*NICE project (IST-2001-35293)*

# Natural Interactive Communication for Edutainment

# NICE Deliverable D7.2a

# Evaluation of the First NICE HCA Prototype

*19 April 2004*

*Part 1 authors*

*Niels Ole Bernsen[1] and Laila Dybkjær[1]*

*1: NISLab, Odense, Denmark*

*Part 2 authors*

*Stéphanie Buisine[1] and Jean-Claude Martin[1]*

*1: LIMSI-CNRS, France*

| | |
|---|---|
| **Project ref. no.** | IST-2001-35293 |
| **Project acronym** | NICE |
| **Deliverable status** | Public |
| **Contractual date of delivery** | 1 January 2004 + agreed 6 weeks delay = 15 February 2004 |
| **Actual date of delivery** | 19 April 2004 |
| **Deliverable number** | D7.2a |
| **Deliverable title** | Evaluation of the First NICE HCA Prototype |
| **Nature** | Report |
| **Status & version** | Final |
| **Number of pages** | 30 |
| **WP contributing to the deliverable** | WP7 |
| **WP / Task responsible** | WP7/NISLab |
| **Editor** | - |
| **Author(s)** | Part 1: Niels Ole Bernsen and Laila Dybkjær<br>Part 2: Stéphanie Buisine and Jean-Claude Martin |
| **EC Project Officer** | Mats Ljungqvist |
| **Keywords** | User testing and evaluation, domain-oriented speech and gesture conversation, embodied conversational agents |
| **Abstract (for dissemination)** | This report, Deliverable D7.2a of the EC Human Language Technologies project NICE (Natural Interactive conversation for Edutainment), presents results of the January 2004 user evaluation of the first NICE Hans Christian Andersen system prototype. Part 1 focuses on evaluation of the spoken conversation, Part 2 on evaluation of gesture input processing. |

# Table of Contents

# Part 1, Evaluation of the Spoken Conversation, NISLab

## 1 Introduction

The first integrated NICE Hans Christian Andersen (HCA) prototype (PT1) was completed in December 2003. This report, NICE Deliverable D7.2a, presents our evaluation of PT1 based on a user test with the running prototype conducted at NISLab in late January 2004. The user test is described in detail in NICE Deliverable D2.2a, *NISLab's Collection and Analysis of Multimodal Speech and Gesture Data in an Edutainment Application.*

Briefly, 18 users from the target user group of 10-18 year olds used the system in two different conditions. In the *first condition*, they had unconstrained conversation with HCA based only on instructions on how to change the virtual camera angle, control HCA's locomotion in his study, and speak and input gesture to the system. This enabled them to become familiar with the system. In the immediately following, *second condition*, the users spent 20 minutes trying to solve as many problems as possible from a hand-out problems list which included 13 problems common for all subjects. Immediately after the second-condition-interaction with HCA, each user was interviewed by a NISLab HCA system developer. The structured interviews were based on a common set of questions, cf. Deliverable D2.2a.

Considering the requirements to test users proposed in NICE Deliverable D7.1, *Evaluation criteria and evaluation plan*, Section 5, the following observations may be made on how the actual user test conformed with those requirements (italicised below):

- *Each prototype should be evaluated by at least 12 test users.* The NICE HCA PT1 was evaluated with 18 users.
- *Age: at least 8 users should belong to the primary target group.* All users belonged to the primary target group of 10-18 year olds.
- *Both genders should be represented approximately equally.* The test group included 9 girls and 9 boys.
- *User background diversification.* The user group shows a good spread in computer game literacy, from zero game hours per week to +20 hours game per week. They were all school children but this is what you do nowadays if you are in the target group.
- *Language background diversification.* Only a single user was not Danish (an 18 years old Scotsman). However, the large-scale (approx. 500 users) in-field Wizard of Oz studies conducted in the summer of 2003 at the HCA Museum in Odense, Denmark, included users of 29 different nationalities, cf. NICE Deliverable D2.2a. WE believe, therefore, that we already have voluminous data on the differential behaviour in conversation of users with different nationalities and first languages.

Following the user test, we have analysed the user test logfiles and the interview results from various perspectives and using various methodologies. In the following, we describe those results of the analyses made which contribute to an evaluation of NICE HCA PT1. Prior to that, and in order to better enable readers to judge the evaluation results, Section 2 briefly describes the NICE HCA PT1 which was tested in the user test. In Section 3, we present a comprehensive evaluation summary of PT1 following the NICE evaluation criteria proposed in NICE Deliverable D7.1. Section 4 presents the results of analysing the user interviews.

Section 5 briefly presents ongoing work on in-depth analysis of the user-HCA conversations logged during the user tests. Section 6 presents some main conclusions and briefly discusses next steps in our work.

If we were to summarise the NICE HCA PT1 evaluation described below, it seems justified to state that:

- with the exceptions made in the two following bullet points, the user-tested PT1 conformed to the PT1 requirements and design specification presented in NICE Deliverable D1.1, *Requirements and design specification for domain information, personality information and dialogue behaviour for the first prototype*, and expanded in NICE Deliverables D1.2a, *Analysis and representation of domain information, personality information and conversation behaviour for H.C. Andersen in the first prototype,* and D5.1a, *First Prototype Version of Conversation Management and Response Planning for H.C. Andersen;*
- in one important respect, i.e. the inclusion of natural language understanding, the user-tested PT1 had more functionality than planned;
- in another respect, i.e. input fusion, the user-tested PT1 had less functionality than planned; and
- as a whole, the user-tested PT1 performed as we expected at this stage of development.

We are of course aware that the evaluation to be presented in this and the following sections is not an *independent* one. That is, it is we, the PT1 system developers, who have done the evaluation and presented our results in the present report.

If the question is asked why we have chosen to evaluate the NICE HCA PT1 with target group users in a controlled laboratory test, the answer is the following. A field test, like the one done at the HCA Museum in the summer of 2003, is much harder to control than a laboratory test. It is difficult or impossible to instruct users adequately in the field, to ensure a strict dual-condition experimental regime, to interview the users, and to video record the users in action with all that this entails in terms of informed consent, permission signatures, and rights to use the recorded data. When conducting costly testing of a first system prototype, it is critically important to collect data based on a corpus design which is optimised for the purpose of getting the interaction information which is most needed in order to judge how the system performs, since analysis of this data will be crucial for the process of functional extension, re-specification, and re-design which is planned to follow the user test. To ensure full control of the corpus collection process, laboratory experimentation seems to be the only viable approach.

# 2 The NICE HCA PT1 system

Figure 2.1 shows the overall NICE HCA system architecture.

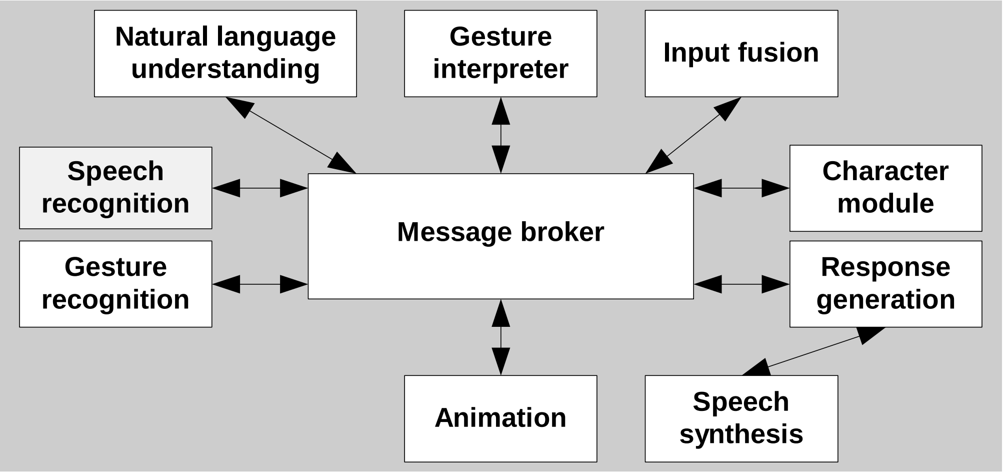

**Figure 2.1.** Overall NICE HCA system architecture.

Compared to the architecture in Figure 2.1 and according to plan, NICE HCA PT1 does not yet have speech recognition. However, PT1 is ahead of plan in that it already includes a natural language understanding module. We have accelerated the development from scratch of a natural language understanding module for PT1 in order to be able to conduct sufficiently realistic and informative user testing at this stage in the project. On the other hand, PT1 is behind plan in that it does not yet have a semantic input fusion module. Instead, gesture input is sent via gesture recognition, gesture interpretation, and an input fusion module which merely forwards this input data, to the character module which takes care of managing gesture input in the conversation context. For more information on the NICE HCA PT1 system, and in addition to the information available in NICE HCA deliverables, we refer to the following papers which have been published already or accepted for publication: [Bernsen 2003], [Bernsen et al. 2004a], [Bernsen et al. 2004b], [Bernsen and Dybkjær 2004], and [Corradini et al. 2004].

