Evaluation of Spoken Multimodal Conversation

Niels Ole Bernsen and Laila Dybkjær
Natural Interactive Systems Laboratory
University of Southern Denmark
Campusvej 55, 5230 Odense M, Denmark
+45 65 50 35 51
nob@nis.sdu.dk, laila@nis.sdu.dk

ABSTRACT
Spoken multimodal dialogue systems in which users address face-only or embodied interface agents have been gaining ground in research for some time. Although most systems are still strictly task-oriented, the field is now moving towards domain-oriented systems and real conversational systems which are no longer defined in terms of the task(s) they support. This paper describes the first running prototype of such a system which enables spoken and gesture interaction with life-like fairytale author Hans Christian Andersen about his fairytales, life, study, etc., focusing on multimodal conversation. We then present recent user test evaluation results on multimodal conversation.

Categories and Subject Descriptors
H5.1 [**Information Systems**]: Multimedia Information Systems – *animations; artificial, augmented, and virtual realities; audio input/output; evaluation/methodology.* H5.2 [**Information Systems**]: User Interfaces – *evaluation/methodology; graphical user interfaces; natural language; voice I/O.*

General Terms
Design, Experimentation, Human Factors, Management, Measurement, Performance.

Keywords
Natural interaction, spoken conversation, evaluation.

1. INTRODUCTION
Research in spoken dialogue systems (SDSs) has been ongoing for many years – first in unimodal (speech-only) SDSs and later also in multimodal SDSs. However, nearly all SDSs so far have been task-oriented [6]. It is only quite recently that researchers have begun to look into non-task-oriented SDSs. In the late 1990s, a Swedish group created the August system [14] which afforded spoken interaction with the expressive face of Swedish author August Strindberg about topics, such as restaurants in Stockholm and the Royal Technical University. Even if the system had limited conversational abilities and was rather close to being task-oriented, the experiment was an interesting one which pointed beyond task-orientation and towards entertainment systems.

Tutoring systems for education and training have been researched for some time, especially in the USA. Some of the most advanced systems include talking faces or embodied characters, spoken dialogue, and sometimes also computer vision. Work on integrating these technologies and the modalities they enable is described in [8]. A major ongoing effort pushing the boundary of task-orientation is the US Army-sponsored systems for tactical situation control and tactical Arabic training [16].

In Europe, the NICE (Natural Interactive Communication for Edutainment) project [18] is developing a non-task-oriented multimodal edutainment SDS enabling users to have spoken and input gesture conversation with life-like fairytale author Hans Christian Andersen.

The issues addressed both in the projects referred to and in similar efforts illustrate the many development challenges that lie ahead before we can build systems which are able to communicate with their human interlocutors in the same ways as a humans communicate with one another. As regards evaluation, we are beginning to realise that many issues remain to be solved as well. Since there are only few non-task-oriented systems around and since these are still under development, little experience exists concerning the criteria to use for their evaluation.

This paper describes the NICE Hans Christian Andersen system in Section 2 and presents the underlying theory of multimodal conversation in Section 3. We then discuss how to evaluate multimodal conversation in Section 4. Section 5 presents evaluation results from the user test of the first HCA prototype. Section 6 discusses the results and describes next steps in our work.

2. THE NICE HCA SYSTEM
The Hans Christian Andersen (HCA) system is one of two systems under development in the NICE project (2002-2005). The second system aims to enable spoken interaction with some of HCA's fairytale characters in a virtual fairytale world [13]. The systems use the same graphics rendering engine and their second prototypes will be linked so that the user can move from HCA's study into the fairytale world. The HCA system discussed in this paper uses English spoken conversation whereas the fairytale world system is in Swedish.

2.1 HCA in his Study
The primary use setting of the HCA system is in museums and other public locations at which interactions with an international user audience are expected to have an average duration of, say, 5-15 minutes. Target users are 10-18 years old children and teenagers. The system may be partly viewed as a new kind of compu-

ter game which integrates spoken conversation into a professional computer games environment and aims to entertain through emulated human-human conversation. However, the system also has an educational purpose which is being pursued by providing ample correct factual information through story-telling and otherwise, and both visually and orally. Figures 1 and 2 show 55-year old HCA surrounded by artefacts in his study. The study is a rendering of HCA's study on display in Copenhagen, modified so that he can walk around freely and so that a pair of doors lead into the fairytale world (Figure 2). Pictures relating to HCA's knowledge domains have been hung on the walls. The user may point to them and ask questions about them; HCA can tell stories about the pictures and also about some other objects in his room, such as his travel bag, HCA being an active traveller (Figure 2). Lacking locomotion autonomy in the first prototype, HCA's locomotion is controlled by the user who is also in control of four different virtual camera angles onto his study.

Figure 1. HCA gesturing in his study.

Figure 2. HCA's study. View on desk, double doors leading to the fairytale world, and travel accessories.

