User Evaluation of Conversational Agent H. C. Andersen

Niels Ole Bernsen and Laila Dybkjær

Natural Interactive Systems Laboratory
University of Southern Denmark
nob@nis.sdu.dk, laila@nis.sdu.dk

10.21437/Interspeech.2005-406

Abstract

The Hans Christian Andersen (HCA) system is an example of a new generation of embodied conversational characters which are aimed to faithfully represent a familiar historical individual and carry out human-style conversation as that individual would have done had he or she lived today. A first prototype of fairytale author HCA was tested with representative users in January 2004. This paper reports on the user test of the second prototype which was done in February 2005, focusing on the structured user interview results.

1. Introduction

Today's embodied conversational agents still tend to carry out task-oriented spoken dialogue in which they help users accomplish one or several particular tasks. In several cases, the task involved is not even a complete and realistic one developed to the rigorous standards of task-oriented spoken dialogue systems [1] but merely an illustrative task fragment [4]. Instead, the developers focus on exploring various aspects of the enormous challenges faced in developing increasingly realistic natural interactive non-verbal output behaviours in context, including facial expression, gaze, lip synchrony, gesture, body posture, action on objects of discourse, etc. The spoken interaction often used to play second fiddle in such systems but there is now an increased focus on the importance of combining strong spoken interaction and good animation.

Traditionally, task orientation is synonymous with the provision, or gathering, of a well-circumscribed body of specific information deemed useful to the intended users. Complementary to the huge field of task-oriented systems is an equally huge field of systems for entertainment, learning, and for making friends through conversation, for getting to know others through in-depth conversation, etc. The boundary between task-oriented and non-task-oriented systems is not a strict one. Arguably, good teaching has an element of edutainment to it and even a dead-serious flight booking system might benefit from fully mixed-initiative spoken interaction, some amount of spoken social interaction and even, if we can get things right, embodied animated characters.

In the context outlined above, and with emphasis on both speech and animation, rich experimentation is taking place, moving the field of spoken multimodal education and entertainment systems towards non-task-orientedness. Not least in the USA, research in tutoring systems has been going on for some time. A major effort is the army training simulations which are being developed at USC, such as the mission rehearsal scenario system [9]. Other examples are the language tutors described in [8] and [5], the physics tutor described in [7], and the reading tutor presented in [11].

On the entertainment side there are not many examples. An interesting experiment in the late 1990s which pointed beyond task-orientation and towards entertainment systems was the Swedish August system that offered spoken interaction with the talking face of Swedish author August Strindberg about topics, such as restaurants in Stockholm and the Royal Technical University, task orientation being somewhat secondary to having fun [6].

Compared to the systems mentioned, the HCA system is aimed at edutainment and is clearly non-task-oriented. HCA is generally historically reliable as regards his looks, articulated personality, visible environment, etc. Yet he is far from being a historical person Q&A (question-answer) system. Rather, he is back and wants to make new friends amongst today's children and adolescents to the point of asking them if they might know a woman he could marry. We call the HCA system a *domain-oriented* system because HCA clearly is not task-oriented but aims to engage users in conversation about the domains of discourse he is familiar with or interested in, such as his life, fairytales, himself and his study, the user, and the user's favourite games. HCA has been developed in the European Human Language Technologies NICE project on Natural Interactive Communication for Edutainment (2002-2005). Computer games company Liquid Media, Sweden, did the graphics, Scansoft, Germany, trained the speech recogniser with children's speech, LIMSI, France, did the 2D gesture components and the input fusion, and we at NISLab did the natural language understanding, conversation management, and response generation.

Domain-oriented systems, no matter if primarily meant for entertainment or education, pose new demands on usability evaluation. Success can no longer be measured in terms of whether the user could solve the task(s). We rather have to somehow measure the extent to which the character manages the conversation successfully. In this paper, we look at the users' comments. Section 2 describes the system in more detail. Section 3 describes the user test of the second prototype. Section 4 discusses the post-trial interview data gathered. Section 5 concludes the paper.

2. The HCA system

The main goal of the HCA system is to demonstrate natural human-system interaction for edutainment by developing natural, fun and experientially rich communication between humans and embodied historical and literary characters. The target users are 10-18 years old children and teenagers. The primary use setting for the system is in museums and other public locations. Here users from many different countries are expected to have English conversation with HCA for an average duration of, say, 5-15 minutes.

