerals in the user utterance are detected while both the syntactic and semantic categories for single words are retrieved from a lexicon. Relying upon these semantic and syntactic categories, grammar rules are then applied to the utterance to help in performing word sense disambiguation and to create a sequence of semantic and syntactic categories. This higher-level representation of the input is then fed into a set of finite state automata, each associated to a predefined semantic equivalent according to data used to train the automata. Anytime a sequence is able to traverse a given automaton, its associated semantic equivalent is the semantic representation corresponding to the input utterance. At the same time, the NLU calculates a representation of the user utterance in terms of dialogue acts. At this point of processing, the NLU makes use of a domain ontology that captures knowledge of one particular domain and serves as a more direct representation of the world. To ensure its reusability, the ontology describes the domain knowledge in a generic way and provides a clear understanding of the domain. One of the clear cases of ontological reuse in a conversational system for an historical character is to craft e.g. the concept of *character's life* in a generic way that comprises his birth, childhood, youth, adulthood, and family (which in turn comprises e.g. father, mother, siblings, etc.). Such a general representation of the concept *character's life* can then be extended and reused across different characters. The concept finder relates the representation of the user utterance, in terms of semantic categories, to the domain level ontological representation. Once semantic categories

are mapped onto domain level concepts and properties, the relevant domain of the user utterance is extracted. The domain helps providing a categorization of the character's knowledge set.

The final output in form of concept(s)/subconcept(s), property, dialogue act and domain is sent to the input fusion module and from there to the character module. Table 1 shows an example of how two exemplary sentences are processed as soon as they get through each of the single modules within the NLU.

## 2.4 The Character Module

The CM is in charge of managing the behavior of the animated character and maintaining natural, educational and entertaining conversation with the user. It deals separately with task modeling and dialogue modeling as customary in most of the state of the art dialogue managers [2, 5, 27].

The behavior of the animated character is controlled via a finite state network implemented into the CM. The finite state network has three states, each of which implements a different set of behaviors that reflect a simplified model of a typical dialogue. This model assumes that either (1) no conversation is going on (referred to as a *resting state*), which is also the default state at system start up, (2) a conversation is on going and the character is paying attention to the user (referred to as a *listening state*), (3) a

conversation is taking place and the character is talking to the user. This latter state is also implemented via an embedded finite state network.

The embedded network can be in any of five states: the start state at the very beginning of the conversation and after system initialization, the end state to indicate that the user has terminated the conversation (e.g. by saying goodbye or after several attempts of the system to continue the conversation with a user that does not reply), the continuation state where the CM is pausing after determining a topic for the next turn in the conversation; and a dialogue state where the CM has particular expectations for the next turn, this is the case when the agent has asked a question to the user or is engaged in a mini-dialogue, which is a short predefined in-depth dialogue about a particular topic. An additional state is entered when meta-communication input is processed, after which the system will always move on. Regardless the state, an output is always produced and sent to the response generator.

Figure 3 depicts a detailed architecture of the components within the CM. A few of its components are discussed in more details the following subsections.

**Table 1. Example of how two sentences are processed as soon as they go through the single NLU components.**

<i>Speech Recognizer Output</i>	What did your father do to earn a living
<i>Keyphrase Spotting</i>	What did your father do to <profession:general>
<i>Semantic Analyzer</i>	<question:general><family:father><profession:general>
<i>Concept Finder</i>	<dialogue_act:question><dialogue_act_type:general><concept:family><subconcept:father><subconcept:profession><subconcept:general>
<i>Speech Recognizer Output</i>	Did your tell me something about it to be the fairy tale the little mermaid
<i>Keyphrase Spotting</i>	Did your tell me something about it to be the <fairytale:little_mermaid>
<i>Semantic Analyzer</i>	<question:general><fairytale:little_mermaid>
<i>Concept Finder</i>	<dialogue_act:question><dialogue_act_type:general><concept:fairytale><subconcept:little_mermaid>

### 2.4.1 Conversational Mover and Conversational Moves

During interaction the conversational mover receives a description of the input in terms of concepts, sub-concepts, properties, dialogue acts and dialogue act types. Using a set of rules over the occurrence in the utterance of certain combinations of those descriptors, a conversational move is the output of the processing by the conversational mover.