# 3 Evaluation according to the NICE evaluation criteria

## 3.1 Introduction

In this section, we apply the NICE evaluation criteria specified in NICE Deliverable D7.1 to the first NICE HCA prototype. It should be noted that D7.1 envisions that "Only a (relatively large) subset of the evaluation criteria ... will be applicable to the first prototype ...". Below, we have tried to apply most of the criteria to the NICE HCA PT1 system. Application has been done based on:

- user, system and equipment observation during the user tests at NISLab;
- analysis of the interviews made with all 18 users, cf. Section 4 below;
- user test logfile analysis, cf. Section 5 below; and
- extensive internal discussion of the evaluation table presented in the next section. Obviously, the evaluation performed is to a large degree qualitative and judgmental rather than quantitative. Such is the nature of many of the criteria in the Table 3.1 below.

## 3.2 User test evaluation table

Table 3.1 shows the results of applying the NICE evaluation criteria.

The numerical scores in the third column from the left of Table 3.1 have been assigned following user test evaluation at NISLab. '1' is the lowest score, '5' the highest score. The scores are commented upon in Column 3 to make clear how each score allocated relates to what we believe should be expected of PT1. For instance, in PT1, one is entitled to expect a better approximation to real-time performance than to perfect handling of domain-oriented conversation, the latter problem being one of the main research challenges in the NICE project. If real-time performance is a serious problem in PT1, we may have an unpleasant and unexpected problem on our hands at this stage, whereas if conversation management is not perfect in PT1, this is only what everyone would be entitled to expect. More generally speaking, we consider a score of '3' for all main challenges addressed in the NICE HCA system clearly adequate at this stage of development. Still, we need to stress, again, the qualitative and judgmental nature of many of the scores assigned in Table 3.1.

The difference between 'as planned' and 'as planned for PT1' in Column 3 is that the former expresses what has been planned for the NICE HCA system for the project as a whole.

Compared to the criteria presented in NICE Deliverable D7.1, the criteria in Table 3.1 have sometimes been (i) slightly re-worded or (ii) split into several criteria for increased clarity of the evaluation made. The former is marked by an '**r**' in the first column from the left, the latter by an '**s**'. Finally (iii), new criteria are marked by an '**n**'.

NICE Deliverable D7.1, Section 4, presents some key planned properties of the NICE HCA PT1 system. When Table 3.1 Column 3 includes 'ahead of plan', this means that a certain property evaluated was not planned to be included in NICE HCA PT1. This applies, in particular, to the natural language understanding component which, according to the Work Programme, is due only by Month 32.

| **Criterion** | **Evaluation** | **Score 1-5** |
|---|---|---|
| Technical criteria | | |
| Technical robustness | Some crashes and a number of loops, improvement needed, see Section 5 | 3 acceptable for PT1 |
| Handling of out-of-domain input | Further improvement needed, see Section 5 | 2 acceptable for PT1 |
| **r,s** Real-time performance, spoken part | OK, natural language understanding is fast | 5 very good for PT1 |
| **r,s** Real-time performance, gesture part | Further improvement needed | 3 basic for PT1 |
| Barge-in | - No barge-in in PT1 | As planned for PT1 |
| Number of characters | 1 | As planned |
| Number of emotions which can be expressed by characters | 4 basic emotions | 4 good for PT1 |
| Actual emotion expression verbally and non-verbally | Much improvement needed, particularly in rendering capabilities: scripts, synchronous non-verbal expressions, speed, amplitude | 1 as planned for PT1 |
| **s** Number of input modalities | 3, i.e.: speech, 2D gesture, user key haptics inconsistent with character autonomy | As planned for PT1 |
| **s** Number of output modalities | Natural interactive speech, facial expression, gesture. More rendering capability needed | As planned for PT1 |
| Synchronisation of output | Speech/gesture/facial OK. More rendering capability needed. No lip synchronisation | As planned for PT1 |
| Number of domains | 6, i.e. HCA's life, fairytales, physical presence, user, gatekeeper, meta | As planned |
| Number of different plots/scenes available | N/A: HCA will have no plots | As planned |
| Basic usability criteria | | |
| Speech understanding adequacy | No speech recognition in PT1<br>Natural language processing in PT1: limited but better than basic | As planned<br>3 acceptable for PT1. Ahead of plan |
| Gesture understanding adequacy | Further improvement needed | 3 basic for PT1 |
| **n** Combined speech/gesture understanding adequacy | - No semantic input fusion module in PT1 | Behind plan for PT1 |
| Output voice quality | Mostly OK, intelligible, not unpleasant, modest syllable swallowing | 4 good for PT1 |
| Output phrasing adequacy | Mostly OK, no user remarks | 4 good for PT1 |
| Animation quality | Further improvement needed in rendering capabilities and output design, cf. above | 3 acceptable for PT1 |
| Quality of graphics | Rather good, only a (true) user remark on too dark graphics due to the study light sources | 4/5 very good for PT1 |
| Ease of use of input devices | Microphone, mouse, touch screen, keyboard: users generally quite positive | 4/5 very good for PT1 |
| **s** Frequency of interaction problems, spoken part | A larger number of bugs, primarily loops, found than was expected. A total of 13.3% of the output was found affected by bugs. The non-bugged interaction, on the other hand, showed better performance than expected. See | Bugged interaction: 2 barely adequate for PT1<br>Non-bugged |

| | | |
|---|---|---|
| | also Section 5. | interaction: 3/4 acceptable for PT1 |
| **s** Frequency of interaction problems, gesture part | Some bugs, an algorithm problem, a stack problem, no waiting function | 3 basic for PT1 |
| **s** Frequency of interaction problems, graphics rendering part | Two serious generic bugs found: users get lost in space outside HCA's study, HCA immersed in furniture | 2 barely adequate for PT1 |
| Sufficiency of domain coverage | Approx. 300 spoken output templates and 100 primitive non-verbal behaviours: further improvement needed | 3/4 acceptable for PT1. Ahead of plan |
| Number of characters the user interacted with in the fairy tale world | N/A HCA's study is distinct from the fairytale world | N/A |
| **r** Number of objects the subject(s) interacted with through gesture | 21 pointable objects in HCA's study: in general, the users pointed to most of them. | 3 acceptable for PT1 |
| Navigation in the fairy tale world | N/A HCA's study is distinct from the fairytale world | N/A |
| Number of topics addressed in the conversation | All generic topics (approx. 30), not all topic details | As expected |
| Core usability criteria | | |
| **r** Conversation success | Most users pointed out that HCA's responses were sometimes irrelevant. Due to loops and core research difficulties. See Sections 4 and 5 | 3/4 acceptable/ good for PT1 |
| **r** How natural is it to communicate via the available modalities | Very positive user comments overall | 4/5 very good for PT1 |
| Output behaviour naturalness | Very complex criterion, hard to score. Still, users were surprisingly positive, see Section 4 | 3/4 quite acceptable for PT1 |
| Sufficiency of the system's reasoning capabilities | Capabilities are basic at this stage | 3 acceptable for PT1 |
| **r** Ease of use of the game: How well did users complete the scenario tasks? | Difficulties mainly due to loops and conversation management | 3 acceptable for PT1 |
| **s** Error handling adequacy, spoken part | Limited in PT1. User test data and speech recogniser addition needed for identifying problems and designing improvements | 2 acceptable for PT1 |
| **s** Error handling adequacy, gesture part | - No error handling involving gesture | Behind plan for PT1 |
| Scope of user modelling | User age, gender and nationality collected, age used | Ahead of plan for PT1 |
| Entertainment value | User test very positive, see Section 4 | 4 good for PT1 |
| Educational value | User test very positive, see Section 4 | 4 good for PT1 |
| User satisfaction | User test very positive, see Section 4 | 4 good for PT1 |
| Technical component evaluation | | |
| Speech recogniser | | |
| Word error rate for English | N/A No speech recognition in PT1 | N/A |
| Vocabulary coverage for English | N/A No speech recognition in PT1 | N/A |
| Perplexity of English language model | N/A No speech recognition in PT1 | N/A |
| **r** Real-time performance | N/A No speech recognition in PT1 | N/A |
| Gesture recogniser | | |
| Recognition accuracy regarding gesture | See LIMSI evaluation report in Part 2. | |