The user sees HCA in his study in Copenhagen (Figure 1) and communicates with him in fully mixed-initiative conversation using spontaneous speech and 2D gesture. Thus, the user can change the topic of conversation, back-channel com-

ments on what HCA is saying, or point to objects in HCA's study at any time, and receive his response when appropriate. 3D animated HCA communicates through audiovisual speech, gesture, facial expression, body movement and action. The high-level theory of conversation underlying HCA's conversational behaviour is derived from analyses of social conversations aimed at making new friends, emphasising common ground, expressive story-telling, rhapsodic topic shifts, balance of "expertise", etc. [2]. When HCA is alone in his study, he goes about his work, thinking, meandering in locomotion, looking out at the streets of Copenhagen, etc. When the user points at an object in his study, he looks at the object and then looks back at the user before telling a story about the object.

Figure 1. HCA gesturing in his study.

Summarising, the HCA system may be viewed as a new kind of computer game which integrates spoken conversation into a professional computer games environment and aims to edutain through emulated human-human conversation. It is not meant for many hours of home-playing, though, the cover story being that HCA is just back and still has a hard time remembering all of what he once was, so, e.g., he only remembers details of three of his most famous fairytales, lots of things about his childhood but far less about his youth and adult life in Copenhagen and his travelling across Europe.

The user test of the first HCA prototype (PT1) confirmed that we were on the right track as regards the story-telling, let's-make-friends theory of conversation. However, we also learned that, as expected, PT1 was still too inflexible in its management of the conversation [2]. In addition, the PT1 user trial only simulated the system's speech recognition which was substituted by wizards typing in what the users said in real time. Finally, the PT1 character rendering was somewhat bugged and limited to one non-verbal primitive action at a time. The second prototype (PT2) user test was designed to test the system improvements made to remedy those deficiencies.

3. The PT2 user test

PT2 was tested with 13 users (six boys and seven girls) from the target user population. All users were Danish school kids aged between 11 and 16 and with an average age of 13 years. The interviews to be reported were conducted immediately after the users' interaction with the system.

3.1. User test setup

The PT2 user test was carried out in much the same way as the PT1 test, using two different test conditions and similar sets of user instructions in both conditions, cf. below. Significantly, the latter meant that the users were *not* instructed in how to speak to the system. In the PT1 test, this did not matter much since the wizards would simply type in what the users said, ignoring contractions, pronunciation variations, disfluencies, etc., only rarely committing typing errors. However, in the PT2 test which included a running speech recogniser, the lack of instruction on how to speak to the system was likely, a priori, to produce far more recognition errors than would have been the case had the subjects been thoroughly trained in how to speak to the system. Moreover, the PT1 user test had demonstrated the unfortunate effects or providing (mostly) computer game literate users with a mouse for pointing to objects of conversation, the result often being that a user would click on everything in sight, all the time, creating a pointing-to-objects ambience very far from that of pointing to objects during human-human conversation [3]. For this reason, the PT2 user test involved touch screen-only pointing which seems far closer to how people do (3D) pointing to objects in real conversation.

Users often arrived two at a time so two test rooms were prepared with the following setup: a touch screen, a keyboard (for changing virtual camera angle), a headset, and two cameras for recording the user-system interaction. The HCA software was running on two computers for practical reasons. The animation part was running on the computer connected to the touch screen and the rest of the system was on the second machine which was being monitored by a developer behind the user's back. In case of problems, the developer would take immediate action by, e.g., restarting a module causing the problem. Only rendering engine problems would require operations via the screen in front of the user. User input, system output, and interaction between modules was logged.

Figure 2. A user in action.

Each user test session took 60-75 minutes. Sessions began with a brief introduction to the system setup and the input modalities available, and calibration of the headset microphone to the user's voice. Then followed 15 minutes of free-style interaction in which it was entirely up to the user to decide what to talk to HCA about. In the following break, the

user was asked to study a handout which listed 11 proposals on what the user could try to find out about HCA's knowledge domains, make him do, or explain to him. Some examples are that the user could make HCA tell about his life and family relations, tell HCA about games the user likes, collect as much information as possible about the place where HCA lives, or be rude to him and see what happens. It was stressed that the user was not required to try to follow all the proposals. Rather, the user could pick those he or she liked, having a good time in the process. The second session had a duration of 20 minutes. Figure 2 shows a user in action during this session. A total of 26 conversations corresponding to 8 hours of speech were recorded, logged and captured on video.

Following the two sessions with HCA, each user was interviewed separately about his/her background, experiences from interacting with HCA, views on system usability, proposals for system improvements, etc., as detailed below.

4. The user interviews

In the PT2 user interviews, we asked a total of 31 questions. Eight initial questions dealt with the user's identity, background, computer game experience and experience in talking to computers. We had no substantial input on the final question on any other comments. This leaves 22 questions about the HCA system itself and how it was to interact with it, which are presented in abbreviated form in Figure 3. Compared to the 16 questions about the system in the PT1 interviews, new questions in the PT2 interviews addressed matters, such as, for input, talking and pointing at the same time, for output, HCA's audiovisual speech, and, as regards conversation management, how HCA dealt with errors and misunderstandings during conversation. Question re-phrasings primarily reflected a less HCA-centric question style.