There are two main kinds of conversational moves: generic and domain-related. Generic moves are, for instance, *user\_praise*, *opening\_greeting*, *ending\_greeting*, and those for meta-communication such as e.g. *clarify*, *repeat*, and *correct*. The names are self-explanatory. Domain-related moves are instead linked to a conversational domain.

Depending on the current conversational move, the state of conversation, and after consulting the ontology, the conversation history and, if necessary, the meta-communication handler, the conversational move processor determines the next conversational move.

### 2.4.2 Knowledge Base and Tree Structure

The knowledge base is organized in terms of conversational moves per domain. There is also an additional domain where the generic moves are stored and a few sub domains where the moves corresponding to the mini-dialogues are included.

In the knowledge base, each conversational move has defined three pieces of additional information that the conversational intention planner (CIP) uses during conversation; they are action code, continuation code and expectations. These components are used in a similar way as described in [5]. Other information associated to each conversational move in the knowledge base is the corresponding output text, which might include variations of the same output template, as well as emotional increments for some kinds of output. This information is used by the response generator for producing the output.

During initialization the CM loads the domain-related conversational moves from the knowledge base into a tree structure, one for each domain. Once a conversational move is triggered by the user or selected from the tree by the system, the

conversational move processor uses the domain agent to retrieve from the knowledge base the action codes, continuation codes and expectation values for the move. Action codes mainly relate to activation of pauses and blocking of nodes or groups of nodes in the tree. Continuations are codes that the CIP uses to calculate a conversational move for the next turn. Actions and continuations are executed, if they are not null, and if the move has expectations they are kept for the next turn. The move processor also retrieves from the knowledge base information related to the output text and its corresponding emotional increments. Once the emotional calculator processes this latter piece of information, the result is sent by the CM to the RG to produce and output.

### 2.4.3 Conversation history

The conversation history is a combination of classical slot-based conversation history and mechanisms for keeping track of what topics have been already exhausted and those the system can still talk about to the user. Similarly to [29] we use the tree structure to keep track of this information, but additionally we use a classical slot-based conversation history to keep track of the current conversation state, last and current domain and conversational move, number of consecutive turns involving meta-communication, last output sent to the response generator, mini-dialogue status, etc. Differently from [29], we do not use weights on the topics nodes to determine if they can be used or not in the future. We simply mark the nodes that have been used for the system not to use them any more during conversation, although the user might trigger them as many times as he wants. Also in our implementation, large sections of the ontology-tree can be marked as used depending on the user's input. For instance, if the user does not want to talk about the life domain after the system has suggested it twice in succession, the entire life domain is marked as used for the system.

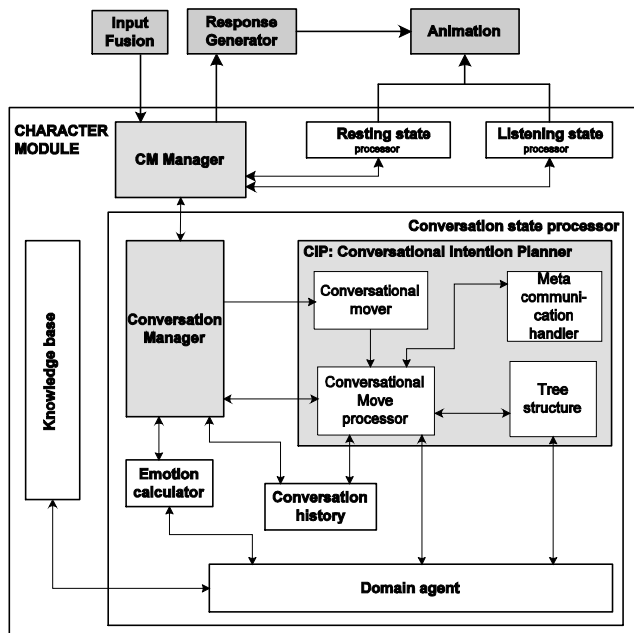


Figure 3. Detailed CM architecture.

## 2.5 Gesture for Object Manipulation

The system is capable of generating actions that influence objects in the 3D graphical world. Objects can be selected, talked about, deleted, highlighted, hidden, carried on and dropped by the agent.