| | | |
|---|---|---|
| type | | |
| Number of recognition failures | See LIMSI evaluation report in Part 2. | |
| Number of interpretation errors | See LIMSI evaluation report in Part 2. | |
| Natural language understanding | | |
| Lexical coverage, English | 66% | Ahead of plan |
| Parser error rate, English | 16% | Ahead of plan |
| Topic spotter error rate, English | Not evaluated for PT1 | As planned |
| Anaphora resolution error rate, English | Not in PT1 | As planned |
| Gesture interpretation | | |
| Selection of referenced objects error rate | See LIMSI evaluation report in Part 2. | |
| Input fusion | | |
| Robustness to temporal distortion between input modalities | - No semantic fusion. No fusion of data structures because no waiting function for NLU input when gesture input | Behind plan for PT1 |
| Fusion error rate | - No semantic fusion | Behind plan for PT1 |
| Cases in which events have not been merged and should have | - No semantic fusion | Behind plan for PT1 |
| Cases in which events have been merged and should not have | - No semantic fusion | Behind plan for PT1 |
| Recognised modality combination error rate | - No semantic fusion | Behind plan for PT1 |
| Character module | | |
| Meta-communication facilities | Repeat, low CS, (insults) | As planned for PT1 |
| Handling of initiative | Limited free user initiative: not in mini-dialogues | As planned for PT1 |
| Performance of conversational history | Support for meta-communication and mini-dialogues | As planned for PT1 |
| Handling of changes in emotion | HCA's emotional state updated for each user input | As planned for PT1 |
| Response generation | | |
| Coverage of action set (non-verbal action) | Too limited since only one non-verbal action can be realised at a time. | 2 acceptable for PT1 |
| Graphical rendering (animation) | | |
| Synchronisation with speech output | Works for a single non-verbal element at a time<br>No lip synchronisation | As planned for PT1 |
| **s** Naturalness of animation, facial | Overlapping non-verbal elements missing<br>Limited number of animations<br>First experimental implementation | As planned for PT1 |
| **s** Naturalness of animation, gesture | Overlapping non-verbal elements missing<br>Limited number of animations<br>First experimental implementation | As planned for PT1 |
| **s** Naturalness of animation, movement | Users: strange HCA walk | As planned for PT1 |
| Text-to-speech | | |

| | | |
|---|---|---|
| Speech quality, English | OK | 4 good for PT1 |
| Intelligibility, English | Some syllables swallowed | 4 good for PT1 |
| Naturalness, English | OK | 4 good for PT1 |
| Non-speech sound | | |
| Appropriateness in context of music/sound to set a mood | N/A None in PT1 | N/A |
| Integration | | |
| Communication among modules | PT1 is reasonably well-tested | 4 good for PT1 |
| Message dispatcher | OK | 4/5 good for PT1 |
| Processing time per module | Real-time overall, except for gesture modules | 5/3 fine/basic for PT1 |

**Table 3.1.** Evaluation of the NICE HCA first prototype based on the NICE evaluation criteria, test observations, user interviews, and conversation analysis.

## 3.3 Conclusion

It seems fair to conclude from Table 3.1 that, overall, the NICE HCA PT1 has worked reasonably well during the user tests and that it has been received remarkably well by the target users. The primary exception is the gesture input components which, we hope, will exceed expectations in the final year of the NICE project. Also, we would have preferred to find a smaller number of bugs than was actually found wrt. (a) the spoken interaction and (b) the workings of the rendering when users made HCA do locomotion in his study.

Given the novelty of the NICE approach compared to the state-of-the-art, and despite the fact that we have conducted, in two iterations, very substantial WoZ simulations earlier in the project, cf. NICE Deliverable 2.2a, the user tests with the implemented first system prototype, only excepting the speech recogniser, were replete with uncertainties as to how the users would receive, perceive, and react to, the system. To mention but a few, those uncertainties concerned:

- the system architecture and its innovative use of a personalised conversational agenda for early output planning;
- the innovative mini-dialogues processed by a dedicated mini-dialogue processor;
- a specialised domain agent for meta-communication;
- a new natural language processor;
- real-time performance and robustness;
- the abandonment of traditional hard-coded dialogue structures, using instead a knowledge base approach for conversational output identification;
- the collaboration with graphical computer game professionals;
- the collaborative graphics rendering of HCA, his study, and his non-verbal behaviours;
- the underlying theory of social conversation which we have developed for the purpose of the NICE HCA system; and
- our interpretation of HCA’s personality as reflected in his communicative behaviour and otherwise.

Now that we have established the target users’ perceptions of the first prototype system, we are, more than anything, surprised and relieved to find that, essentially, those design decisions were rather sound ones which do not have to be re-made for PT2 purposes. On the contrary, the users were enthusiastic about them, stating that the NICE HCA PT1 system made them glimpse a new generation of more entertaining, fun and immersive computer edutainment applications.

This suggests that we can go on to develop PT2 on far more solid foundations than we had at the start of the NICE project, avoiding much frustrating design error-correcting re-design and re-implementation. Instead, we can dedicate the final year of the project to (i) further addressing the key challenges of the NICE project, including conversation management for domain-oriented spoken conversation, life-like embodied non-verbal behaviour, conversation domain extensions, improved natural language understanding, speech recogniser incorporation, improved non-verbal behaviour design-for-rendering, flexible verbal and non-verbal emotion expression, ontology development for HCA's knowledge representation, improved user modelling, system portability, possibly machine learning of domain ontologies, and more.

# 4 Evaluation according to the user interviews

During the structured interview which was made after interaction with PT1, each user was asked the following questions:

**User information**

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?

**Interaction**

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?

13.1.1. Did you use the mouse or point onto the screen?

13.1.2. How was it to do the gestures?

13.1.3. Would you like to be able to do more with gesture? If yes, what?

**Usefulness and improvements**

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?

**Other**

20. Any other comments?

The list below is a revised interview questions list adapted to serve as explanation for Table 4.1. Included in the list are the three-point scales which were qualitatively applied to the user's responses to the major interview questions which required a subjective reply. In addition, we have added one or two examples of user responses per interview question. Each response comes with the three-point scale score we have given it, illustrating our use of the three-point scales. The number in parenthesis after an example refers to the user number in Table 4.1.

i. User number

ii. Name

iii. Gender

iv. Age

v. Nationality

vi. How often do you play computer games: hours per week?

vii. Did you ever talk to a computer before?

viii. How well do you know HCA?

**Rough scale:** 1: top, 2: well, 3: poorly.

ix. Was it easy or difficult to use the system? Why?

**Rough scale:** 1: easy, 2: somewhat difficult, 3: difficult.

**Examples:** Score 2: Somewhat difficult to formulate oneself in a way so that HCA understands it (7). Score 3: Difficult to understand the English. Navigation was difficult (9).

x. What do you think of HCA?

**Rough scale:** 1: fine, 2: qualifications, 3: no good.

**Example:** Score 1: He looks much the way I think he used to look (authentic) (5).

xi. Could you understand what he said?

**Rough scale:** 1: yes, 2: qualifications, 3: difficult.

**Examples:** Score 2: Yes, but the voice seemed to break from time to time (2). Score 2: Most of it. Sometimes he talked a bit fast (12).

xii. How did it feel to *talk* to HCA?

**Rough scale:** 1: fun, natural, OK, 2: qualifications, 3: negative.

**Examples:** Score 1: Different. Somewhat strange to talk to a computer, but fun (2). Score 2: Strange, unusual. Managed to get used to it to some extent (10).

xiii. Could he follow what you wanted to talk to him about?