2.2 System Architecture

The NICE project has five partners: LIMSI-CNRS (Paris, France), Liquid Media (Stockholm, Sweden), NISLab (Odense, Denmark), Scansoft (Aachen, Germany), and TeliaSonera (Stockholm, Sweden). Scansoft delivers speech recognition optimised for children and adolescents for Swedish and English, trained on data (mostly) collected in the project. LIMSI is responsible for gesture recognition (based on freeware), gesture interpretation, and speech-gesture input fusion. Liquid Media is responsible for animation and system integration. NISLab develops natural language understanding, character modelling [4], and response generation [9] for HCA. TeliaSonera develops natural language understanding, character modelling, and response generation for some of HCA's fairytale characters. Speech synthesis is off-the-shelf software.

The general architecture of the first prototype HCA system (PT1) is shown in Figure 3. The modules communicate via a central Message Broker which is publicly available from KTH [17]. The Broker is a server that routes function calls, results, and error codes between modules, using the Transmission Control Protocol (TCP) for communication. The Broker coordinates input and output events by time-stamping all messages from the modules and associating them to a certain dialogue turn. The behaviour of the Broker is controlled by a set of message-passing rules, specifying how to react when receiving a message of a certain type from one of the modules.

Speech recognition is simulated in HCA PT1 because the speech recogniser still needs to be trained on 40-50 hours of speech data recorded with mostly non-native English speaking children. Pre-PT1 Wizard of Oz speech data collected at the HCA Museum in his native city, Odense [5], as well as data collected with PT1 is being used for training purposes. Also, HCA PT1 has no semantic input fusion. The input fusion module simply passes on to the character module 1-best input semantics, topic, and domain information from natural language understanding and n-best gesture interpretations from the gesture interpreter. The character module manages the conversation, eventually passing on to response generation sets of spoken output template references, input values, non-verbal behaviour references, and HCA's updated emotional state. Response generation retrieves synchronisation timing information from the speech synthesiser whereupon the animation module takes care of speech/non-verbal synchronisation. For a comprehensive system description, see [2].

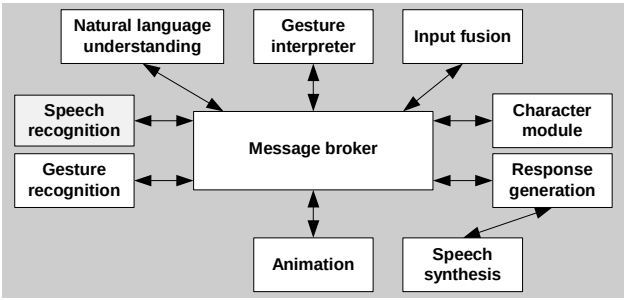

Figure 3. General NICE HCA system architecture.

2.3 Domain-oriented Conversation

The ultimate goal of developing non-task-oriented spoken and multimodal conversational systems is to be able to create systems which, like humans, can have conversation about virtually everything [15]. While such systems are still a long way off, a more approachable "half-way" goal is to build *domain-oriented* SDSs, i.e., SDSs which can conduct an approximation to human conversation but only *within* one or several *particular* domains. We use the term 'conversation' to refer to the spoken, and possibly multi-

modal, interaction which takes place when there is no common task to be solved by the interlocutors. Presumably, the absence of a common task implies the absence of a single, shared user-system goal in any operational sense. Jointly, these negated constraints imply completely new design challenges, cf. below. As defined by its domains and conversational principles, the HCA system may be regarded as a sub-type of a domain-oriented conversation system, cf. below.

In the HCA system, the user can address, in any order, any part of any domain or topic within HCA's knowledge domains, using spontaneous speech, mixed-initiative dialogue, and pointing gesture. In PT1, the domains are: HCA's fairytales, his childhood life in Odense, his physical presence and his study, the user, and HCA's role as "gate-keeper" for access to the fairytale world. In addition, HCA has a 'meta' domain in order to be able to handle repair meta-communication during conversation. In conversation about HCA's study, the user may use 2D gesture input to indicate via touch screen or mouse an object which HCA might want to tell a story about. HCA may react emotionally to the user's input by, e.g., getting angry or sad because of what the user says, or by getting happy if the user likes to talk about his fairytales. The underlying emotion state-space model is updated for each user input which modifies HCA's emotional state from the default friendly state towards happiness, anger, or sadness. The resulting emotional state is used by response generation. However, although the mechanisms for emotion shifts have been implemented in PT1, there is only little verbal and non-verbal rendering of emotions. This is partly due to the limitations of the PT1 rendering engine which mean that HCA can only display a single graphical non-verbal behaviour at a time, and partly due to the lack of emotional variants for most of HCA's verbal output. One of the few examples of emotional output from the user test is shown in the dialogue fragment in Figure 4 (cf. Section 5). With some justification, the user described HCA as being impolite in the post-test interview.

