Comparative 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

Abstract

The Hans Christian Andersen (HCA) system is an experiment in resurrecting a familiar historical character and make this individual carry out human-style natural interactive conversation. We tested the first research prototype of fairytale author HCA with representative users in January 2004 and near-replicated the test setup when testing the second prototype in February 2005. This paper compares the results of the structured user interviews made right after each of the two tests.

1. Introduction

While most spoken dialogue systems – unimodal as well as multimodal – are still task-oriented, cf., e.g., [2][5] some recent research initiatives address applications which might be called domain-oriented systems. These systems do not aim to help the user solve particular tasks, whether well-structured or ill-structured but, rather, engage the user in conversation about one or several semi-open domains of knowledge and discourse. Examples of systems heading towards domain-orientedness are: the army training simulations which are being developed at USC, such as the mission rehearsal scenario system [9], various tutoring systems, see, e.g., [6][7][8], and the Hans Christian Andersen (HCA) system described in this paper.

The HCA system is aimed at edutainment. 3D animated fairytale author HCA is generally historically reliable wrt. his looks, articulated personality, visible environment, etc. He wants to have conversation with the user about the domains he is familiar with or interested in, such as his life, fairytales, himself, 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, CNRS-LIMSI, France, did the 2D gesture modules and the input fusion, and NISLab developed natural language understanding, conversation management, and response generation.

Domain-oriented systems pose new demands on evaluation. Success can no longer be measured in terms of, e.g., task completion rate. Rather, we have to somehow evaluate the extent to which the system successfully manages the conversation *as a conversation.* The users' opinions of the system are an important part of this evaluation. Two prototypes (PT1 [1] and PT2) of the HCA system were developed and tested with representative users in January 2004 and February 2005, respectively. The user tests were carried out following similar protocols. In both, subjective user data was gathered in post-trial interviews. Due to added functionality in PT2, the PT2 interviews included more questions than the PT1 interviews. In this paper, we compare the interview results. Section 2 describes the two prototypes and their main differences. Section 3 describes the user tests. Section 4 compares the interview data from the two tests. Section 5 concludes the paper.

2. The HCA prototype systems

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 an embodied historical character. Target users are 10-18 years old children and teenagers. The primary use setting is in museums and other public locations where users from different countries can have English conversation with HCA for an average duration of, say, 5-15 minutes. HCA, of course, is far from knowing everything the historical HCA knew. The cover story is that HCA is back but still has a hard time remembering all of what he once was, so, e.g., he only recalls details of three of his most famous fairytales, many things from his childhood but far less about his youth and adult life in Copenhagen and his travelling around Europe.

Common to the prototypes is that the user sees HCA in his study in Copenhagen (Figure 1) and has conversation with him, using spontaneous speech and 2D gesture. 3D animated HCA communicates through speech, gesture, facial expression, and body movement. The high-level theory underlying HCA's conversation is derived from analysis of social conversations aimed at making new friends, emphasising common ground, expressive story-telling, rhapsodic topic shifts, balanced interlocutor "expertise" (stories to tell), etc. [3].

The differences between PT1 and PT2 are significant. Perhaps the most important difference is that PT2 uses automatic speech recognition. In PT1, speech recognition is emulated by human wizards typing what the user says whereupon the system runs. PT1 thus has near-perfect speech recognition whereas PT2, as a complete running system, has to deal with the plethora of additional technical difficulties that arise from recognising the speech of children who, moreover, have English as their second language. Some other important differences between PT1 and PT2 are: PT2 enables full mixed-initiative conversation whereas PT1 only represents a limited approximation. Thus, in PT2, the user can change the topic of conversation, back-channel comments on what HCA is saying, or point to objects in HCA's study at any time, and receive his response when appropriate. This yields a far more flexible conversation than was possible in PT1. Also, the handling of miscommunication has been improved in PT2. Although HCA's domain knowledge has been slightly extended in PT2, the major change is in the re-structuring of his knowledge, i.e., in how the user can converse with HCA and get access to his knowledge, and in what HCA does when he has, or takes, the initiative.

Figure 1. Second prototype HCA in his study.

HCA's looks have become slightly more friendly in PT2 and his animation more articulate. In PT1, it was only possible to show one movement (animation primitive) at a time. For instance, HCA could not move his lips and lift an arm at the same time. PT2 HCA can display several gestures simultaneously and has semi-natural lip synchrony as well as some amount of face, arm and body movement. In PT1, HCA has a single output state, i.e., the one in which he produces conversational output. When no user is present, HCA does nothing but wait. In PT2, when alone, HCA walks around thinking, looks out his windows, etc. Unfortunately, this new output state is not properly integrated with the conversational output state, and HCA's behaviour when alone is also sometimes rather weird. The conversational output state was designed to consist of the active phase described above and a listening, or input-attentive, phase. However, the latter phase has not been integrated and the only listening behaviour HCA shows is that, when a user points to an object, he turns towards the object to look at it before he responds.