**Rough scale:** 1: yes, 2: qualifications, 3: not really.

**Example:** Score 2: Sometimes and sometimes not. For instance I had to ask him three times before he told me his age (7).

xiv. What do you think of his behaviour on the screen?

**Rough scale:** 1: fine, fun, realistic, 2: qualifications, 3: negative.

**Example:** Score 2: Great, but a couple of errors (walk on ceiling and in furniture) (6).

xv. 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?

Rough scale: 1: fine, 2: qualifications, 3: negative.

c. Would you like to be able to do more with gesture? If yes, what?

Rough scale: 1: yes, 2: don’t know, 3: no.

xvi. Was it fun to talk to HCA? If yes, what was fun? If no, can you imagine what could make it fun?

**Rough scale:** 1: yes, 2: sometimes, 3: no.

**Examples:** Score 1: Yes, it was entertaining to learn more about HCA. I like to learn about persons. This is fascinating (5). Score 2: There was too much of the same. Basically what worked was the area around HCA’s desk. It would be nice if there had been more things one could get a story about and maybe also if there were cross-reference between things (2).

xvii. What did you learn from talking to with HCA?

**Rough scale:** 1: a lot, 2: some things, 3: don’t know/nothing, not even memory-refreshing reminders.

**Examples:** Score 2: Learned something about HCA and his family (4). Score 2: English. Fairytale names in English. Knowledge about HCA as a person (15).

xviii. What was bad about your interaction with HCA?

**Rough scale:** 1: nothing/don't know, 2: some things, 3: a lot.

**Examples:** Score 2: That he walks into the graphics (10). Score 2: It was annoying that he often talked about something different than what one asked for (11). Score 3: Not being able to understand him (9).

xix. What was good about your interaction with HCA?

**Rough scale:** 1: a lot, 2: some things, 3: nothing/don't know.

**Examples:** Score 2: Very nice to talk to him when he understood the input (16). Score 2: The windows work in 3D (10).

xx. What do you think we should make better?

**Rough scale:** 1: minor things, 2: substantial improvements, 3: most or all of it.

**Examples:** Score 2: HCA should be able to understand more. The graphics should be corrected so that HCA does not walk on the ceiling or in his furniture (6). Score 3: Have him speak Danish. (9).

xxi. How interested would you be in playing computer games with speech and gesture?

**Rough scale:** 1: very, 2: maybe, 3: not interested.

**Examples:** Score 1: Very, both wrt. touch and speech. Would make games more interesting (18). Score 2: It was fine with HCA but otherwise I don't know (12).

Compared to the original list of interview questions above, we have left out the following two questions from Table 4.1:

1. (If relevant) Which computer games do you like (types of game or concrete games)?
2. Any other comments?

The first of these questions was asked to get an impression of the type(s) of game the users knew about or were familiar with. The games mentioned spanned a broad range of computer games available today, frequent game players typically mentioning the most games (including playstation II games).

The "other comments" is quite a varied group. It includes the following comments (numbers refer to subject numbers in Table 4.1): It is nice that one can move around in his study. (1). I only got proper answers in the works area (desk). Other input typically did not generate an answer. It was entertaining that he knew something about Frederik and Mary and about the statue of the Little Mermaid. (2). Facial expressions were fine. (3). It would be good if HCA reacted when one asks him to stop [authors' comment: request for barge-in]. HCA's life story should be told up front. It helps to create a context and makes it easier to understand the pictures. It would be desirable to have more things to point to with creative stories attached which could even be a bit surprising. (4). Good that the furniture is old. (6). I was in doubt about the English titles of HCA's fairytales. However, I recognised them when he told about the fairytales. I would like to be able to point to more things and get a story. HCA should have a larger vocabulary. (7). I tried to get to the fairytale world by clicking on his hat. He told about the fairytale world when clicking on his hat. It is a good way in which to make a game. (11). I would like to be able to get more information about, e.g., Napoleon and HCA's father. It was a bit strange that when I asked about his preferred game he started to talk to me about his family. (12). It would be nice also to be able to enter the fairytale world. It may quite soon become a bit boring in his office. (13). Might be used for learning English. Liked it. Well done. (18).

| i | ii | iii | iv | v | vi | vii | viii | ix | x | xi | xii | xiii | xiv | xv-a | xv-b | xv-c | xvi | xvii | xviii | xix | xx | xxi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Alexandra | Girl | 12 | Dane | 21 | Yes | 2 | 2 | 2 | 2 | 1 | 2 | 1 | Mouse | 2 | 3 | 2 | 1 | 2 | 2 | 2 | 2 |
| 2 | Nojin | Girl | 17 | Dane | 1 | No | 2 | 1 | 2 | 2 | 1 | 2 | 3 | Mouse | 1 | 3 | 2 | 2 | 2 | 1 | 2 | 2 |
| 3 | Rikke | Girl | 17 | Dane | 1 | No | 2 | 1 | 1 | 2 | 2 | 2 | 1 | Both | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 2 |
| 4 | Simon | Boy | 15 | Dane | 0 | No | 3 | 1 | 1 | 2 | 2 | 2 | 2 | Touch | 2 | 3 | 3 | 2 | 2 | 1 | 1 | 3 |
| 5 | Stefan | Boy | 15 | Dane | 23 | No | 2 | 1 | 1 | 2 | 1 | 2 | 2 | Mouse | 2 | 3 | 1 | 2 | 1 | 1 | 1 | 2 |
| 6 | Sissel | Girl | 12 | Dane | 0 | No | 2 | 2 | 1 | 2 | 1 | 2 | 2 | Touch | 2 | 3 | 1 | 2 | 2 | 1 | 2 | 2 |
| 7 | Bettina | Girl | 15 | Dane | 1,5 | No | 2 | 2 | 1 | 1 | 1 | 2 | 1 | Touch | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 |
| 8 | Paul | Boy | 18 | Scot | 20 | No | 2 | 1 | 1 | 2 | 2 | 1 | 1 | Touch | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 |
| 9 | Camilla | Girl | 12 | Dane | 0 | No | 2 | 3 | 3 | 3 | 3 | 3 | 1 | Mouse | 1 | 2 | 3 | 3 | 3 | 3 | 3 | 3 |
| 10 | Simone | Girl | 14 | Dane | 7 | No | 2 | 1 | 2 | 2 | 1 | 2 | 2 | N/A | N/A | 2 | 1 | 2 | 2 | 2 | 2 | 1 |
| 11 | Mads | Boy | 13 | Dane | 2,5 | No | 2 | 2 | 1 | 2 | 1 | 2 | 1 | Touch | 2 | 3 | 1 | 2 | 2 | 1 | 1 | 1 |
| 12 | Christoffer | Boy | 11 | Dane | 24,5 | No | 3 | 2 | 1 | 2 | 1 | 2 | 1 | Touch | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 2 |
| 13 | Anders | Boy | 15 | Dane | 7 | No | 2 | 2 | 1 | 2 | 2 | 2 | 1 | Touch | 2 | 1 | 2 | 2 | 3 | 2 | 1 | 2 |
| 14 | Tanja | Girl | 17 | Dane | 1 | Yes | 2 | 2 | 1 | 2 | 2 | 2 | 2 | Mouse | 1 | 3 | 1 | 3 | 2 | 2 | 1 | 2 |
| 15 | Mathias | Boy | 13 | Dane | 7 | No | 2 | 1 | 1 | 1 | 1 | 2 | 1 | Mouse | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 |
| 16 | Søren | Boy | 10 | Dane | 9 | No | 3 | 3 | 3 | 1 | 1 | 2 | 2 | Mouse | 1 | 2 | 3 | 3 | 2 | 2 | 3 | 2 |
| 17 | Emil | Boy | 14 | Dane | 7 | No | 2 | 1 | 1 | 1 | 1 | 3 | 2 | Mouse | N/A | 3 | 1 | 2 | 2 | 2 | 2 | 1 |
| 18 | Camilla | Girl | 17 | Dane | 3 | Yes | 2 | 2 | 1 | 1 | 1 | 1 | 1 | Touch | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 |

**Table 4.1.** Overview of the interview data.

Quite obviously, the process of converting users' oral comments into scores is fraught with problems which conspire to reducing the scores obtained to subjective estimates signed by whoever allocated the scores. Still, the scores we have allocated and which are shown in Table 4.1 enable a first, coarse overview of the subjects' opinions on the system.

