

*NICE project (IST-2001-35293)*



# **Natural Interactive Communication for Edutainment**

## **NICE Deliverable D2.2a**

### **NISLab's Collection and Analysis of Multimodal Speech and Gesture Data in an Edutainment Application**

*15 April 2004*

*Author*

*Niels Ole Bernsen<sup>1</sup>, Laila Dybkjær<sup>1</sup>, Svend Küllerich<sup>1</sup>*

*1: NISLab, Odense, Denmark*

<b>Project ref. no.</b>	IST-2001-35293
<b>Project acronym</b>	NICE
<b>Deliverable status</b>	Public
<b>Contractual date of delivery</b>	29 February 2004+ agreed 6 weeks delay = 15 April 2004
<b>Actual date of delivery</b>	15 April 2004
<b>Deliverable number</b>	D2.2a
<b>Deliverable title</b>	NISLab's Collection and Analysis of Multimodal Speech and Gesture Data in an Edutainment Application
<b>Nature</b>	Report
<b>Status &amp; version</b>	Final
<b>Number of pages</b>	30
<b>WP contributing to the deliverable</b>	WP2
<b>WP / Task responsible</b>	Joakim Gustafson, TeliaSonera
<b>Editor</b>	N/A
<b>Author(s)</b>	Niels Ole Bernsen, Laila Dybkjær, Svend Kiilerich, all NISLab
<b>EC Project Officer</b>	Mats Ljungqvist
<b>Keywords</b>	Wizard of Oz data collection, user test, embodied conversational agent for edutainment, domain-oriented conversation
<b>Abstract (for dissemination)</b>	This report, Deliverable 2.2a from the NICE (Natural Interactive Communication for Edutainment) project, describes three corpus-gathering exercises conducted in order to specify and test the first running prototype of the NICE Hans Christian Andersen system. The system enables children, young people, and everyone else to have edutaining conversation with fairytale author Hans Christian Andersen.

# Table of Contents

<b>1</b>	<b>Introduction .....</b>	<b>1</b>
<b>2</b>	<b>WoZ1 .....</b>	<b>3</b>
2.1	Setup.....	3
2.2	The simulation.....	3
2.3	Conversation example .....	6
2.4	Analysis .....	7
<b>3</b>	<b>WoZ2 .....</b>	<b>8</b>
3.1	Preparations .....	8
3.2	Setup.....	10
3.3	The simulation.....	10
3.4	Conversation example .....	12
3.5	Initial data handling.....	14
3.5.1	Diaries and lessons learned .....	14
3.5.2	Transcription and additional markup .....	14
3.6	Analysis of age, gender, turn length, relative target group success .....	14
3.6.1	Age, gender, turn length .....	15
3.6.2	Relative target group success .....	16
3.7	In-depth analysis .....	16
3.7.1	Domain/topic ranking and expertise balance .....	17
3.7.2	Phase initiative and drive symmetry .....	17
3.7.3	Test-driving the PT2 specification .....	18
3.7.4	Future work on WoZ2 corpus analysis .....	19
<b>4</b>	<b>WoZ3 .....</b>	<b>20</b>
4.1	Setup.....	20
4.2	The simulation.....	21
4.3	Conversation example .....	24
4.4	Analysis, basic data .....	25
4.5	In-depth conversation analysis, interview analysis .....	25
<b>5</b>	<b>References .....</b>	<b>26</b>



# 1 Introduction

This report, NICE Deliverable D2.2a, describes the work done at NISLab to collect and analyse speech and gesture data for the purpose of developing and evaluating the Hans Christian Andersen (HCA) part of the NICE system. Our use of part of the data in evaluating the first HCA prototype (PT1) is described in NICE Deliverable D7.2a.

Due to the rather innovative nature of the HCA system, the NICE Work Programme envisions substantial data collection in the project. NICE HCA data collection serves three main purposes:

1. English speech recogniser training;
2. system requirements and design specification;
3. system prototype evaluation.

As the target users of the system are the 10-18 year olds, special emphasis has been put throughout on collecting data with this target user group. At NISLab, three data collection efforts have been made so far. In the following, these will be called WoZ1, WoZ2, and WoZ3, respectively. “WoZ” refers to the Wizard of Oz systems simulation methodology by which users are made to believe that they are communicating with a real system rather than a more or less simulated one [Bernsen et al. 1998]. A fourth data collection exercise carried out in early 2003 in which video data was collected with a professional actor who acted as HCA, is not described in the present report. The three WoZ simulations described in this report are compared in Table 1.1.