User 2a: picturejennylind [no accompanying speech]

HCA: This is a picture showing the Swedish nightingale Jenny Lind. She sang beautifully and is one of the most beautiful women I have ever met. I fell very much in love with her. Unfortunately, she didn't fall in love with me but she became a dear friend of mine, a dear friend. You should try to point at the things in my study. I like to tell my guests about some of them.

User 2a: stop it

HCA: Don't piss me off!

Figure 4. Dialogue snippet from user test. "picturejennylind" means that the user points to a picture of Jenny Lind.

The HCA system's character module is always in one of three output states, called NCA, CF, and CA, respectively. When in the NCA (non-communicative action) state, HCA goes about his work in his study and is not engaged in conversation. He transits into the CF (communicative function) state whenever he is being addressed by the user, paying attention to the user's conversational contribution. In the CA (communicative action) state, he responds to the user's input [1].

In PT1, HCA has fairly limited knowledge about his domains of conversation. His output behaviours are being composed on-line from approx. 300 response templates and 100 primitive non-verbal behaviours. We have implemented his domains breadth-first in order to maximally explore the issues involved, rather than first implementing a single domain in depth. The cover story is that HCA is back! However, he still has to re-learn much of what he once knew. If the user would do him the favour of visiting him later, he is convinced that he will have become much more of what he once was. In addition to the very true information provided by the cover story, the story may help convince users that HCA is not (yet) a full virtual person and make them behave accordingly. HCA does not tell the cover story up front to new users and does not, more generally speaking, instruct users on how to interact with him or inform them of what he is able to have conversation about. Rather, users will be told his cover story if they either explicitly ask what HCA knows about or can do, or if they show too much interest in things he does not know about (yet).

3. THEORY OF CONVERSATION

As argued above, conversation, domain-oriented or otherwise, is essentially different from task-oriented dialogue. By comparison with task-oriented dialogue with its shared-goal principles of cooperativity, serving to maximise efficiency and successful task completion, spoken conversation has a different, and often contrary or even contradictory, richness. For instance, the Gricean maxims of shared-goal cooperativity cannot be assumed any more [3][12], task completion is irrelevant, and the serendipity of leisurely conversation would seem to render efficiency irrelevant as well. Moreover, in the case of HCA, we are addressing conversation for edutainment between young users and a famous historical character who might be somewhat familiar to many of them in advance. Clearly, we need a new model of spoken dialogue on which to base the design of HCA's conversational behaviour.

Based on the problem at hand and the literature on social conversation, we defined the notion of *prototypical successful human-human conversation* to designate our model for HCA's conversational behaviour. The term 'prototypical' means that we aim to capture, at least, part of the essence of having "a good conversation" in which two people get to know each other. The model makes use of the seminal notion of common ground [7] and has significant commonality with models of how to make friends through conversation [11]. To these properties, we added a normative symmetry requirement of our own and highlighted the rhapsodic character of conversation. The latter is based on analysis of the Wizard of Oz field simulation data collected at the HCA Museum. The data shows that, in conversation with simulated HCA acting in accordance with the PT1 specification, users tend to change domain and topic quite frequently [5]. Thus, HCA has been designed to follow these principles of prototypical successful human-human conversation:

1. initially, in a polite and friendly way, the interlocutors search for common ground, such as basic personal information, shared interests, shared knowledge, and similarity of character and personality, to be pursued in the conversation;
2. the conversation is successful to the extent that the interlocutors find enough common ground to want to continue the conversation;
3. the interlocutors should be able to provide, by and large, symmetrical contributions to the conversation by, e.g., taking turns in acting as experts in domains of common interest, so that one partner (the user) does not have to end up in the role of passive hearer/spectator, like, for instance, the novice who

is being educated or trained by the other(s). Conversely, we are not developing a user-driven Q&A machine;

4. to a significant extent, the conversation is characterised by the participants taking turns in telling stories, such as descriptions of personal experiences and items within their domains of expertise, observations, anecdotes, jokes, etc.;
5. conversation is rhapsodic, i.e. highly tolerant to digression, the introduction of new topics before the current topic has been exhausted, etc. Yet conversation also requires a reasonable amount of conversational control and coherence in order not to fall apart into disjoined semi-monologues;
6. conversation, when successful, leaves the partners with a sense that it has been worthwhile.