Many of the objects in the study have been designed to resemble events experienced by the character and/or works that he created during his real life. For instance, a picture of a duck hanging over his desk serves as a link to his famous fairy tale 'The ugly duckling'; another one depicting the Colosseum in Rome is hung on the wall to refer to his love for and frequent travels to Italy; books are stored in shelves all over the study, while a few personal object of the writer, such as his umbrella and walkstick, stand by the door. These objects have not been placed in the room just for embellishment purposes. They had a central role in the character's real life and thus offer a topic of conversation to the user. From a technical perspective, those objects form the basis for multimodal interaction. Object behaviors are used as visual feedback in deictic utterances as well as for their selection when gesture input is employed. Anytime the user selects an object, the character turns to it and, accordingly to the accompanying speech command, if any, it performs an actions to the object signaled.

## 2.6 Visual Representation and Multimodality

The use of gesture enables people's interaction with the visual representation of objects. This visual representation reinforces the verbal information conveyed to the user and provides important cues and topics for the ongoing conversation. This also results in an interaction that is more intuitive and flexible. Moreover, late recall of information is facilitated by the visual stimuli.

The multimodal interface improves the accessibility for diverse users, contexts and information domains. Based on multimodal communicative acts, our embodied agent permits the building of a kind of relationship with the interactive environment to facilitate its exploration. An efficient integration of visualization and multimodal interaction techniques ensures good user experience and an improved learning experience what is very important particularly in an educational system such as ours.

## 2.7 Response Generation

The ability to realistic display of face, body language, emotion and speech to regulate the dialogue process and to achieve the ultimate goal of suspension of disbelief [15] on the part of the user is central to the concept of ECA. The highly believable animated characters that have been created for the movie industry demand a great deal of manual design work from professional animators based on a predefined storyboard. However, the behavior of an ECA cannot be created in advance because it must reflect and accommodate for the on-going conversation with the user. Our approach to behavior generation is based on a simplification of the idea of gesture and speech as modalities on a par like in the theoretical frameworks proposed in [1, 4].

When dialogue occurs within one of the domains covered, the RG receives a message from the CM with the necessary information to retrieve the next contribution to the ongoing conversation from a knowledge base. The knowledge base stores predefined sentences along with encoded non-verbal classes of behaviors, semantic classes and the system's domains ontology. Unification between the incoming information from the CM and the information stored in is carried out to retrieve the correct template.

Numerous, ever-expanding, canned templates guarantee broad domain coverage but also require manual maintenance and have a limited variability by design. Each template is a compact representation of a predefined spoken output with embedded tags for non-verbal behaviors and placeholders to textural values to be filled in at run-time. Both textural values and non-verbal behavioral elements are un-instantiated. The binding of non-verbal behavior to actual gesture and variables to text occurs at run-time enabling a sentence to be synthesized on the fly at different times with different accompanying non-verbal elements and/or words. Tables that map semantic categories onto non-verbal behaviors are maintained to bound non-verbal behavior tags to actual gestures. The RG operates on the selected canned template and encodes the modalities to display into a unified XML string representation. That assembled representation is forwarded to the graphical engine for rendering and to the TTS.

Currently, we store some 450 templates, many of which are non-variable stories to be told by the character, and 110 different non-verbal primitives. We designed the templates by hand after the analysis of data from recordings of an actor impersonating H.C. Andersen and interacting with kids in a children's theatre class in the fairy tale writer's hometown Odense, Denmark. Complex non-verbal behaviors can be easily created, combined, sequenced, assigned a name, and stored via a generative approach which is based on a layered composition of primitives [6, 7].

### 3. EVALUATION

We run a pilot study involving kids in the attempt to shed light on the factors that contribute in creating better computer games. How computers are able (or perceived) to play, the degree of challenges, entertainment and interaction they offer, and the believability of the game characters, seem to be among such factors. However, neither it is easy to quantify those variables nor there is much empirical evidence [30] that support the claim that those are indeed the factors that make a game worth playing. We further examined also the 'goodness' of the technical system in term of reliability and accuracy of all its single components.