Table 1.1 shows that NISLab has collected what was planned and may be estimated as a sufficient amount of speech data on the NICE HCA application for training the dedicated English Scansoft recogniser for use in future versions of the NICE HCA system.

	<b>WoZ1</b>	<b>WoZ2</b>	<b>WoZ3</b>
<b>Purpose</b>	recogniser training system specification	recogniser training design specification test	recogniser training PT1 evaluation
<b>When collected</b>	1-31 October 2002	21-30 July 2003	20-22 January 2004
<b>Where collected</b>	local schools	HCA Museum, Odense	NISLab
<b>WoZ type</b>	controlled in-field	field	controlled laboratory
<b>Users</b>	7-19 year olds	all ages	10-18 year olds
<b>Number of users</b>	60-70	approx. 500	18
<b>User input</b>	speech	speech	speech and gesture
<b>System output</b>	speech, small animated HCA face, text reminders	speech, modestly animated HCA in his study	speech, 3D embodied HCA
<b>Implemented system parts involved</b>	none	none	all except speech recogniser
<b>Role of wizard(s)</b>	Act as HCA	Act as HCA	Type spoken input
<b>Assistant(s)</b>	One acting as experimenter and technician	Student helper in museum, inviting young visitors to talk to HCA	Experimenter, two typing wizards, technician, observers, two interviewers
<b>Raw data collected</b>	acoustic no logfiles	acoustic no logfiles	acoustic video complete logfiles
<b>Speech data volume</b>	approx. 6 hours	approx. 30 hours	approx. 11 hours
<b>Video data volume</b>	N/A	N/A	approx. 11 hours
<b>No. of conversations</b>	35-40	502	36
<b>No. of user utterances</b>	2.047	6.870	1.206
<b>Raw transcription</b>	yes	yes	yes
<b>Transcription coding</b>	yes	yes	yes, underway
<b>Transcription tool</b>	Transcriber	Transcriber	Transcriber
<b>Semantics annotation</b>	no	yes approx. 50% of data	no
<b>Topic annotation</b>	yes	yes approx. 70% of data	no
<b>User language skill evaluation</b>	yes	yes	yes, underway

**Table 1.1.** Comparison of the three NICE HCA Wizard of Oz experiments.

## 2 WoZ1

### 2.1 Setup

As most of the recordings took place in various schools, the setup had to be easily transportable by two persons (wizard and assistant) and also fairly quick to set up on the spot. A few dry runs were performed before the first recording at a school. The two main purposes were to check how much time it took to set up the equipment and of course to check whether the set up worked satisfactorily.

The WoZ setup is illustrated in Figure 2.1. The users spoke through a microphone and listened to HCA responses via headphones, which allowed recording of the user only. This recording was meant for the acoustic model. Software included Samplitude 6.0 for recording and PC Remote 1.0 for communication with the wizard.

The wizard spoke through a microphone and listened to the user via speakers. To render the voice of the wizard as neutral and computer-like as possible, the interface Sound Blaster Live! was used along with the Sound Blaster card for distortion. The wizard was responsible for controlling the mouth, eye and head movements of a cartoon-like HCA face, which could be seen via a mini-hub connection by the user. A programme called FaceWorks, version 1.0 was used to create the cartoon and control it.