HCA pursues the six principles (Ps) above as follows. The principles do not mention entertainment or education, despite the system's edutainment goal. We simply assume that successful conversation will be perceived as entertaining by the target users (P6) and that these will concur that they have learned from HCA's stories and physical appearance in his study (P4). HCA assumes, of course, that the user takes an interest in his life and fairytales as well as in himself and his study (P1,P2). However, he is aware that common ground is a dual-aspect notion which has both an HCA aspect and a user aspect. As for the user aspect, HCA asks polite questions about the user early on, such as about the user's name, age, gender, and nationality. He also tries to elicit user opinions on his fairytales, his visible persona, and his study. These HCA initiatives serve the goal of conversational symmetry as well (P3), as does the following. HCA makes an effort to make the user the expert in conversation by asking about games played by children and adolescents today, demonstrating interest in football, computers, and otherwise. In our Wizard of Oz (WoZ) field collection of 30 hours of approx. 500 conversations with mostly young users at the HCA Museum, we found that the users did show keen interest in telling HCA about contemporary game-playing. They were equally happy telling him about technical inventions made after HCA's times in response to his interest in photography, trains, and other contemporary inventions. HCA, in his turn, does not just answer questions, or ask them, but tells stories and anecdotes and offers personal views about his life, his fairytales, pictures and other objects in his room, etc. (P3,P4).

Developing for domain-oriented conversation is an exploratory exercise about which, prior to the user test, we primarily knew that our first solution would be far from perfect. At development time, HCA's main problem in conducting human-style conversation seemed to us to be that he cannot always pursue in depth a topic launched by himself or his interlocutor because, at this stage, at least, (i) his knowledge and conversational skills are quite limited, and (ii) we do not have sufficient information about the key interest zones of his target audience. This is where the rhapsodic nature of conversation (P5) might come to his rescue to some extent. When, during conversation, and despite his conversational agenda (Section 4), HCA gets lost and does not understand what the user is saying, he changes topic or even domain in order to try to recover some amount of conversational control. It is in this situation that he also cracks (non-situated) jokes. However, we did not know to which extent HCA's rhapsodic behaviour may catapult him into conversational incoherence and semi-monologue. Judging from the WoZ corpus referred to above, the target users are pretty tolerant to digression. However, users who insist on pursuing a topic beyond HCA's current knowledge got frustrated and tended to thwart HCA's attempt at re-gaining control.

In the following sections, we discuss how to evaluate conversational systems, such as the HCA system, and present and discuss results from the user test of HCA PT1, focusing on the conversational interface.

4. EVALUATION OF THE HCA SYSTEM

System evaluation includes technical and usability evaluation. As regards technical evaluation, testing a domain-oriented SDS is no different from testing a task-oriented SDS. For instance, both types of system must be robust, i.e. they should rarely crash, and if there is output in several modalities which must be time-aligned, the output must be properly synchronised. Today, we have a reasonable baseline for evaluating the usability of unimodal (speech-only) task-oriented SDSs but there are still many unknowns involved in the evaluation of multimodal, task-oriented SDSs [10]. For domain-oriented conversation systems for edutainment or otherwise, however, there is hardly any dedicated usability evaluation best practice advice available. As argued above, task performance efficiency and effectiveness are hardly relevant to having a good time together in conversation. Moreover, edutainment is as irreducible to education as is a museum visit to class-room occupational training, hence comparative or absolute learning metrics are probably irrelevant in this case. Finally, even if one possible measure of entertainment success could be the sheer length and duration of the target users' conversations with HCA, the user test protocol did not allow its application, cf. below.

Even if we, like all developers, have a preference for quantitative metrics, the subjective information provided by user interviews are always an important source of information about the success of an SDS. In the present case, the user interviews following the test of the first HCA prototype, were crucial because of the lack of established quantitative metrics.

5. THE USER TEST

The first HCA system prototype was tested in January 2004 with 18 users (nine girls and nine boys) from the target user group of 10-18 years old kids and teenagers. In the user test setup, the recogniser was replaced by a wizard who typed what the user said. The rest of the system was running. Users arrived in parallel, so there were two test rooms, two wizards, and two interviewers. In one room, the user had a mouse and a touch screen for gesture input while in the other room only a mouse was available as pointing device. In the room with the touch screen, the user could also watch HCA on a 42" flat-panel screen. An observer was present in this room as well, cf. Figure 5.

Each user test session had a duration of 60-75 minutes. A session included conversation with HCA in two different conditions followed by an interview. In the first, 15-minutes condition, the users only received basic instructions on how to operate the system, i.e. to speak using the headset, control HCA's movements, control the virtual camera angles, and gesture using mouse or touch screen. In the second condition, the user received a set of 13 brief scenarios, such as "Find out if HCA has a preferred fairytale and which it is", "Make HCA tell you about two pictures and two other objects in his study", and "Tell HCA about games you like or know". The users fully decided on the order and number of scenarios to solve. They tried to solve them all, of course.

Figure 5. User interacting with HCA, cameras, observer.