Thirteen kids evenly split between males and females (6 and 7 subjects, respectively) were recruited as participants through local schools in the city of Odense, Denmark. Each user test session had a duration of 50-60 minutes. A session included conversation with the agent followed by a post-test interview. The participants' average age was 13.1 years (12.8 for males and 13.3 for females). All kids were Danish native speakers with advanced skills in speaking English. Fifty-three percent of them (100 percent of males and 43 percent of females) indicates themselves as being a frequent (i.e. more than 1 hour/week) videogame and/or console player, with a peak of 45 hours/week spent in gaming by a male teenager. A 38.5 percent of the participants (28.6 percent of females and 50 percent of males) had been exposed before to computing systems able to process speech and/or gesture; all of them were acquainted with an earlier version of our system. With regard to pre-interaction knowledge about the writer, his life, his fairy tales and the historical period he lived, 53.8 percent of the children (42.8 percent of females and 66.7 percent of males) declared to have a fair to very good knowledge of these historical

and literacy facts and events<sup>1</sup>. When asked about their favorite games, children said they like to play with games of any genre, ranging from shoot-'em-up (66.6 percent of males and 0 percent of females), action, platform, to sports and strategy games (40 percent of males and 50 percent of females).

For playing with the system, each subject had to wear a microphone headset to enter spoken utterances and could choose among a touch screen, a mouse and a keyboard for entering gesture marks. Initially, the participant was given a 15 minutes session to get accustomed with the system. During this time an assistant was present to help out in case of questions about system functioning. At the end of the introductory session, after a short break, each subject was given a set of tasks to carry out during an additional interaction session lasting for 20 minutes without human assistant support. We video- and audio-taped each session while system events were all automatically logged into XML files for further dialogue analysis. Kids were allowed to break up the game at any time for any reason. At the end of the interaction each kid was interviewed according to a set of predefined questions and then were handed out (without being told in advance) a theater ticket as reward for their time spent in the interaction. The questions were used to survey four main aspects, namely user's gaming habits, system interaction capabilities, system's educational and entertainment values, as well as open-ended questions for the subject to provide us with valuable insights and suggestions for creating a better systems.

Two persons independently evaluated the questionnaires. Qualitative quantities for the categories we wanted to analyze were mapped onto numerical values on a Likert scale from 1 to 5. For instance, when looking at the subjective entertainment degree experienced by the user, we mapped sentences such as e.g. '*I had no fun at all*' and '*the interaction was very entertaining, amazing!*' to 1 and 5 respectively. Interrater reliability for second scoring of the entire questionnaire data was 94%. Data analysis over the single categories revealed numerical value distributions of sufficient regular shape. Thus, despite the limited sample size, the obtained results can be shown in terms of statistical measures like the mean and the standard deviation. These values for a few categories, each characterized by a reasonably symmetric distribution of and no outliers among its numerical values, are:

- Interface easy of use (difficult = 1, very easy = 5): mean = 3.9, stdev = 0.28
- Graphics and quality of animations (bad = 1, great = 5): mean = 3.38, stdev = 0.75
- Agent's understanding skills (almost nothing = 1, almost anything = 5): mean = 3, stdev = 0.57
- Interface entertaining degree (not at all = 1, very exciting = 5): mean = 3.77, stdev = 0.44
- Interface educational content & degree of learning (nothing = 1, much = 5): mean = 3.08, stdev = 0.64
- System's overall rate (very bad = 1, great = 5): mean = 3.62, stdev = 0.87

---

<sup>1</sup> To justify this high figure, it may be useful to remind that Odense is Andersen's hometown. Many events take place every year to celebrate the writer and special after school programs are offered having a topic related or somehow linked to him

In other words, the system was overall rated fairly well. It was perceived as exiting and funny, with a reasonable degree of added educational value. With regard to the educational content, most of the users did not indicate what exactly they have learnt, yet when they did they mostly referred to the writer's life and family while stating that they already knew a great deal about his fairytales and therefore there was nothing new to learn about this topic. On the light of that, more specific questions on what side of the writer (such as life or family or historical period, etc.) subjects have learnt about while playing should be considered in future studies.

The interaction with the character is driven primarily by the speech modality however a small set of pen gestures is available to operate on objects in the room as well. Interestingly, 53.8 percent of the subjects (50 percent males and 57.1 percent females) stated they liked the gesture modality and/or wanted to do more with it. Despite gestures were not extensively used by the subjects, we hypothesize that they ease shy users into the conversation (shy users generally start with clicking on a picture and then just wait for something to happen; rarely they ask about it). Gestures may help breaking the initial hesitance on the part of the user and help form a relationship with the interactive character, which forms the basis of a smooth overall conversation.