The handling of gesture input is improved in PT2. A major problem in PT1 was that the gesture recogniser was always open for input. When using the mouse (Section 3), users tended to create large queues of gestures waiting to be processed, which generated internal system problems as well as, sometimes, contextually inappropriate conversational contributions by HCA. The PT2 gesture recogniser does not "listen" while processing input. The same is true for the speech recogniser which does not have barge-in.

3. The user tests

PT1 was tested with 18 users (9 girls and 9 boys) from the target user group. Except for a young Scotsman, all users were 10-18 years old Danish school kids with an average age of 14 years. PT2 was tested with 13 users (7 girls and 6 boys) from the target population. All users were Danish school kids aged between 11 and 16 and with an average age of 13 years.

The user tests were carried out in much the same way, involving two test conditions and similar sets of user instructions for both conditions. Two test rooms were prepared with the following setup: a touch screen, except that for PT1 one of the rooms had a standard screen and the user used a mouse for pointing; a keyboard for changing virtual camera angles and make HCA walk; a headset; and two cameras for recording user-system interaction. Providing (mostly) computer game-literate users with a mouse for pointing to objects of conversation had the unfortunate result that they would tend to click on everything in sight, creating a pointing-to-objects ambience far from that of pointing to objects during human-human conversation [4]. The PT2 user test afforded touch screen pointing-only, which seems far closer to how people do (3D) gesture references to objects in conversation.

The software was running on two computers for practical reasons. The animation part was on the computer connected to the user's screen and the rest of the system was on the second computer which, for PT1, was operated by the wizard and for PT2 was being monitored by a developer out of sight of the user. In case of problems, the wizard/developer would take immediate action by, e.g., restarting a module causing the problem. Only rendering engine problems required operations via the user's screen. User input, wizard input (PT1-only), system output, and interaction between modules was logged.

Each user test session took 60-75 minutes. Sessions began with a brief introduction to the input modalities available. For PT2, the headset microphone was calibrated to the user's voice. 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., and only rarely committing typing errors. However, in the PT2 test with the speech recogniser included, the lack of instruction on how to speak to the system was likely, a priori, to produce more recognition errors than would have been the case had the subjects been trained in how to speak to the system. We wanted to collect baseline data on how second-language speakers of English, most of whom had not spoken to a computer before, manage to talk to a conversational system without prior instruction.

After the introduction followed 15 minutes of free-style interaction. It was entirely up to the user what to talk to HCA about. In the following break, the user was asked to study a handout which listed 13 (PT1) and 11 (PT2) proposals, respectively, for what the user could try to find out about HCA's knowledge, make him do, or explain to him. It was stressed that the user did not have to follow all the proposals. Rather, the user could pick those s/he liked whilst having a good time. The second session had a duration of 20 minutes. In total, approx. 11 hours of interaction were recorded on audio, video, and logfiles, respectively, for PT1, while about 8 hours of interaction were recorded in the same ways in the PT2 test.

4. The PT1 and PT2 user interviews

The PT1 and PT2 interviews comprised a total of 20 and 29 questions, respectively. In both cases, the first six questions concerned the user's identity, background, computer game experience and experience in talking to computers. For PT2, we also asked about the user's experience in touch screen use. In what follows, we focus on the part of the interviews which addressed system interaction and usefulness issues.

Seven PT1 and 14 PT2 questions dealt with the user's interaction with the system. Six PT1 and seven PT2 questions, concerned usefulness and improvements. The added PT2 usefulness question was on overall system evaluation. The final common question as to whether the user had any other comments did not give us any new information.