One reason for the two-condition test design just described is the partner symmetry requirement described in Section 3. In the Wizard of Oz simulation mentioned in Section 2.2, we found (i) large individual differences in the extent to which a user from the target group would contribute to drive the conversation forward, and (ii) that, on average, the target users tended to be significantly less active than HCA himself [5]. As the Wizard of Oz test condition was very similar to the first user test condition described above, we had reason to believe that the first user test condition would show a similar pattern of user drive in the conversation. This was confirmed, cf. Section 5.2. We therefore designed the second user test condition in a way which would be likely to make the users drive the conversation forward much more strongly than in the first condition. For this purpose, the users were given written scenarios. We expected them to put HCA under strong user-initiative pressure to provide the information and exhibit the behaviours requested in the scenarios. This assumption was confirmed as well, cf. Section 5.2, enabling us to collect data on how well HCA's conversation would stand up to a closer approximation to symmetrical conversation.

All interactions were logged, audio recorded, and video recorded. In the room with the touch-screen, a video camera pointed at the user and a second camera recorded the screen. In the second room, a single camera recorded the user. In total, approximately 11 hours of interaction were recorded on audio, video, and logfile, respectively. In addition, 18 sets of structured interview notes were collected. In the next sub-section, we present the interview results with particular focus on evaluating the conversation and the theory of conversation underlying the system, cf. Section 3.

5.1 Qualitative, Interview-based Evaluation of Conversation

The structured post-session interviews took between 15 and 30 minutes per user. Each user was invited to simply report what came to mind when asked each of the 20 questions listed below. Structurally, questions (1) through (6) collect user information, questions (7) through (13) collect information on how the users experienced the interaction, questions (14) through (19) elicit information on the system's perceived usefulness and how it could be improved, and the final open question (20) invites any comments which were not elicited so far.

1. User identity: Name, age, gender.
2. Occupancy.
3. How often do you play computer games: hours per week?
4. (If relevant) Which computer games do you like (types of game or concrete games)?
5. Did you ever talk to a computer before? If yes, which program did you use?
6. How well do you know HCA?
7. Was it easy or difficult to use the system? Why?
8. What do you think of HCA?
9. Could you understand what he said?
10. How did it feel to talk to HCA?
11. Could he follow what you wanted to talk to him about?
12. What do you think of his behaviour on the screen?
13. How did it feel to be able to use input gesture? (a) Did you use the mouse or point onto the screen? (b) How was it to do the gestures? (c) Would you like to be able to do more with gesture? If yes, what?
14. Was it fun to talk to HCA? If yes, what was fun? If no, can you imagine what could make it fun?
15. What did you learn from talking to with HCA?
16. What was bad about your interaction with HCA?
17. What was good about your interaction with HCA?
18. What do you think we should make better?
19. How interested would you be in playing computer games with speech and gesture?
20. Any other comments?

The answers to the user information questions (1) through (6) told us that all users were school children and native Danes except for a single Scotsman. Their average age was 14.3 years with a slightly higher average for the girls than for the boys, i.e. 14.8 years vs. 13.8 years. On average, the boys spent more time per week playing computer games than did the girls, i.e. 11.1 hours vs. 3.9 hours. Only three users had spoken to a computer before. The large majority of users said that they were reasonably familiar with HCA prior to the user test.

We now look at the information provided through the interviews on HCA and his conversation skills. Qn refers to question number n in the list above. Plain numbers in parentheses show how many users shared a particular comment.

The bulk (6) of the critical comments on how easy the system was to use (Q7) concerns HCA's occasional difficulties in understanding what the users said, as evidenced by, e.g., irrelevant output and unnecessary repetitions, cf. Figures 6 and 7. HCA himself (Q8) was generally received quite positively, being realistic, lifelike, imaginative, and fun to watch (15). HCA's spoken intelligibility (Q9) was received surprisingly positively by his mostly non-native English-speaking interlocutors. The main criticism (6) was that the RealSpeak synthesiser sometimes "swallowed" or did not properly segment some syllables. Fifteen users had not spoken to a computer before (Q10). They found the experience strange, surprising (10), fun (6), or easy, like talking to a person (3). Three users found it embarrassing to talk to HCA while being observed.

User 5a: what is your name

HCA: My name is Andersen. My friends call me Hans Christian. I hope we will become friends.

User 5a: how old are you

HCA: I am fifty-five years old. And I am in a good shape occasionally I ride on horses and go for long walks and when I travel I even sometimes hike.

User 5a: where do you live

HCA: What about you don't you think that I am a good-looking man?

User 5a: yes

HCA: My headmaster at the grammar school in Elsinore called me Shakespeare with the vampire eyes.

User 5a: why

HCA: Here in my study I have everything I need. What if it was your study would you be missing anything?

Figure 6. Conversation illustrating HCA's strategy when he does not understand the user's question.