From a dialogue management point of view we were interested in evaluating aspects like conversation success, domain coverage, robustness, etc. As mentioned in section 2.4.1 every input processed by the CM is converted into a conversational move that falls into one of the domain covered in the knowledge base. Based on the amount of conversational moves from real conversational data, we have determined the average number of turns over each domain as well as their percentage of domain coverage. Table 2 presents this statistics.

**Table 2. Average Domain Coverage.**

Domain Name	Average # of turns	Percentage
Fairy tales	8.2	9.6
Life	6.9	8.1
Physical Presence	4.8	5.7
Study	13.3	15.6
User	7.7	9.0
Generic	44.2	51.9
<b>Total # of turns</b>	<b>85.1</b>	<b>100</b>

The study domain has the information about the objects in the study, so every time the user points at something in the study the study domain is triggered. This can be a good indicator of the multi-modality input behavior of the users. The generic domain is the most addressed one by the user. Being the output in this domain not always dependent on the user input, it is not always the user the cause for triggering a reply from it. The generic domains covers all the meta-communication involved in the dialogue and is called upon anytime a low confidence score occurred in the speech or gesture recognizer or the NLU. This is actually an indication that sustaining dialogue is a difficult task.

We haven't performed any data correlation analysis because of the limited number of subjects and thus the lack of a large set of data. A very preliminary examination about the correlation between entertainment and favorite game genre and gameplay expertise, respectively, proved itself inconclusive.

#### 4. CONCLUSIONS AND FUTURE WORK

Kids across gender, age groups and culture, like to play video games. Games that support characters behaving like human beings seem to enhance gamers' experience [30].

We have presented an entertaining and educational game system for children to play and interact with. The system consists of an ECA impersonating the Danish fairy tale writer Hans Christian Andersen. The system is targeted to children of both genders of age 10 to 18 years. It is easy to use and appeals to most of the subjects we had in a user test regardless of their gender. This may also indicate that girls' attitude to play computer games has changed. What once was seen as a 'boy territory' is now not anymore so, even if the amount of hours spent playing computer games by our participants is still higher for boys than for girls.

A user study carried out with a group of thirteen kids showed a high degree of satisfaction in the system appearance and interface. We recognize that, due to the small number of subjects, the conclusions that can be drawn from our user study are of limited statistical relevance. We are also aware that the set of subjects was homogeneous (only Danish children) while our system is targeted to kids of any nationality able to speak English. Thus, we need a larger and more heterogeneous group of participants to validate our preliminary results. Nevertheless kids' responses to our post session questions seem to indicate that our system may contribute toward entertain kids and teenagers while providing them with useful educational content. Several youngsters reported that it was fun to have the computer tell about the writer even if they already knew what they were listening to. It was an entertaining manner to brush up their knowledge, thus making the system a complement to school lessons rather than a replacement for them. We will need a larger and more heterogeneous group of participants to validate these preliminary results.

Several studies of children exposed to video games, especially in the case of violent ones [9, 11, 26], have shown that repeated or prolonged immersion to computer games results in cycles of reinforcement of the experience that are potentially origin of powerful learning experience as kids learn certain behaviors and social rules from observation. Furthermore, those rules are then reinforced from first person practice and tend to emulate them. Becoming absorbed in performing a specific activity is a common reported experience in playing video games. This phenomenon is correlated with a frame of mind of accentuated attention and application referred to as flow state [8], which has been associated with enhanced learning [21]. None of the participants to our pilot study showed indicators of absorption very likely due to the lack of both a definite goal as big motivator and the emotional, self-gratification implications deriving from competitive games such as those played against other humans or the computer. Yet, most of the kids reported that they did learn about the agent's life, family, fairy tales and/or his historical period, showing that interaction with our prototype makes learning an active experience in a relaxed and receptive mental state for learning.

From the subjects' wish lists, we can notice a demand for cunning, challenging and adapting environments. Participants want more fun, interaction, action and competition. They wish for more clickable objects and entities, mobility and actions.

As far as it concerns the system, we plan to improve integration of spoken interpretation and gesture interpretations, which at this stage, was implemented in a very simply way being pen gestures used only to select on-screen objects.