Figure 3 presents a quantified summary of the PT2 interview results. Each user's verbatim response to each question was scored independently on a three-point scale by two raters. The general scoring principle followed may be roughly presented as 1 = high, with minor or no qualifications, 2 = reasonable but with qualifications, and 3 = low/negative. The general scoring principle was instantiated to each interview question, taking the specific contents of the question into account. Rating differences were negotiated by the two raters until consensus was reached. Finally, all user ratings per question were averaged to arrive at the summary shown in Figure 3. Admittedly, new raters might have rated some user answers slightly differently, at least initially, on the basis just described. Nevertheless, despite its qualitative and judgmental nature, the methodology does provide a means of summarising large amounts of user interview data in order to build a coarse-grained profile of how an entire user population views a system and their interaction with it.

Grouping the issues raised in the interviews, the following picture emerges, using 'Qn' for Question n.

As regards *pointing input,* users were very positive about using the touch screen (Q4). In general, HCA was aware of their pointing gestures (Q3). Half of the users were happy with the 2D gesture affordances in PT2 while the other half wished to be able to gesture towards more objects in HCA's study (Q5). Only a couple of users never tried to talk and

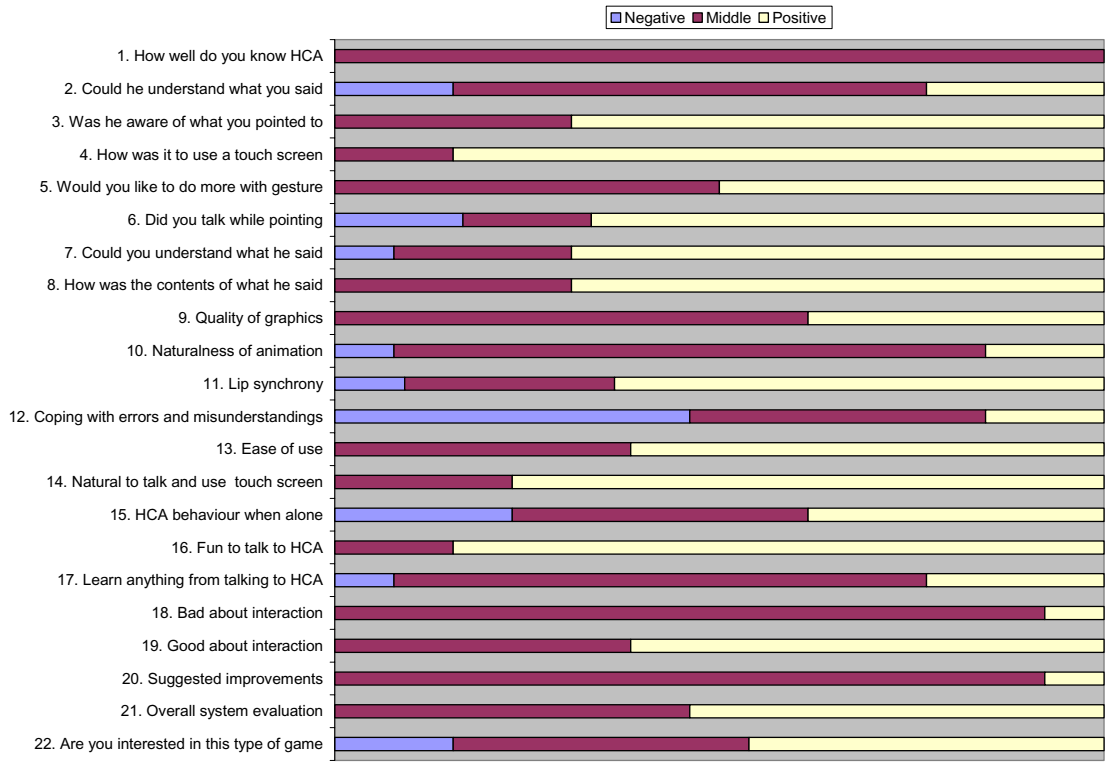

Figure 3. Summary of interview results form the PT2 user test.

point at the same time (Q6). The large majority of users found it natural to combine spoken and gesture input (Q14).

On *graphics and animation,* the overall quality of the graphics was viewed as rather good (Q9). So was the lip synchronisation which many compared to what they are used to in computer games. Only a single user remarked on the time delay between speech onset and lip movement onset (Q10). The naturalness of animation (Q11) received critical comments from most users. The key targets was HCA's walk which is often a gliding movement as if on rails. A couple of users found the animation fairly natural and one praised his facial movements. Users were the most critical of animated HCA when he was alone in his study (Q15). Part of this was due to an overheating graphics card in the first sessions, which made parts of HCA disappear. However, several users did not appreciate various antics made by the 55-years old man, such as squatting, jumping, gliding around bent forward, or negotiating a wall by repeated body impact.