One of the key interview questions (Q11) was if HCA could follow what the user wanted to talk to him about. One user was largely happy with HCA's conversational abilities and two users were rather dissatisfied. One of these was a 12-years old girl who only managed to input two phrases to the system. All other users managed the conversation pretty well, grammatical errors and all. The main criticisms were that HCA's output was sometimes irrelevant (15), cf. Figure 6, or unnecessarily repetitive (3), cf. Figure 7. Analysis of the transcribed conversations shows that these problems were aggravated in the second test condition in which the users put HCA under heavy-handed direct interrogation in order to quickly get through the scenarios. The scenarios had been designed to make this simple strategy fail. Two users observed that HCA stuck too much to some of his pet topics. Two users noted that he could understand one input formulation but not another, equivalent one.

Overall, the users very much liked the graphics (Q12) in general and HCA's non-verbal behaviours in particular (13). Many users spotted the bugs in the graphics which enabled HCA to walk into the middle of his furniture, through a wall, or disappear through the ceiling into open space. Five users commented on the bugs. Several users made very particular observations on HCA's non-verbal behaviours, demonstrating that they took a keen interest in the output graphics, including observations on funny movements, strange walk, mysterious eye movements, and a wish for HCA to be able to sit down. Only two users remarked that HCA should be made more lively, i.e. that he should exhibit more frequent non-verbal behaviours.

No user was really negative towards the use of gesture (Q13). Even the girl who understood close to nothing appreciated the gesture opportunity. However, one user did not get the opportunity to use gesture input because the experimenter forgot to make the mouse available. A second user does not seem to have used gesture at all. Although the users seemed to appreciate 2D gesture input into HCA's study, and several of them would have liked HCA to be able to tell stories about more objects in the room, there were no requests for increasing the opportunity to use 2D gesture. In fact, many users did not really have anything to say to the matter. It is tempting to conclude that the users much preferred spoken interaction over gesture interaction with HCA.

HCA: What do you think of the prince the character from the fairy tale The Little Mermaid?

User 6b: ehm I have not read that fairy tale so I do not know

HCA: I heard they made a statue of the little mermaid in Copenhagen. Have you seen it?

User 6b: eh yes I have it was pretty eh beautiful

HCA: I have met lots of princes. Have you ever met a real prince?

User 6b: ehm no I have not

HCA: I hear that the present crown prince Frederik has now found his princess. What do you think about it?

User 6b: it is pretty good then we can get a queen and a king sometimes sometime

HCA: Can you tell me about the fairy tale The Little Mermaid?

User 6b: ehm no I have not read it

HCA: If you know it what do you think of the character the little mermaid?

User 6b: ehm I do not know much because I have not read the fairy tale so

HCA: What do you think of my fairy tale The Little Mermaid?

User 6b: ehm I have not read it so I do not know the history so

HCA: Can you tell me about the prince the character from the fairy tale The Little Mermaid?

User 6b: eh no because I have not read the story the fairy tale ehm can you tell me about the ugly duckling

HCA: Could you give me a summary of one of my fairy tales?

User 6b: ehm yes eh the ugly duckling eh lives with a duck family and gets grown into a swan when its gets old

HCA: As a child I heard for the first time of a mermaid when I saw a play at the theatre in Odense.

Figure 7. Conversation illustrating how HCA was sometimes found to repeat himself and stick too much to his pet topics.

The question if it was fun to talk to HCA (Q14) addresses the system's entertainment qualities. The bulk of the comments were that it was entertaining, fun, exciting, or great to talk to HCA (7), fine that he told long stories (4), and fun to get stories about objects by pointing to them (3). The answers on the system's educational import (Q15) included several surprises. A minor surprise, perhaps, was that most Danish users considered HCA's fairytale knowledge as reminders rather than novelties. More surprisingly, most users (11) strongly valued HCA's stories about his life and said that they learned a lot from them. The real surprise was that five users pointed out the system's value for training their English skills, casting an entirely different light on the system's educational potential from what we had anticipated.

An important part of the system criticisms (Q16) addressed HCA's less-than-human linguistic and conversational skills, with 11 comments. Four users admitted their English language difficulties at this point. The second-largest target of criticism were the graphics bugs. The system praise (Q17) may be summarised by

quoting the user who said that the system is on the right track overall. Essentially, the rich data on system improvement (Q18) expresses a wish for more of the same, with 14 comments. This time, probably having spent most of their ammunition already, only two users pointed out the need for better spoken input understanding. To the question (Q19) on the users' interest in speech-/gesture input computer gaming, no less than 12 users felt that spoken conversation might make all or some of their favourite games more entertaining, interesting, and immersive. Finally, the any other comments question (Q20) did not add much to the above.

5.2 Conclusions on Conversation Evaluation

In Section 3 we listed six principles of prototypical successful human-human conversation, including (1) search for common ground, i.e. shared knowledge, interests, etc., because (2) success depends on it; (3) interlocutor contribution symmetry in terms of activity and expertise-sharing; (4) expressive story-telling of, e.g., personal experiences, anecdotes, humour; (5) the permissibility of rhapsodic topic-shifts on a baseline of coherence; and (6) the perception of the conversation as having been worthwhile. In the following, we comment on the user test findings in relation to each of these principles.