Questions + scoring criteria	Average PT1/PT2 scores with comments and explanations
1. Was it easy or difficult to use the system? Why? 1. easy, 2. qualifications, 3. difficult	**1.7/1.4.** Clearly better for PT2. However, the question focus is different, probably due to the different question orderings. The question was asked very early in the PT1 interviews but rather late in the PT2 interviews. As a result, users seem to have focused on different issues. Wrt. PT1, they focused on slow gesture understanding and deficiencies in how the system understood them. Wrt. PT2, they focused on minor difficulties of manual control of camera angles and HCA locomotion.
2. Could you understand what he said? 1: almost always, 2: qualifications, 3: difficult	**1.8/1.4.** Clearly better for PT2: the synthesis has fewer pronunciation errors and is perceived as being natural rather than basically intelligible. A male voice is used rather than a female voice. Due to our misunderstanding of available Scansoft synthesis, HCA had to use a female voice in PT1.
3. Could he understand what you wanted to talk to him about? 1: almost always, 2: qualifications, 3: many problems	**2.1/1.9.** Better for PT2: although speech understanding is significantly worse, conversation management is fully mixed-initiative, there is metacommunication support, and the number of crashes and bugs is far smaller. Gone are the PT1 problems of: many questions not answered, several unwanted repetitions, HCA did not follow user topic change due to an overly inflexible conversation structure. HCA's domains of knowledge are only slightly more extensive in PT2.
4. What do you think of the naturalness of the animation? 1: fine, fun, realistic, 2: qualifications, 3: negative	**1.6/1.9.** PT2 is clearly doing worse: even if PT1 has more graphics bugs, so that HCA, e.g., walks on the ceiling and stands amidts his furniture, there is strong user reaction to PT2 HCA's unnatural walk in which he often glides rather than walks, and his occasional unnatural posture in which he glides along bent forward. These problems are due to lack of output state integration. Lip synchrony is appreciated, though (cf. 11), but HCA's emotion display was only noticed by a single user.
5. How was it to do the gestures? 1: fine, 2: qualifications, 3: not so good	**1.5/1.2.** Clearly better for PT2: the touch-screen is now praised as giving more control than the mouse, even though this contradicts the PT1 scoring for mouse vs. touch screen in which the touch screen was felt to be strange and difficult to use. In addition, some gestures did not work in PT1.
6. Would you like to be able to do more with gesture? If yes, what? 1: no, 2: some more, 3: much more	**1.9/1.4.** Clearly better for PT2. This may be due to (i) the absence of the PT1 "anonymous objects" about which HCA always only says that he does not know much about them yet. PT2 HCA thus also avoids repeating himself in these cases; (ii) the fact that it is far more work to point using the touch screen than using the mouse may also have reduced the wishes for more gesturable objects. Note that the number of objects HCA could tell a story about is the same in PT1 and PT2.
7. Was it fun to talk to HCA? If yes, what was fun? If no, what could make it fun? 1: yes, 2: sometimes, 3: no	**1.7/1.2.** Clearly better for PT2: the conversation is far better, smoother and more flexible, and there are significantly less bugs and crashes.
8. What did you learn from talking to with HCA? 1: a lot, 2: some things, 3: don't know/ not much	**2.0/1.8.** Better for PT2. The reasons for the improved average scoring are not clear since the users' actual comments are very similar: they learnt quite a lot about HCA himself, his life and his family whereas his fairytales were known already, and they very much appreciated the opportunity to learn English by speaking with HCA.
9. What was bad about your interaction with HCA? 1: nothing/ don't know, 2: some things, 3: a lot	**2.1/1.9.** Better for PT2 which has far better conversation with much less of: did not change topic when the user wanted, irrelevant replies, too much repetition, did not answer questions; less graphics bugs, lip synchrony added, better synthesis. The bad walk and posture is still a strong negative. New, smaller problems are now remarked upon, such as inconsistent locomotion control and that HCA takes offence too easily. For both PTs, the users want HCA to have more knowledge.
10. What was good about your interaction with HCA? 1: several things, 2: one thing, 3: nothing/don't know	**1.8/1.4.** Clearly better for PT2: far better conversation, the touch screen is more natural than the mouse for the application, the synthesis is excellent. The PT1 and PT2 users agreed that it was good to talk to HCA in English, get information about himself and his life, and point to objects and get stories about them.
11. What do you think we should make better? 1: minor things, 2: substantial improvements, 3: most or all of it	**1.7/1.9.** Worse for PT2: the animation was deemed worse than in PT1 despite more functionality, such as lip synchrony and more facial expressions, and fewer bugs. The users, moreover, imposed new, more advanced requirements, thus raising the stakes of system development. Examples are: HCA should ask more questions of the user, he does not hear and understand as well as a human, he should be more active, and he should understand more ways of expressing things.
12. How interested would you be in playing computer games with speech and gesture? 1: very, 2: depends, 3: not interested	**1.6/1.7.** The PT1 and PT2 users provided very similar replies: spoken computer games are good for strategy games, the SIMS if one could participate oneself, mission negotiation, adventure games, museum games, wargames, and school teaching. Computer games are more life-like, entertaining, and immersive with speech and pointing. Speech may be better for learning than for entertainment, though.

Figure 2. Comparison of PT1 and PT2 interview results.