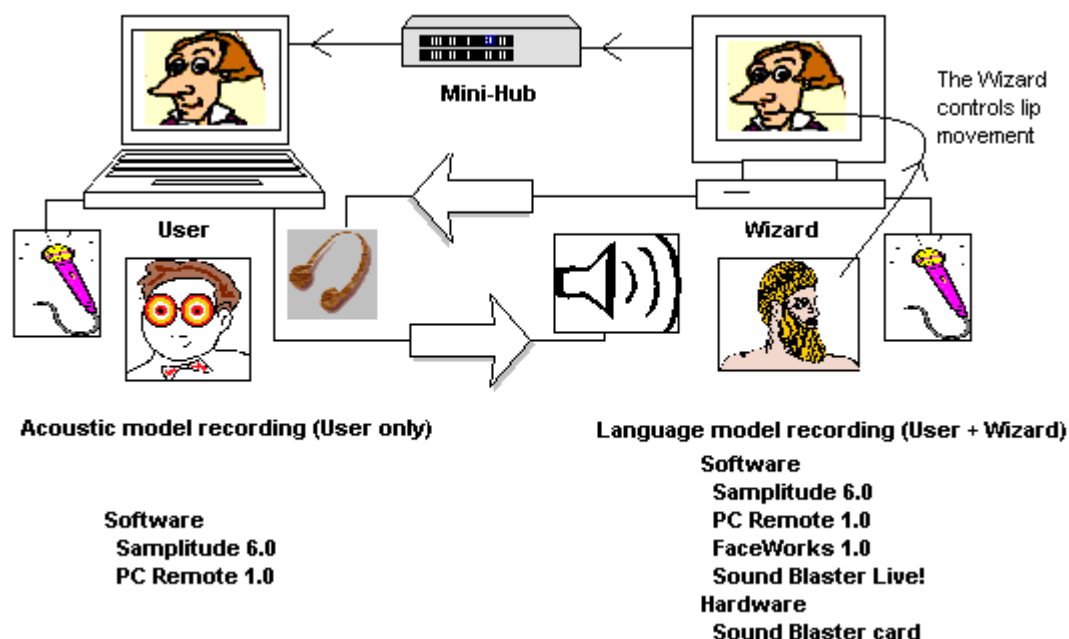


Figure 2.1. The Wizard of Oz set-up.

### 2.2 The simulation

The recordings took place at four local schools in Odense and, in one case, with a single user at NISLab, in October 2002. 2047 user utterances were recorded equalling around six hours with users aged between 7 and 19. The average utterance length was seven words. In total between 60 and 70 users took part in approx. 35-40 dialogues, each lasting between five and 30 minutes. In each dialogue one or two users took part. The reason for the fact that up to two users took part in each dialogue was twofold:

- it took pressure off the shoulders of the users who were not so strong in English;

- it ensured longer dialogues as time was in some cases limited due to the school class schedules.

Most of the computer screen was covered with paper (Figure 2.2). Through a hole in the paper the animated cartoon-like HCA face was visible (Figure 2.3). The topics could be seen below the HCA face.



**Figure 2.2.** The user in front of the graphics.



**Figure 2.3.** The graphics visible to the user.

The wizard (Figure 2.4) could see two versions of the cartoon-like HCA face of which the small one in the upper left-hand corner could be manipulated as regards mouth, eye and head movements (Figure 2.5).

The assistant informed the users on the distance to keep from the microphone. The users were instructed by the assistant that they could talk to HCA about whatever they wanted to. However, if they needed inspiration a list of topics was visible on the computer screen just below the HCA animation. The topics were:

- HCA and his childhood in Odense
- HCA and his youth in Copenhagen
- HCA and the women
- HCA and the travels



- HCA and his peers
- HCA and the inventions at his time
- HCA and Odense at his time
- HCA and Copenhagen at his time
- HCA and Denmark at his time



**Figure 2.4.** The wizard in front of the screen with graphics.



**Figure 2.5.** The graphics visible to the wizard.

The opening questions to the user were about their age, gender and mother tongue. The initiative was given back to the user with the question “what would you like to speak to me about”. The initiative was mostly on the user side. If there were long breaks in the conversation the wizard took initiative and asked, e.g., “what is your favourite fairy tale” and

“why is this your favourite fairy tale”. After another break in the conversation, the wizard asked “do you have more interesting questions to me”. If the user answered no, the wizard would end the conversation with “it was nice talking to you, good bye”.

## 2.3 Conversation example

In the example below, speaker 1 (spk1) is the wizard and speaker 3 (spk3) is the user.