When looking at the conversations from the user test it seems fair to conclude that, in general, there is a search for common ground (1). HCA is designed to ask questions to the user from time to time in order to continue an ongoing conversation or venture into a new topic, and he is also designed to sometimes only make observations, thus offering the initiative to the user. The users sometimes only answer or comment on what HCA says but, in many cases, they also ask questions to HCA concerning topics that it would be fair to expect he knows about, such as his age, his fairytales, and objects in his study. However, success (2) is only partial as evidenced by answers to Q7 (easy or difficult to use the system) and Q11 (could he follow what you wanted to talk to him about). Figure 8 shows smooth conversation with mainly HCA initiative in which HCA tries to give away initiative at the end. Figure 6 shows HCA's strategy of talking about something else when he is not able to understand what the user says and when the lack of understanding is not due to a (simulated) low recognition score. In the latter case, HCA will ask for repetition as illustrated in Figure 9 (Excuse me?). Figure 6 also illustrates that, given the rather sparse domain knowledge HCA has in PT1, he is not able to answer "why?" questions requiring explanations of his utterances, and also that there is still much he does not know about himself. Figure 9 illustrates successful user initiative. Figure 7 shows how HCA sometimes tends to focus on the same topic for too long and that he lacks the ability to leave a topic when the user has said that s/he does not know about it or is not interested in it.

Interlocutor symmetry (3) is a bit difficult to judge due to the test conditions. The data from the second test condition is, as expected, clearly influenced by the fact that the users asked very many questions to solve as many of the scenario problems as possible. The data from the first test condition typically shows a user who, at least to begin with and often to a considerable extent, lets HCA take the lead, cf. Figure 8.

HCA is designed to tell a goodly number of stories (4), e.g. about his childhood, his fairytales, and objects in his room. The user rarely tells stories and only when explicitly encouraged to do so by HCA. It is possible that the lack of user-volunteered stories is due to the experienced problems in having human-human-like in-depth conversation with HCA and his perceived difficulties in understanding the user (Q11).

HCA: What is your name?

User 3a: Paul

HCA: How old are you?

User 3a: have you forgotten

HCA: I am fifty-five years old. How old are you?

User 3a: eighteen

HCA: When I was your age I was a big fan of the theatre. Whenever there was a possibility I went to the theatre. It was my favourite place number one. Are you a boy or a girl?

User 3a: boy

HCA: I guess I was not a typical boy. I didn't like boy's games. Even at school I didn't take part in their games. Instead I stayed indoors. You might say that as a boy I was a peculiar child - a kind of an ugly duckling. Which country do you come from?

User 3a: Scotland

HCA: Well, now that you have told me something about yourself it is your turn. What do you want to talk to me about?

User 3a: fairy tales

HCA: Which of my fairy tales do you prefer?

User 3a: the emperors new clothes

Figure 8. Initial part of a conversation.

User 1a: picturecolosseumrome [no accompanying speech]

HCA: This is a picture showing Colosseum in Rome, one of the most impressive buildings of the world. I have been there! Italy is a very wonderful country and definitely one of my favourite destinations. I love to walk among the ruins of Rome, it inspires me so much.

User 1a: how many books have you wrote

HCA: I wrote around 180 fairy tales. That is actually quite a lot. And many of them are known all over the world. Some of the most well-known are The Little Mermaid The Ugly Duckling, and also The Princess and the Pea.

User 1a: what is your favourite fairy tale

HCA: Excuse me?

User 1a: what is your favourite fairy tale

HCA: I like several of my fairy tales very much. One of my favourites is the ugly duckling. The ugly duckling is not accepted by the other ducks and animals because it looks different. It has a hard time of suffering and troubles. Eventually, it wishes only to die till it discovers that it is in fact a beautiful swan. The morale of the fairy tale is that if you are different then you will get bullied.

Figure 9. A user asking HCA various questions, including pointing to a picture.

Rhapsodic topic-shifts (5) are only acceptable to a certain extent. It is fine to change to another topic if the current topic has, in some sense, been exhausted or closed. However, when HCA does not understand the user's input, he will start talking about some-

thing else which is usually not related to the current topic at all. The reason for this design is HCA's limited domain knowledge and conversational abilities in PT1. Our hypothesis was that when HCA starts to talk about something else, the conversation will move to a topic which HCA is more able to handle. However, when this happens too often, and especially when an unanswered question has been asked, users tend to get annoyed as expressed in the answers to Q11 (could he follow what you wanted to talk to him about), Q16 (what was bad about your interaction) and Q18 (suggested improvements).

Despite the problems experienced, nearly all users found that the conversation with HCA had been worthwhile (6) as illustrated, not least, through the positive answers to Q14 (entertainment value) and Q15 (educational value). Most users also thought that the addition of speech has the potential to enrich computer games (Q19) and make them more interesting and entertaining.