Basically, the scores have been designed in a way such that (i) the score '1' reflects the top and '3' reflects the bottom, or most negative, judgment, such as that the system must be completely re-designed to be of any use or that its prospects are dull indeed. Score '2' reflects the middle ground, i.e. users who are not completely excited by some property or prospect of the system, on the one hand, and are not completely critical of that property of prospect, on the other. Ignoring property viii on the users' knowledge about HCA, we obtain an average score of 1.72, reflecting 99 x 1, 121 x 2, and 30 x 3.

Defined as explained above, the average, researcher-judgmental score of 1.72 would seem to carry some promise, reflecting, for instance, that relatively few user judgments issue in a bottom score of '3'. Moreover, if we look into the 30 bottom '3's, we find the following. 11 of the '3's were produced by subject 9, Camilla, a 12 years old Danish girl, who turned out to be incapable of understanding virtually anything HCA said and who only managed to input her name and age. For obvious reasons, she was unlikely and mostly unwilling, although humorously so, to pass any positive judgment on the system. Remarkably, for the system functionality she actually did manage to use, i.e. the gesture input, she passed more positive judgment, cf., columns xv-b and xv-c.

The other very critical user is Søren, a 10 years old Danish boy with substantial knowledge of English due to a recent 18-months stay in Washington, DC. Søren passed six ‘3’s and obviously judged the system as one intended for multi-hour home use just like the computer games he is familiar with. In our view, Søren is quite right in pointing out that the system, as it stands, is not suitable for multi-hour use because it still lacks the necessary richness of contents for this purpose.

The system’s intended setting of use is in museums and other public locations in which each user will have limited use time as compared with a private use setting in which it may take the user at least 30 hours to fully explore a game. Obviously, the difference between an intended use for, say, 15 minutes per user, and 30 hours (15 minutes x 120) is very significant. The reason why we did not, rightly or wrongly, tell the young subjects about the intended use setting for the present system prototype version, was that we did not believe that the target users would be interested in, or perhaps even capable of, taking such abstract requirements into account when judging the system’s capabilities and prospects. The consequence of that user instruction decision is that we do not know the extent to which the subjects have judged the system from the point of view of the average 30-hour home use. Our suspicion is that many, if not most, users have judged the system from this point of view. Given this assumption, it does seem warranted to conclude that the system was judged rather favourably overall as shown in Table 4.1.

Interestingly, having already accounted for 11+6=17 of the 30 bottom ‘3’ scores, 8 of the remaining 30 ‘3’s appear in column xv-c in Table 4.1. The question here was whether the subject would like to be able to do more with gesture input. Clearly, a substantial fraction of the subjects were against pursuing the option of offering more 2D gesture functionality in the next system version.

In spite of the qualitative, judgmental, and, arguably, sometimes subjective nature of the scores allocated to the subjects’ interview replies, it may be interesting to note that close to 84% of the ‘3’s allocated by the subjects are accounted for by reference to (9) Camilla’s and Søren’s replies as well as by the subjects’ replies concerning the desirability of adding more gesture functionality to the present system.

Overall, we are actually stricken by the positive reception by the users of the first NICE HCA prototype, as evidenced by the 99 (= 39.6%) ‘1’ scores collected. Reinforcing this impression is the fact that, apart from a single native-English user (user 8 in Table 4.1), all subjects have English as a second language and are actively learning English at school. Conceivably, all or most of these users might have found the system too difficult to deal with, simply because of the language skills it requires. Table 4.1 strongly suggests that this is not the case.

Viewed in a 10-year futuristic perspective, the NICE HCA first prototype system would, no doubt, be considered obsolete. At present, however, there do not seem to be any competing systems around. The positive reception of the system by its test target users constitutes a major encouragement for us to develop the system further.

# 5 Evaluation through in-depth analysis of user-HCA conversations

So far, all 18 conversations (9 users, two conditions) recorded with users who had access to making gesture input with the touch screen, have been annotated using MS Excel. The annotation purposes pursued so far include:

1. basic data gathering on the user test conversations;
2. interaction problems identification and analysis in order to (i) aid diagnostic evaluation and debugging and (ii) prepare the Conversation Management design specification for PT2;
3. theory-based, iterative coding scheme development and application in order to identify and appropriately code phenomena relevant to quantifying the extent to which the user test conversations conform to the theory of conversation underlying the NICE HCA system;
4. development of appropriate metrics for measuring quantifying the extent to which the user test conversations conform to the theory of conversation underlying the NICE HCA system;
5. iterative development of a coding scheme and quantitative metrics for measuring conversation success, one of the research “holy grails” of the NICE project.

The iterative development of coding schemes and metrics mentioned above has been necessary because there are no appropriate coding schemes or metrics available in the literature for the NICE purposes described above. In the following, we briefly describe the, still largely unpublished, work done.

## 5.1 Basic conversation data

The 2x9 conversations between 9 users and HCA comprise 1296 turns half of which were made by the users and half of which were made by HCA. The higher female turn average of 81 turns per conversation compared to the male turn average of 65 suggests an effect of the higher English second-language skills to be expected of the female users, cf. Section 1. 82.3% of the user input turns were spoken and 17.7% were gestural.

## 5.2 Interaction problems

Of the 1296 turns in the corpus, 172 turns, or 13.3%, were affected by system errors (bugs) and wizard input errors (cf. Table 3.1). Our analysis has squarely focused on separating system errors *in the NISLab components* (natural language understanding, conversation management, response generation), from all other interaction problems. All other interaction problems have been classified as NISLab design errors to be addressed in re-designing for PT2. We are aware that this approach is a “maximum” one in the sense that a non-NISLab-component-bug is not necessarily a *NISLab component design error*. In particular, some of the interaction problems which are not due to NISLab component bugs might be due to gesture input processing bugs or gesture component design errors rather than to NISLab component design errors. We have not studied the extent to which this is the case, leaving the analysis of gesture interaction with NICE partner LIMSI, cf. Part 2 of this report. This implies that the conversation success results presented below (Section 5.4) are “minimal” ones, i.e., they might turn out to be higher if one removed from the metrics applied the interaction problems affected by gesture input processing bugs and gesture input processing design errors. This has only been done to a limited extent, cf. the next paragraph. The “minimal” interpretation of the conversation success measures made and presented in Section 5.5 also applies to the effects of the parsing errors made by the NISLab natural understanding component. Due to our time schedule, the parsing errors were only quantified after the analysis reported here. Thus, it is possible that the 16% parsing error rate reported in Table 3.1 includes parsing errors which have been classified as conversation design errors in the interaction problems analysis presented here.

Analysis showed that most of the system errors affecting HCA-user spoken interaction, i.e. 7.9%, were due to loops where HCA continued to produce the same output independently of what the user said. Following diagnosis, the loops themselves turned out to have a range of different causes. These have all been fixed. Another 3.1% of the errors were due to a single bug in the Life/Games mini-dialogue which has now been fixed. Of the remaining, approx. 2.3%, errors, 1.4% were due to the intriguing gesture/speech timing behaviour of a single user. Arguably, this problem is not a system error (bug) but a sophisticated design error in the gesture input processing components.

With the qualifications noted above, interaction problems analysis revealed a number of different types of interaction failures made by HCA, most of which, at least, can be attributed to conversation design errors, or, otherwise expressed, to the failure of our current PT1 conversation design implementation to adequately enable free, domain-oriented spoken conversation. For instance, HCA sometimes fails to reply to a user question, replies irrelevantly, or asks two questions in a single turn. The design error typology achieved through conversation analysis is obviously of paramount importance to our conversation management re-design for the NICE HCA PT2 system.

## 5.3 Evaluating our theory of conversation

The theory of conversation underlying HCA's spoken conversational behaviour is described in [Bernsen and Dybkjær 2003], [Bernsen and Dybkjær 2004], and [Bernsen et al. 2003]. The user tests have given us no reason to revise this, admittedly rather general, theory. On the contrary, the user interviews provided strong confirmation of some of the main tenets of the theory, such as the edutainment value of HCA's story-telling, the rhapsodic nature of non-task-oriented conversation, and the key importance of conversational coherence. Rather, the user test analysis, in particular, the typology of HCA's spoken interaction errors, has provided important clues to how the theory should be enhanced and rendered more concrete and specific.