Figure 2 compares replies to questions on interaction and usefulness common to PT1 and PT2. The left-hand column shows the interview question followed by the scoring criteria used for rating each user's reply. The right-hand column shows, first, for each question, the average scores for PT1 and PT2, respectively. Each user's verbatim response to each question was scored independently on a three-point scale by two raters. The general scoring criterion applied may be presented as 1 = high, with minor or no qualifications, 2 = reasonable but with qualifications, and 3 = low/negative. This general criterion was instantiated to each interview question, cf. Column 1, 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 scores for PT1 and PT2 shown in Figure 2. Secondly, Column 2 shows our comments and proposed explanations for the scoring differences between PT1 and PT2. Grouping the issues raised in the interviews, the following picture emerges, using 'Qn' for Question n and 'QAv' for a question's average score.

It seems clear that HCA's spoken *conversational abilities* have improved significantly in PT2. Conversation management problems do not enter into the replies to PT2-Q1 and PT2-Q7, and only rarely into the replies to PT2-Q9, but figure prominently in the corresponding replies regarding PT1. Moreover, Q3 and also Q10 show a decrease in average score from PT1 to PT2 despite the significant decrease in speech recognition performance in PT2. Interestingly, criticism of HCA's conversational abilities surfaces for both PT1-Q11 and PT2-Q11. In both cases there is a wish that HCA can understand more. However, for PT1 the main criticism is that he should know more while for PT2 it is his vocabulary size that is criticized. Conversely, the strong increase in *animation* articulation and expressiveness, and the reduction of the number of graphics bugs, in PT2 over PT1, is not rewarded by the users, cf. Q4, Q9 and Q11. Rather, PT2's unnatural animation is punished quite severely. Overall *interaction* is generally praised for PT2 (Q1) as is PT2's speech synthesis (Q2), the touch screen as input device (Q5) and the use of pointing (Q6). We have no immediate explanation of why the PT1 users were far more critical of the touch screen than the users of PT2 (Q5). Finally, the users' views on *learning* (Q8) from the system and on *future prospects* (Q12) for speech/-gesture computer gaming are similar for PT1 and PT2.

Two questions were only asked for PT1 and nine only for PT2. These are not shown in Figure 2. Nearly all of them concern interaction and reflect new PT2 functionality. One PT1-only question asks what the user thinks of the HCA character. He is basically viewed as authentic apart from his (female!) voice at QAv=1.4. For PT2 we included instead a more general question about the quality of the graphics, which was rated as good overall (QAv=1.6). Two more questions on the visual impression of PT2 were: one on the lip synchrony which users found quite good (QAv=1.5), and a question about HCA's behaviour when he is alone in his study. The QAv at 1.8 confirms what we have observed about animation already.

The second PT1-only question asked how it feels to *talk* to HCA. The QAv at 1.7 mainly reflects that users' took some time getting accustomed to speaking to the system. The more general PT2 question asks how natural it is to talk and use the touch screen. At QAv=1.2, users were clearly very positive. A new, related PT2 question was if the user talked while pointing and if it worked. Half of the users did not talk while pointing while the rest did so occasionally. The QAv=1.5 reflects that the multimodal input worked for almost all users who tried. The related question about HCA's understanding of pointing input was answered very positively at QAv=1.3.

Of the three final PT2-only questions, one was about the quality of the contents of what HCA says. The users were generally very positive at QAv=1.3. The question about how easy it was to cope with errors and misunderstandings received the harshest QAv of all (2.3). Only two users found this to be rather easy while half of them found it difficult. Our main explanation so far is that this is where the system's imperfect speech recognition and limited vocabulary and domain knowledge take centre-stage. The users' overall evaluation of the system was good at QAv = 1.5. This is close to their averaged rating of PT2 across all questions, at 1.59, whereas PT1 was rated at 1.74 overall.

5. Conclusion

This paper has reported first results from what may be a relatively rare exercise of performing similarly protocolled user tests with two subsequent research prototypes. The next step in our work is to correlate the subjective evaluations reported with objective analysis of the conversations.

6. Acknowledgements

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

7. References

[1] 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. Proceedings of AVI 2004, Gallipoli, Italy, 2004, 458-461.

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

[3] Bernsen, N. O. and Dybkjær, L.: Evaluation of Spoken Multimodal Conversation. Proceedings of ICMI 2004, Penn State University, USA, 2004, 38-45.

[4] 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).

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

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

[7] Litman, D. and Silliman, S.: ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. Proceedings of 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.: Advances in Natural Multimodal Dialogue Systems. Springer. Series: Text, Speech and Language Technology, Vol. 30, Springer, 2005.