Speaker	Turn
spk3	i'm a girl i am thirteen years old i speak Dane nish and come from Denmark
spk1	hello i'm Hans Christian Andersen i come from Denmark too what would you like to speak to me about
spk3	why do you begin to read histories
spk1	well at first i wanted to be an actor i tried to get into the theatre but they didn't want me so instead i started to write
spk3	where do you live the where do you live that you was little where you was little
spk1	my childhood i spent in Odense i lived in Odense 'till i was fourteen years old then moved to Copenhagen
spk3	where do you live in Copenhagen
spk1	oh i lived different places do you know Copenhagen well
spk3	yes i do
spk1	at first i lived in a very small room without any windows it was very cheap later on when i got famous i lived in Nyhavn do you know Nyhavn
spk3	yes i have been there before
spk1	do you like Copenhagen
spk3	yes
spk1	why do you like Copenhagen
spk3	i don't know there's many things you can do and i don't know why
spk1	when i'm in Copenhagen i like to go around in the city and watch people and faces what do you do when you are in Copenhagen
spk3	i'm i'm out to shop and i i'm together with my family
spk1	have you read any of my fairy tales
spk3	yes very many of them i think i've read the all
spk1	all one hundred and eighty
spk3	i no i don't think so i don't know there was so i don't know there was so many
spk1	oh yes
spk3	

spk1	i wrote a lot of fairy tales
spk3	yes what's of yu what's of this hi what's of your history do you best like
spk1	i think i like the Ugly Duckling the best do you know the Ugly Duckling
spk3	yes that's a s that is a sweet history
spk1	i'm pleased to hear that you think that
spk3	
spk1	why do you think it's a sweet history
spk3	
spk1	okay thank you do you have anybody with you
spk3	yes i have i have a girl
spk1	i would very much like to speak to her
spk3	all right

## 2.4 Analysis

Two students were used as transcribers and for topic tagging. The dialogues for the five recording days were transcribed using the freely available transcription software Transcriber version 1.4. The transcribers were instructed orally and received also two documents, one containing instructions and guidelines for transcription and one with the topics to use. One supervisor was responsible for checking the work of the students and for answering all upcoming questions.

Spell-checking was made after the dialogues had been transcribed and topic tagged by one of the supervisors using Xemacs with ispell from the Linux platform.

The speech recognition partner in the NICE project wanted markup of the speaker's level of English inserted in the annotation files. We established a set of criteria for evaluating the English level of the user. The criteria include English level, accent and fluency. Each of these criteria are given three or four levels. Transcriber version 1.4 was used for inserting the evaluation tagging. The annotation was done by a phonetician from NISLab.

## 3 WoZ2

### 3.1 Preparations

The HCA Museum in Odense kindly allowed us to collect the WoZ2 data at the museum premises during the summer of 2003. Due to ongoing reconstruction work at the museum in preparation for HCA's 200 years anniversary in 2005, the wizards would have to work in the museum's basement and to connect wirelessly with the laptop with HCA on it in the museum itself.

In preparation of the WoZ experiments some support documents were developed and training of the wizards was conducted. The key support document for the wizards was the system output specification, re-structured in a way which allowed the wizard fast access to any specified output during conversation with the users. Having found the output needed in context, the wizard would have to speak it to the user. Since HCA would be visibly present to the user at all times, HCA should move his mouth when speaking. The wizard would have to take care of that as well.

Another document written to the wizards was a guide which emphasises the importance of close adherence to the system specification in order to gather reliable data on it. The guide also stresses the importance of making users believe that they are communicating with a real system despite the important differences between a system and a human wizard. To this end, the wizard's speech behaviour was prescribed in detail, including that wizards were permitted to show happy/sad/angry emotions during conversation. To enable the gathering of some amount of error communication data, the guide proposes that, during each conversation, the wizards should sometimes fail to understand the user and either ask for repetition or clarification, exhibit more or less extravagant misunderstanding of the user's input, or patch up having lost the thread in conversation by producing out-of-domain output.

It was judged that a single wizard at-a-time could do the simulation without the need for an assistant. However, since the simulations would go on all day from 9 AM to 7 PM for ten days in succession, we needed two wizards who would take turns, each of them working approx. 5 hours per day, including at weekends.