Earlier work on the WoZ2 corpus, cf. NICE Deliverable D2.2a and [Bernsen et al. 2004d], had generated metrics for measuring a number of conversation properties relevant to our theory of conversation, such as domain symmetry, domain initiative symmetry, and conversation drive symmetry. The latter metrics has been applied to the user test corpus discussed here, showing a drive symmetry average of 0.62. Drive symmetry compares the extent to which the user and HCA, respectively, contribute to driving the conversation forward by asking questions and volunteering information for the interlocutor to react to. AS expected, the drive symmetry ratio was significantly higher in the second user test condition in which the users had to induce HCA to provide particular pieces of information or to react in specified ways, making the user the "hard-driving" interlocutor. Also, confirming the WoZ2 findings [Bernsen et al. 2004d], we found very significant individual user differences when measuring drive symmetry.

## 5.4 Conversation success

Nobody knows how to measure the success of domain-oriented (non-task-oriented) conversation. In analysing the user test data, a first, tentative metrics for measuring conversation success has been established, based on the typology arrived at of HCA's conversational interaction errors. To compute conversation success, we first subtracted the known system bug-affected turns, i.e. 13.3% of the turns, cf. Section 5.2 above, from the corpus, arriving at an 86.7% corpus fraction which, barring gesture input processing bugs and design errors as well as natural language understanding bugs and design errors, could be subjected to quantification as regards conversation design errors. We then measured the percentage of identified conversation design errors and subtracted those from 100% conversation success. To our surprise, we found an overall conversation success average in the 18 user test conversations of 73.5%. This, frankly, is way above what was expected of the NICE HCA PT1 system. More explicitly, this average score means that:

- *given what we currently know about conversational coherence and interaction problems in conversation* - because we have tried to take all of it into account in corpus annotation and metrics;
- *adding the uncertainty as to the cumulative contribution to speech/gesture interaction problems of gesture input processing bugs and design errors as well as natural language understanding bugs and design errors;*
- *only subtracting the effects of identified conversation management and processing bugs from the corpus studied;*
- *HCA managed to conduct flawless free, domain-oriented conversation in an average of 73.5% of the conversational turns.*

Obviously, we consider it mandatory to continue to sharpen our notion of conversational success in order to try to identify additional interaction problems which should detract from conversation success as measured with respect to NICE HCA PT1. Still, the preliminary results reported here do seem to demonstrate that systems which conduct free domain-oriented conversation about particular domains of discourse are not beyond the reach of advanced state-of-the-art spoken and gesture conversation management. Unless we have missed some really essential factors contributing to conversation success, we consider this result a surprisingly positive preliminary result of the NICE project.

# 6 Conclusion and future work

Part 1 of this report has presented the knowledge we now possess concerning the virtues and shortcomings of the first NICE HCA prototype. Five general conclusions are that:

- the NICE HCA PT1 is on or ahead of schedule apart from input fusion;
- the NICE HCA PT1 has in general performed acceptably in the user test;
- the users were quite enthusiastic as regards the potential of including spoken conversation in future computer games;
- the system platform, including the broker, cf. Figure 2.1, performs as expected or even better than that;
- NISLab's HCA modules are on the right track.

These conclusions imply that we have a good basis for the second phase of NICE HCA system development. We are in a position to specify and implement NICE HCA PT2 without having to engage in costly and cumbersome basic re-design, focusing instead on the many challenges which have either been evident from the start of the project or discovered in the user test and other exposures of the system specification to user interaction (see NICE deliverable D2.2a).

Since the user tests conducted in January 2004, we have been working along the following lines:

- analysing the user test results as described in this report;
- performing diagnostic testing and debugging of the HCA PT1 system version used in the user test. At the time of writing, this phase has been completed, having led to the removal of all identified system errors except for those to do with natural language understanding module inadequacies. These will be removed in PT2;
- developing non-communicative action and communicative functions for PT1 to be demonstrated at the NICE project review;
- incorporating the Scansoft speech recogniser into the system in order to get hands-on experience with the speech recogniser before completing the NICE HCA PT2 requirements and design specification; and
- carrying out PT2 requirements analysis, PT2 requirements specification, PT2 design analysis, and PT2 design specification based on the WoZ1, WoZ2, and user test data analyses. The results will be reported in the second version of NICE Deliverable D1.1: *Requirements and design specification for domain information, personality information and dialogue behaviour for the second prototype.*

Following the steps just mentioned, we will proceed through rapid prototyping, addressing a single HCA conversational domain at-a-time and starting with the fairytales domain, towards implementing, as soon as possible, a first NICE HCA PT2 version which can be tested with particular emphasis on speech recogniser performance, meta-communication handling, improved natural language understanding, and improved communication management. The NICE HCA PT1 response generation has not been prominent above because, so far, it works as expected.

# 7 References

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

[Bernsen et al. 2002] Niels Ole Bernsen, Johan Boye, Svend Kiilerich and Ulrich Lindahl: Requirements and Design Specification for Domain Information, Personality Information and Dialogue Behaviour for the First NICE Prototype. *NICE Deliverable D1.1.* NISLab, Denmark: September 2002. 30 pages.

[Bernsen and Dybkjær 2003] Niels Ole Bernsen and Laila Dybkjær: First Prototype Version of Conversation Management and Response Planning for H.C. Andersen. *NICE Deliverable D5.1a.* NISLab, Denmark, October 2003. 20 pages.

[Bernsen and Dybkjær 2004] Niels Ole Bernsen and Laila Dybkjær: Domain-Oriented Conversation with H.C. Andersen. To appear in *Proceedings of the Workshop on Affective Dialogue Systems,* Kloster Irsee, Germany, June 2004.

[Bernsen et al. 2004a] Niels Ole Bernsen, Marcela Charfuelàn, Andrea Corradini, Laila Dybkjær, Thomas Hansen, Svend Kiilerich, Mykola Kolodnytsky, Dmytro Kupkin and Manish Mehta: Conversational H.C. Andersen. First Prototype Description. To appear in *Proceedings of the Workshop on Affective Dialogue Systems,* Kloster Irsee, Germany, June 2004.

[Bernsen et al. 2004b] Niels Ole Bernsen, Marcela Charfuelàn, Andrea Corradini, Laila Dybkjær, Thomas Hansen, Svend Kiilerich, Mykola Kolodnytsky, Dmytro Kupkin and Manish Mehta: First Prototype of Conversational H.C. Andersen. To appear in *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI 2004)*, Gallipoli, Italy, May 2004.

[Bernsen et al. 2004c] Bernsen, N. O., Dybkjær, L., and Kiilerich, S.: NISLab's Collection and Analysis of Multimodal Speech and Gesture Data in an Edutainment Application. *NICE Deliverable D2.2a.* NISLab, Denmark, April 2004. 30 pages.

[Bernsen et al. 2004d] Bernsen, N. O., Dybkjær, L. and Kiilerich, S.: Evaluating Conversation with Hans Christian Andersen. To appear in *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC'2004), Lisbon, Portugal, May 2004.

[Bernsen et al. 2003] Niels Ole Bernsen, Laila Dybkjær and Mykola Kolodnytsky: Analysis and Representation of Domain Information, Personality Information and Conversation Behaviour for H.C. Andersen in the First Prototype. *NICE Deliverable D1.2a.* NISLab, Denmark, October 2003. 26 pages.

[Corradini et al. 2004] Andrea Corradini, Morgan Fredriksson, Manish Mehta, Jurgen Königsmann, Niels Ole Bernsen, and Lasse Johannesson: Towards Believable Behavior Generation for Embodied Conversational Agents. To appear in *Proceedings of the Workshop on Interactive Visualisation and Interaction Technologies*, *IV&IT 2004*, Krakow, Poland, June 2004, in conjunction with the International Conference on Computational Science 2004 (ICCS 2004).

[Dybkjær et al. 2003] Laila Dybkjær, Niels Ole Bernsen, Reinhard Blasig, Stéphanie Buisine, Morgan Fredriksson, Joakim Gustafson, Jean-Claude Martin, Mats Wirén: Evaluation Criteria and Evaluation Plan. *NICE Deliverable D7.1.* NISLab, Denmark, March 2003. 28 pages.