In summary, it seems fair to conclude from the user interviews that the system is on the right track overall but that there are still major challenges in improving HCA's conversational abilities.

6. DISCUSSION AND FUTURE WORK

Nobody who works in the area of (multimodal) spoken dialogue will be in doubt that the successful development of "real" conversational systems is a major challenge. We view the answers collected from the 18 users who participated in the user test of the first HCA prototype as, even surprisingly, encouraging. Overall, the users found that the technology is on the right track and represents a first glimpse of entirely new spoken computer games technology which could significantly improve the entertainment and educational value of computer games as well as attracting a new group of users who have not been so interested in traditional computer games. It should be noted, however, that the system's sheer novelty to the users, and the ease of imagining how future conversational systems could revolutionise computer gaming, museum visits, and self-training, may have carried many of the users away. We need quantitative metrics for conversation success.

The second HCA prototype is now being designed and developed with particular emphasis on increased conversational coherence and flexibility. The design and development is inspired by the data collected in the user test and data collected in an earlier, fully simulated Wizard of Oz setup of the system [5]. The second prototype will be ready by the end of 2004.

7. ACKNOWLEDGMENTS

The work described is being supported by the European Commission's Human Language Technologies Programme (HLT) under Grant IST-2001-35293. The support is gratefully acknowledged.

8. REFERENCES

[1] Bernsen, N. O.: When H. C. Andersen is not talking back In Rist, T., Aylet, R., Ballin, D. and Rickel, J. (Eds.): Proc. of the Fourth International Working Conference on Intelligent Virtual Agents (IVA'2003). Berlin: Springer Verlag 2003, 27-30.

[2] Bernsen, N. O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D., and Mehta, M.: First prototype of conversational H.C. Andersen. Proc. of the International Working Conference on Advanced Visual Interfaces (AVI 2004), Gallipoli, Italy, May 2004. New York: ACM 458-461.

[3] Bernsen, N. O., Dybkjær, H. and Dybkjær, L.: Cooperativity in human-machine and human-human spoken dialogue. Discourse Processes, Vol. 21, No. 2, 1996, 213-236.

[4] Bernsen, N. O. and Dybkjær, L.: Domain-Oriented Conversation with H.C. Andersen. Proc. of the Workshop on Affective Dialogue Systems, Kloster Irsee, Germany, June 2004. Berlin: Springer Verlag: Lecture Notes in Artificial Intelligence, Vol. 3068, 2004, 142-153.

[5] Bernsen, N. O., Dybkjær, L. and Kiilerich, S.: Evaluating Conversation with Hans Christian Andersen. Proc. of the Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, May 2004, Vol. 3, 2004, 1011-1014.

[6] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (Eds.): Embodied conversational agents. Cambridge, MS: MIT Press 2000.

[7] Clark, H.H.: Using Language. Cambridge: Cambridge University Press, 1996.

[8] Cole, R., van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W. and Yan, J.: Perceptive Animated Interfaces: First Steps towards a New Paradigm for Human-Computer Interaction. In van Kuppevelt, J., Dybkjær, L. and Bernsen, N.O. (Eds.): Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Dordrecht: Kluwer, 2004 (to appear).

[9] Corradini, A., Fredriksson, M., Mehta, M., Königsmann, J., Bernsen, N. O., and Johannesson, L.: Towards Believable Behavior Generation for Embodied Conversational Agents. Proc. of the Workshop on Interactive Visualisation and Interaction Technologies, IV&IT, Krakow, Poland, June 2004. Berlin: Springer Verlag: Lecture Notes in Computer Science, Vol. 3038, 2004, 946-953.

[10] Dybkjær, L., Bernsen, N.O. and Minker, W.: Evaluation and Usability of Multimodal Spoken Language Dialogue Systems. Speech Communication, Vol. 43, 2004, 33-54.

[11] Gabor, D.: How to Start a Conversation and Make Friends. London: Sheldon Press 2001.

[12] Grice, P.: Logic and conversation. In P. Cole and J.L. Morgan (Eds.): Syntax and semantics Vol. 3: Speech acts (41-58). New York: Academic Press 1975. Reprinted in Grice, P.: Studies in the way of words. Cambridge, MA: Harvard University Press 1989.

[13] Gustafson, J., Bell, L., Lindström, A. and Wirén, M.: The NICE Fairy-tale Game System. Proc. of 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, MA, 2004.

[14] Gustafson, J., Lindberg, N. and Lundeberg, M.: The August Spoken Dialogue System. Proc. of Eurospeech, 1999, 1151-1154.

[15] Turing, A.: Computing machinery and intelligence. Mind 59, 1950, 433-60.

[16] ISI: http://www.isi.edu/isd/carte/proj_tactlang/

[17] KTH broker: http://www.speech.kth.se/broker

[18] NICE: http://www.niceproject.com/