On *speech understanding,* we found again, as in the PT1 interviews, that Danish kids understand spoken English amazingly well (Q7). Only a single user had a hard time understanding HCA. The question of whether HCA could understand the user's input (Q2) received a rather broad range of answers, from the damaging "Yes, a little more than half of the time" to "Almost all the time". Probably the most adverse comments concerned HCA's meta-communication abilities (Q12). As already remarked, users were not given instructions on how to speak to HCA. Many were initially uncertain as to what to say to him at all, and only few had spoken to a computer before. Disfluencies abound in the data, some users spoke lengthy sentences throughout, and it is our hypothesis that few managed to make significant adjustments to their speech behaviour during the sessions. For these reasons, we are positively surprised by the replies to Q2 but puzzled about their negative replies to Q12. Our hypothesis is that they did not tend to, e.g., rephrase and/or shorten their input when HCA did not understand them.

With respect to *fun and learning,* the users unambiguously found talking to HCA to be fun (Q16). All users except the one who did not understand HCA well, learned something from the conversation (Q17), primarily about his life and person, and about speaking English, rather than about his fairytales which Danish kids know quite well already (Q1). Correspondingly, users were generally positive towards the contents of the conversations (Q8).

On the issues of *what is good or bad and in need of improvement,* negative points (Q18) not made earlier included certain inconsistencies between the user's and HCA's control of his locomotion, and between camera angle and HCA's turning towards an object pointed to. Also, HCA should have more knowledge and improved prosody, and one user felt he takes offence too easily. Several users praised HCA's storytelling (Q19), the chance to have conversation with him, his "easy English" and good voice. The needs for improvements question (Q20) made the users re-emphasise some main messages, e.g. more knowledge to HCA, better walk, less antics, improved understanding and asking more questions of users. The system was generally regarded as easy to use (Q13).

In their *overall evaluation* (Q21), the users scored the system at 1.5 on a scale from 1.0 (great) through 2.0 (interesting) to 3.0 (somewhat negative). Ten users were interested in spoken computer games (Q22) for some or all gaming purposes. Two users simply did not play computer games, and a single user correctly pointed out that HCA is not presently fit for multi-hour home-gaming.

5. Conclusion

In general terms, the user test went quite well, with only a very limited number of module crashes, approx. one per effective hour of interaction, and what seems to us to be a thorough user critique of most aspects of the system and the interaction. The future potential of the kind of conversation illustrated by the HCA system, i.e., conversation for edutainment with famous people from our history, does seem to have been demonstrated by the user test interviews briefly reported in this paper. Our work has now turned towards in-depth analysis of the spoken conversation data and development of a strategy for quickly educating kids and teenagers in how to work effectively with speech recognition-based systems.

6. Acknowledgements

We gratefully acknowledge the support by the European Commission's HLT Programme, Grant IST-2001-35293.

7. References

[1] Bernsen, N.O., Dybkjær, H. and Dybkjær, L.: Designing Interactive Speech Systems. From First Ideas to User Testing. Springer Verlag 1998.

[2] Bernsen, N. O. and Dybkjær, L.: Evaluation of Spoken Multimodal Conversation. Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI 2004), Penn State University, USA, 2004, 38-45.

[3] Buisine, S., Martin, J.-C. and Bernsen, N. O.: Children's Gesture and Speech in Conversation with 3D Characters. Proceedings of HCI International 2005 (to appear).

[4] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (Eds.): Embodied Conversational Agents. Cambridge, MS: MIT Press 2000.

[5] Granström, B. and House, D.: Effective Interaction with Talking Animated Agents in Dialogue Systems. In [10].

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

[7] Litman, D. and Silliman, S.: ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. Proceedings of the HLT Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL), Boston, MA, 2004.

[8] Massaro, D.: The Psychology and Technology of Talking Heads: Applications in Language Learning. In [10].

[9] Traum, D., Marsella, S. and Gratch, J.: Emotion and Dialogue in the MRE Virtual Humans. Proceedings of the Workshop on Affective Dialogue Systems, LNAI 3068, Springer Verlag, Germany, 2004, 117-127.

[10] van Kuppevelt, J., Dybkjær, L. and Bernsen, N. O. (Eds.): Advances in Natural Multimodal Dialogue Systems, Springer Verlag, to appear.

[11] Wise, B., Cole, R., van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., Tuantranont, J. and Pellom, B.: Learning to Read with a Virtual Tutor: Foundations to Literacy. In Kinzer, C. and Verhoeven, L. (Eds): Interactive Literacy Education: Facilitating Literacy Learning Environments through Technology, Mahwah, NJ: Lawrence Erlbaum, to appear.