# Part 2, Evaluation of Gesture Input Processing, LIMSI-CNRS

## 8 LIMSI-CNRS contribution to PT1 evaluation

### 8.1 Introduction

No PT1 user tests were conducted at LIMSI because the gestural and multimodal modules we developed are not easily testable *per se*, and out-of-context evaluations may not be sufficient. Since these modules are integrated in both HC-Andersen and Fairy-Tale-World versions of the first NICE prototype, LIMSI will take advantage of user tests carried out at NIS-Lab and Telia sites to evaluate Gesture Recognition, Gesture Interpretation and Input Fusion.

### 8.2 User tests at NISLab

User tests were conducted at NIS-Lab, January 20-22. 18 users (9 boys, 9 girls; 10 to 18 years old) tried HCA version of PT1. Half of them used the system with a tactile screen as input device, and half with a mouse.

#### 8.2.1 Description of GR, GI and IF modules

In the version tested, gestures that could be recognized by GR were: points, circles, horizontal lines, vertical lines, and diagonal lines. The gesture trace was always displayed on the screen, whatever the input device.

According to the output of GR and to the object tracker, GI sends (or not) to IF the object selected. If the user gestures to a non-referenceable object, no GI frame is sent to IF module. In this version, the IF module used only temporal proximity as a criteria for merging at most one NLU frame and one GI frame.

#### 8.2.2 Log files analysis

LIMSI collected log files from gestural and multimodal modules for 15 users (due to technical problems, we could not process log files from the 3 remaining users). For each user, 3 files were logged (1 by the GR, 1 by the GI and 1 by the IF), which resulted in 45 raw log files.

These logged data include the messages sent by each one of the three modules and the vector of coordinates for the gesture. These log files are not enough to evaluate the modules. They will have to be compared to the videos of user's behaviour in order to check if the modules worked properly. If they did not, the videos will also clarify why.

Figure 1 describes the analysis process of log files, parsed by a log analyser and submitted to analysis of variance.

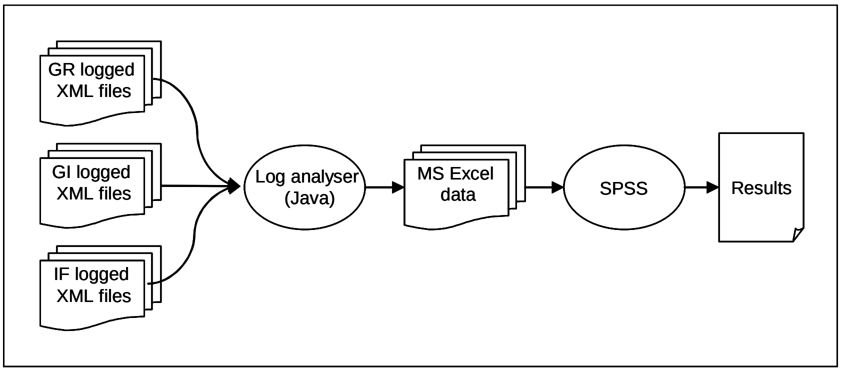

**Figure 1.** The analysis process of log files.

| Date & time | Gender | Input device | Number of GR frames | Number of IF frames | Number of NLU-only IF frames | Number of GI-only IF frames | Number of multimodal IF frames | Number of deictic | Number of ref to objects in NLU |
|---|---|---|---|---|---|---|---|---|---|
| 21/01 4PM | Fem | Mouse | 53 | 32 | 3 | 29 | 0 | 0 | 0 |
| 21/01 5PM | Fem | Mouse | 0 | 80 | 80 | 0 | 0 | 8 | 4 |
| 22/01 5PM | Fem | Mouse | 29 | 44 | 32 | 12 | 0 | 0 | 0 |
| 21/01 4PM | Fem | Tactile | 42 | 85 | 73 | 12 | 0 | 3 | 8 |
| 21/01 5PM | Fem | Tactile | 15 | 83 | 76 | 7 | 0 | 5 | 0 |
| 22/01 5PM | Fem | Tactile | 14 | 62 | 50 | 12 | 0 | 2 | 0 |
| 21/01 3PM | Male | Mouse | 33 | 78 | 69 | 8 | 1 | 12 | 3 |
| 22/01 2PM | Male | Mouse | 73 | 81 | 61 | 20 | 0 | 10 | 2 |
| 22/01 3PM | Male | Mouse | 49 | 125 | 99 | 23 | 3 | 14 | 1 |
| 22/01 4PM | Male | Mouse | 76 | 77 | 70 | 7 | 0 | 6 | 2 |
| 21/01 2PM | Male | Tactile | 61 | 104 | 81 | 23 | 0 | 5 | 3 |
| 21/01 3PM | Male | Tactile | 42 | 85 | 73 | 12 | 0 | 3 | 8 |
| 22/01 2PM | Male | Tactile | 33 | 91 | 77 | 12 | 2 | 21 | 0 |
| 22/01 3PM | Male | Tactile | 22 | 54 | 46 | 8 | 0 | 3 | 0 |
| 22/01 4PM | Male | Tactile | 16 | 48 | 44 | 3 | 1 | 0 | 0 |
| **TOTAL** | | | **558** | **1129** | **934** | **188** | **7** | **92** | **31** |

**Table 1.** Data extracted from GR and IF log files.

Table 1 presents the first data extracted. One striking result is the difference between the number of GR frames (approximately one every minute in users' scenarios) and the number of IF frames containing gesture (see Figure 2). Indeed, 61% of GR frames were not further processed by IF. At this stage of the analysis, we may assume that the users often gestured to non-referenceable objects. This assumption might be confirmed by the videos. In post-test interviews, 5 users (3 girls, 2 boys) mentioned that they would like more referenceable objects.

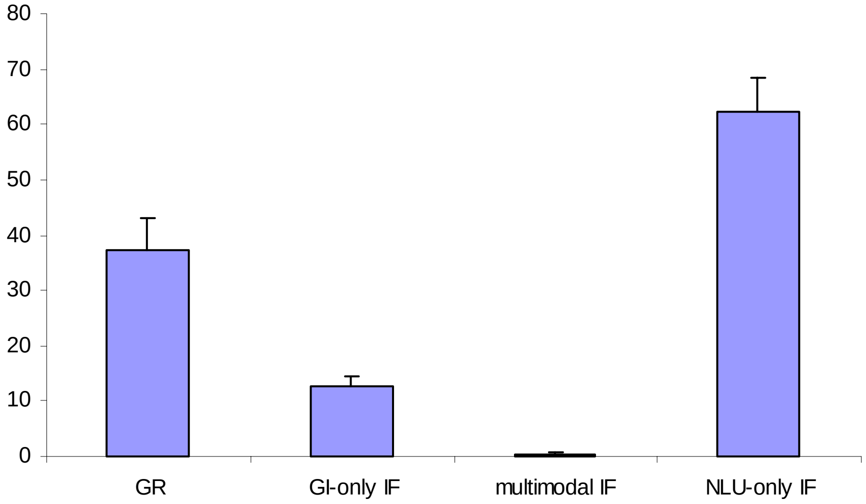

**Figure 2.** Average number of frames and standard error for this sample of 15 users.

We also observe a gender difference in this percentage of “non-processed” gestures (F(1/10) = 3.9, p = .077) and in the initial number of GR frames (F(1/11) = 3.8, p = .076), showing that boys tended to be particularly eager to use gesture.

We may notice that the number of multimodal IF frames is dramatically low. Among IF frames, 82.7% contained only NLU output, 16.7% contained only GI output, and 0.6% were multimodal (see Figure 2). Besides, in these multimodal frames, the two modalities were used to address simultaneously two unrelated topics (e.g. talk about the ugly duckling and gesture towards the picture of Jenny Lind). The seven multimodal frames that were collected are fully transcribed in Annex 1. For future development, we will have to find a way to manage such behaviours.

An important issue for gesture modules is the influence of the input device (mouse vs. tactile screen) on the gestural behaviour. The first logged data do not show such an influence.