Guidelines were produced for the assistants who would be tasked with rounding up relevant subjects in the museum, inviting them to speak to a nice person, and unobtrusively keeping an eye on the proper workings of installed equipment. The assistants were university students who took turns on duty. They were instructed to locate and invite English-speaking youngsters to talk to "a nice person" and answer any questions they might have, except, of course, the one about whether the system is a real system or is faked in some way. Questions might be about the system, about the purpose of exhibiting it in the museum, the purpose of gathering acoustic data, etc. If the system is already occupied by users, whatever their age, the assistant should not interfere. In this way, we would gather data on adult-HCA and underage-HCA interaction.

We planned to put up two posters near the laptop in the museum, one in English and one in Danish, announcing the opportunity to talk to "a nice person", leaving identification of that person to the interested museum visitors.

The graphics on the laptop showed the 55 years old HCA who was able to turn towards the user and to open and shut his mouth in a shallow simulation of speaking. Figures 3.1 and 3.2 show, more or less, the movements of which HCA would be capable. HCA's graphical talking behaviour would have to be controlled by the wizard.



**Figure 3.1.** HCA at his writing desk.



**Figure 3.2.** HCA addressing the user.

One of the wizards had acted as wizard before, i.e. in the WoZ1 simulation in the autumn of 2002 which, compared to WoZ2, was far more wizard free-style conversation on behalf of HCA since no detailed output specification was available at the time. The second WoZ2 wizard had not acted as wizard before.

Six days before the actual start of the user tests, the wizards received the first electronic version of the document containing the HCA domain knowledge, i.e. questions to HCA, and HCA's answers. The wizards were encouraged to make themselves familiar with the contents of the document enabling them to navigate it quickly. Four days before the actual start of the user tests a dry-run took place at NISLab, as this was not possible at the HCA museum. A couple of hours of simulated dialogues were made testing the equipment and making the wizards familiar with the setup. Three days before the actual start of the user tests the wizards

received the first version of the wizard guide describing mainly how to behave like a system, cf. above.

### 3.2 Setup

The setup we used basically consisted of two PCs connected by a wireless network and two headsets with wireless audio connection as shown in Figure 3.3 and as explained in more detail below.

The PC network was used to control the simulation on the user's screen. The two wireless audio links were used to send the audio between the wizard and the user.

The PC at the wizard end recorded the spoken input and output using the program Cooledit2000. The PC was also used for displaying the wizard's manuscript and for remote control of the notebook at the user end.