We have to add an objective to our interactive system to turn it into an actual computer game. We plan to sprinkle the game with several different sub-tasks and riddles of variable difficulty for the user to solve to proceed through the game. If a game is not difficult enough, experienced players lose interest. On the other hand, if it is too difficult, the initial enjoyment may give way to frustration. Several psychological factors hang in a delicate balance: challenge, enjoyment, frustration, and curiosity and has to be taken into account. As Norman [23] puts it '*Once the curiosity is lost and the frustration level becomes too high, it is hard to get a person's interest to return to the game.*' We wish to have many things continually happening to maintain the curiosity motive. In general, we want the game to maintain its appeal for players of any levels, from beginners to experienced players.

Speech recognition errors remain a big problem. For instance, the simple sentence 'bye' used to conclude the interaction was misrecognized several times as 'why'. In such cases it would have been worthwhile to have sensors monitoring the presence of the user. Also long, uninterruptible character's replies as well as those like '*I tried, I give up*' in response e.g. to speech recognizer errors, should be used very sparingly, if at all. Such sentences create a feeling of disappointment in the user and inevitably result in the user becoming more passive to the system.

By using a local knowledge base for a small set of domains, a user quickly uses up topics while the replies from the characters become more and more familiar. Inevitably, such a system does not make a good game for a user would stop playing with it just after a few interactive sessions. To prevent this problem, without having to keep on adding data to the knowledge base, we have developed a module that integrates a set of question answering systems to retrieve information available on the web. This module is triggered whenever the user addresses certain topics not stored in the knowledge base [7] but currently it is not capable of providing contextual replies. We plan to investigate how we can properly integrate our current domain related component with the web-based component to add more depth to the system and make the interaction more entertaining.

Camera control during conversation is partially automatic and partially controlled by the user via keyboard. We need to investigate how much control over the camera we can grant the user without disrupting the interaction flow.

Currently the state of the system cannot be saved for the interaction to be resumed at a later time. Because game play generally takes place over extended periods, it is necessary to make it possible for the user to return to the system at any time without disruption. The users must be able to resume the game at a later time and find the game configuration exactly as it was left.

Another issue that needs attention and careful design is how to add constraints in the dialogue so that false expectations on

system capabilities are not elicited. Discretely guiding and supporting the user in the conversation may help in limiting out of domain dialogue.

## 5. ACKNOWLEDGMENTS

Our thanks go to the Human Languages Technologies programme EU IST-2001-35293 for supporting this research work and to Dymtro Kupkin, Holmer Hemsén and Mykola Kolodnytsky for programming help and Svend Killerich for data entry.

## 6. REFERENCES

- [1] Alibali, M.W., Kita, S., Young, A.J. Gesture and the process of speech production: We think therefore we gesture. *Language and Cognitive processes*, 15(6):593-613, 2000.
- [2] Bohus D., and Rudnicky, A. I. RavenClaw: Dialogue Management Using Hierarchical Task Decomposition and an Expectation Agenda. *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, 597-600.
- [3] Cassell, J., and Stone, M. Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. *Proceedings of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, 1999, 34-42.
- [4] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (eds) *Embodied Conversational Agents*. MIT Press, 2000.
- [5] Charfuelan, M., and Bernsen, N.O. A Task and Dialogue Model Independent Dialogue Manager. *Proceedings of RANLP 4th International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 2003, 91-97.
- [6] Corradini, A., Mehta, M., and Bernsen, N.O. Script-based Multimodal Output Generation for a Conversational Character. *Proceedings of the International Conference on Human Computer Interaction*, Las Vegas, NV, vol. 3, CD ROM, USA, 2005.
- [7] Corradini, A., Mehta, M., Bernsen, N.O., and Charfuelan, M. Animating an Interactive Conversational Character for an Educational Game System. *Proceedings of the ACM International Conference on Intelligent User Interfaces*, San Diego, CA, 2005, 183-190.
- [8] Csikszentmihalyi, M. *Creativity: Flow and the psychology of discovery and invention*. Harper Collins Publishers, 1996.
- [9] Donnerstein, E., and Smith, S.L. *Impact of media violence on children, adolescents, and adults*. In Kirschner, S., and Kirschner, D.A. (eds.): *Perspectives on psychology and the media*, 1997, 29-69.
- [10] Drennan, P. *Conversational Agents: Creating Natural Dialogue between Players and Non Player Characters*, *AI Game Programming Wisdom 2*, Charles River Media, 2004.
- [11] Funk, J.B., Pasold, T., and Baumgartner, J. How Children Experience Playing Video Games. *Proc. of the 2nd Int'l Conf. on Entertainment computing*, Pittsburgh, PA, 2003.
- [12] Isbell, C.L., Kearns, M., Kormann, D., Singh, S., and Stone, P. Cobot in LambdaMOO: A Social Statistics Agent, *Proceedings of the 17th National Conference on Artificial*