We also collected in log files a set of data on the shapes of movements produced as 1st best results by the GR module. Table 2 shows that 44% of gestures were recognized as pointing, 33% as circles, and 23% as lines. Neither the input device nor the gender of users had any influence on this pattern. See Figure 3 for a comparison of shapes of gestures as a function of the input device.

| **Date & time** | **Gender** | **Input device** | **Nb of points** | **Nb of circles** | **Nb of vertical lines** | **Nb of horizontal lines** | **Nb of diagonal-up lines** | **Nb of diagonal-down lines** | **Nb of other shapes** | **TOTAL** |
|---|---|---|---|---|---|---|---|---|---|---|
| 21/01 4PM | Fem | Mouse | 7 | 29 | 12 | 2 | 2 | 1 | 0 | **53** |
| 21/01 5PM | Fem | Mouse | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| 22/01 5PM | Fem | Mouse | 4 | 10 | 7 | 3 | 1 | 4 | 0 | **29** |
| 21/01 4PM | Fem | Tactile | 12 | 17 | 8 | 2 | 2 | 1 | 0 | **42** |
| 21/01 5PM | Fem | Tactile | 6 | 5 | 3 | 0 | 0 | 1 | 0 | **15** |
| 22/01 5PM | Fem | Tactile | 0 | 2 | 7 | 1 | 4 | 0 | 0 | **14** |
| 21/01 3PM | Male | Mouse | 10 | 8 | 10 | 1 | 0 | 4 | 0 | **33** |
| 22/01 2PM | Male | Mouse | 68 | 0 | 4 | 0 | 0 | 1 | 0 | **73** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 22/01 3PM | Male | Mouse | 6 | 34 | 6 | 3 | 0 | 0 | 0 | **49** |
| 22/01 4PM | Male | Mouse | 57 | 4 | 7 | 5 | 2 | 1 | 0 | **76** |
| 21/01 2PM | Male | Tactile | 2 | 52 | 4 | 1 | 1 | 1 | 0 | **61** |
| 21/01 3PM | Male | Tactile | 12 | 17 | 8 | 2 | 2 | 1 | 0 | **42** |
| 22/01 2PM | Male | Tactile | 25 | 3 | 1 | 1 | 2 | 1 | 0 | **33** |
| 22/01 3PM | Male | Tactile | 20 | 2 | 0 | 0 | 0 | 0 | 0 | **22** |
| 22/01 4PM | Male | Tactile | 15 | 0 | 1 | 0 | 0 | 0 | 0 | **16** |
| **TOTAL** | | | **244** | **183** | **78** | **21** | **16** | **16** | **0** | **558** |
| **%** | | | **44** | **33** | **14** | **4** | **3** | **3** | **0** | |

**Table 2.** Shapes of movements extracted from GR log files.

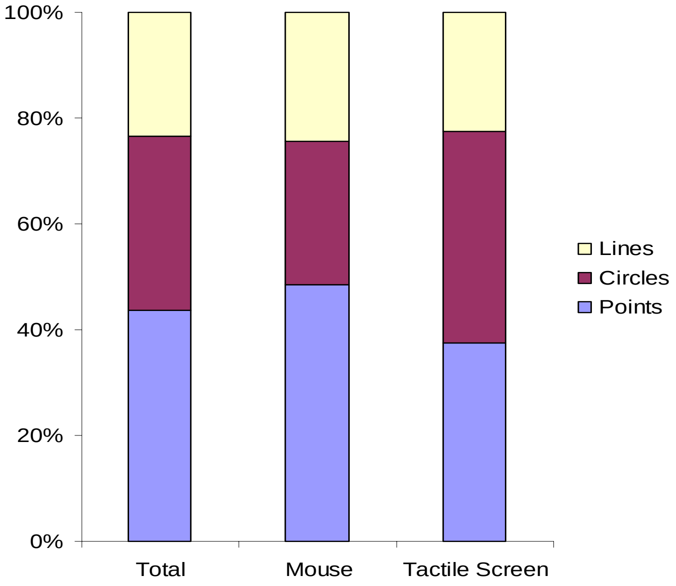

**Figure 3.** Percentage of lines, circles and points in the total corpus (left), with the use of mouse (centre) and with the use of the tactile screen (right). No statistical differences arose from these data.

### 8.2.3 Future directions

#### *8.2.3.1 Log files*

Further analysis of log files will be performed on the data already extracted, and also on forthcoming extractions. In particular, we intend to study the objects targeted in relation to the modality used and the gestures observed. We will also compare the shape recognised by GR to the logged vector of coordinates.

#### *8.2.3.2 Video*

The analysis of log files will greatly benefit from being connected to videos of the users. For half of users, the screen was video-recorded, showing the whole interaction (verbal interactions and gestural input on the touch-screen). LIMSI intends to make a parallel analysis of log files and videos of gestural input, including the study of how each module processed users' behaviours.

#### *8.2.3.3 Interviews*

Finally, gestural and multimodal modules may be improved on the basis of comments made by users during post-test interviews. For example, we may consider that 5 users found the processing of gestures too slow (which suggests to adjust the IF temporal window).

#### *8.2.3.4 Contribution to evaluation of entertainment aspects*

Among the most difficult features to evaluate, the entertaining aspects of PT1 will be mostly assessed from the interviews. However, some behavioural indices may be useful to complement subjective evaluation of users, in particular for children (Hanna et al., 1997).

During user tests at NIS-Lab, facial expressions of users interacting with the system were video-recorded. We assume that they could provide some indications of users' satisfaction and entertainment. LIMSI will take part in the attempt to analyze these facial expressions. We plan to annotate them on a functional rather than a morphological level (Manstead, 1991; Steininger et al., 2002), and try to code them into satisfaction assessment.

If analyses of facial expressions on this first test turn out to be feasible and fruitful, we may re-use the same method for evaluation of the second NICE prototype. We may thus be able to compare the two tests and see whether there is an evolution of entertainment.

## 8.3 User tests at Telia

Log files from Telia user tests will also complement this first analysis. We will work out the same kind of behavioural metrics. In this second corpus, we may observe noticeable differences in gestural and multimodal behaviour of users. Indeed, Telia's scenario was less conversational and rather more object-oriented. The input device (a gyro mouse) may also have influenced the data.

## 8.4 References

Hanna, L., Risden, K., Alexander, K.J.: Guidelines for usability testing with children. Interactions, 4, 1997, 9-14.

Manstead, A.S.R.: Expressiveness as an individual difference. In: R.S. Feldman & B. Rimé. Fundamentals of Nonverbal Behavior, pp. 385-328. Cambridge University Press, 1991.

Steininger, S., Rabold, S., Dioubina, O., Schiel, F.: Development of the User-State conventions for the multimodal corpus in Smartkom. Proceedings of the Workshop on Multimodal Resources and Multimodal Systems Evaluation. LREC'2002, Las Palmas, Canary Islands, Spain, 2002.

## 8.5 Annex 1

This annex details the 7 multimodal IF frames collected during user tests at NIS-Lab.

### 8.5.1 21/01-3PM-Boy-Mouse

```
------
SPEECH: hm i do not know
SEMANTICS: [user_opinion:negative] [verb:know]
GESTURE:pictureJennyLind
```

### 8.5.2 22/01-2PM-Boy-TactileScreen

```
------
SPEECH: oh i remember that now
SEMANTICS: [user_opinion:general] [verb:remember] [diectic:that]
GESTURE:pictureLittleMermaid
------
```

SPEECH: mm
SEMANTICS: [no_semantics]
GESTURE:pictureJennyLind

### 8.5.3 22/01-3PM-Boy-Mouse

------
SPEECH: okay goodbye i will go
SEMANTICS: [user_opinion:positive] [verb:visit] [greeting:ending]
GESTURE:pictureJennyLind
------
SPEECH: can tell me about your dad and your mom
SEMANTICS: [question:general] [user_intent:listen] [family:father] [family:mother]
GESTURE:pictureJonasCollin
------
SPEECH: can you tell me about your dad and your mom and your grandpa
SEMANTICS: [question:yes/no] [hca_old] [family:father] [family:mother]
GESTURE:pictureJonasCollin

### 8.5.4 22/01-4PM-Boy-TactileScreen

------
SPEECH: it is a bout a duck who is not as pretty as the other ones and eh but in the end it becomes much prettier in the other ones
SEMANTICS: [user_opinion:general] [greeting:ending] [fairytale:ugly_duckling] [number:1]
GESTURE:pictureJennyLind