- Intelligence (AAAI)*, AAAI Press, Menlo Park, CA, 2000, 36–41.
- [13] Isobe, T., Hayakawa, S., Murao, H., Takeda, K. and Itakura, F. A Study on Domain Recognition of Spoken Dialogue Systems. *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, 1889-1892.
- [14] Johnson, W.L., Rickel, J.W., and Lester, J.C. Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of Artificial Intelligence in Education*, 11, 47-78, 2000.
- [15] Johnston, O., and Thomas, F. *The Illusion of Life*. Walt Disney Production, 1981.
- [16] Khoo, A., and Zubek, R. Applying Inexpensive AI Techniques to Computer Games. *IEEE Intelligent Systems*, 17 (4): 48-53, 2002.
- [17] Kuck, D.J. Can We Build HAL? Supercomputer Design. In *HAL's Legacy: 2001's Computer as Dream and Reality*, edited by Stork, D.G., MIT Press, 1996.
- [18] Lee, S., Potamianos, A., and Narayanan, S. Acoustics of children's speech: Developmental changes of temporal and spectral parameters, *Journal of Acoustical Society of America*, vol. 105, 1455-1468, 2003.
- [19] Marinelli, D., and Stevens, S. Synthetic Interviews: the Art of Creating a 'Dyad' Between Humans and Machine-based Characters, *Proceedings of the 6th ACM International Conference on Multimedia: Technologies for interactive movies*, Bristol, UK, 1998, 11-16.
- [20] Mauldin, M. Chatterbots, TinyMUDs, and the Turing Test: Entering the Loebner Prize Competition, *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, AAAI Press, Menlo Park, CA, 1994, 16–21.
- [21] Moneta, G.B., and Csikszentmihalyi, M. The effect of perceived challenges and skills on the quality of subjective experience. *Journal of Personality*, 64, 1996, 275-310.
- [22] Morris, T.W., Conversational Agents for Game-Like Virtual Environments. *Papers from the AAAI 2002 Spring Symposium on Artificial Intelligence and Interactive Entertainment, TR SS-02-01*, AAAI Press, 2002, 82-86.
- [23] Norman, D.A. *The Design of Everyday Things*. The MIT Press, London, England, 2001.
- [24] Pan, S. A Multi-Layer Conversation Management Approach for Information Seeking Applications. *Proceedings of the International Conference on Spoken Language Processing*, Korea, 2004, 245-248.
- [25] Pazzani, M., and Billsus, D. Adaptive Web Site Agents. *Proceedings of the 3rd International Conference on Autonomous Agents*, 1999.
- [26] Provenzo, E.F. *Video kids: Making sense of Nintendo*. Cambridge, MA, Harvard University Press, 1991.
- [27] Robinson, K., Horowitz, D., Bobadilla, E., Lascelles, M., and Suarez, A. Conversational dialogue Management in the FASiL project. *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, Boston, USA, 2002, 113-124.
- [28] Rudnicky, A.I., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu W., and Oh, A. Creating natural dialogs in the Carnegie Mellon Communicator system. *Proceedings of Eurospeech*, 1999, 4, 1531-1534.
- [29] Stede, M., and Schlangen, D. Information Seeking Chat: Dialogue Management by Topic Structure. *Proceedings of CATALOG'04 8th Workshop on the Semantics and Pragmatics of Dialogue*. Barcelona, Spain, 2004, 117-124.
- [30] Sweetser, P., Johnson, D., Sweetser, J., and Janet, W. Creating Engaging Artificial Characters for Games. *Proceedings of the 2nd International Conference on Entertainment computing*, Pittsburgh, PA, 2003.
- [31] Weizenbaum, J., ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine, *Communications of the ACM*, Volume 9, Number 1 (January 1966): 36-45, 1966.
- [32] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T.J., and Hetherington, L. Jupiter: A Telephone-based Conversational Interface for Weather Information. *IEEE Transactions on Speech and Audio Processing*, 8(1):100-112, 2000.