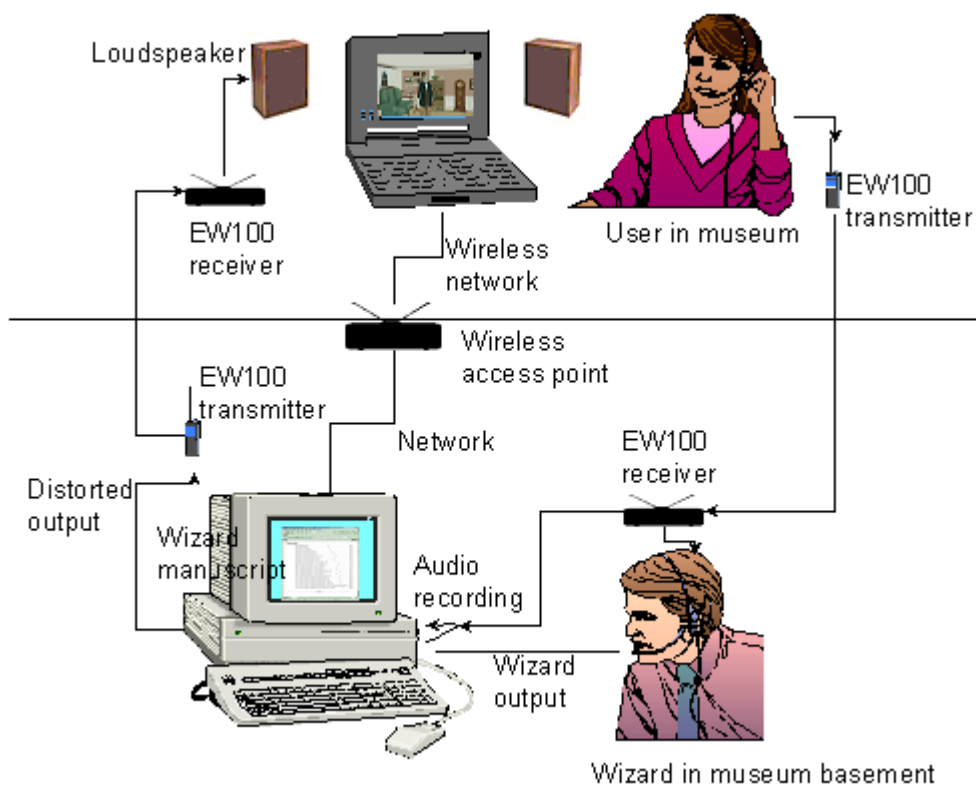


Figure 3.3. Wizard of Oz set-up in museum.

### 3.3 The simulation

The assistant answered questions and helped the users put on the microphone headset, ensuring as much as possible that the microphone had the right distance to the mouth of the user. Then the assistant instructed the users to say hello to the machine in order to start the conversation. During the conversation the assistant kept out of eyesight of the user but remained nearby if the user should have additional questions. Figure 3.4 shows HCA on the laptop in the museum. Figure 3.5 shows users in front of the laptop.

Mostly, the users reacted in a positive way to the posters and the assistants' invitations, provided that they had sufficient confidence in their English communication level. Those who refused the invitation were in many cases not able to speak and understand English.

Often the users had positive remarks after ending their conversations. A few of them even wrote about their experiences in the museum's Guest Book. A few examples are:

"Hans Christian Andersen is very polite but he could just know some more about his life and stories! Thanks a lot! Bye. Salerie, Switzerland".

"Impressive Exhibition HC Andersson really gets to life here!! The children loved talking to him! The Drauwee Family from Holland".



**Figure 3.4.** HCA Museum setup for the second NISLab Wizard of Oz simulation.



**Figure 3.5.** HCA and users in the second NISLab Wizard of Oz simulation.

The wizards chose different strategies for each recording day as they got more confident with their role. During the first couple of days the main goal was to get familiar with the system.

Later on one strategy was to sometimes misunderstand and not understand the user utterances in order to see how the users would react in such cases.



**Figure 3.6.** A wizard in action in the WoZ2 simulation.

One of the two wizards made use of HCA's lip movements in the life domain only, where the HCA answers are quite long. It was considered too time consuming to use it in other domains where HCA utterances were quite short, because each start and stop of the lip movements required two clicks with the mouse. The second wizard did not make use of the lip movements.

Ten consecutive days of 10 hours of being a wizard would mentally and physically have been too demanding for one person. Therefore, it was a good decision to make two wizards share the workload. In the first days of recording the wizards worked half a day each. Later in the recording period the wizards also played the role throughout entire days. This was mentally and physically OK, especially because each recording day had four breaks, three of 20 minutes each and one lasting 30 minutes, so that the wizards would be working at the most for two consecutive hours. A wizard in action is shown in Figure 3.6.

### 3.4 Conversation example

Figure 3.7 shows an example of a conversation from the WoZ2 experiments.

1: Marius: hello
2: HCA: hello welcome to my study
3: Marius: hello
4: HCA: hello welcome to my study
5: Marius: okay is that your study
6: HCA: this is my study
7: Marius: okay
8: HCA: this is where i write my fairy tales
9: Marius: which of your fairy tales have you been written on your study
10: HCA: all of them
11: Marius: all o' them can you tell me about one o' them



12: HCA: i can tell you about the Little Mermaid the Ugly Duckling and the Princess and the Pea  
13: Marius: i haven't heard much about the Little Mermaid could you please tell me 'bout it  
14: HCA: yes the mermaid saves a prince from drowning and falls in love with him to become human and get legs. she pays with her beautiful voice because she is now dumb she cannot tell the prince who she is and he marries another causing the mermaid to die  
15: Marius: okay a  
16: HCA: do do you like the story  
17: Marius: yea i like i've watched a c cartoon about it a ca a cartoon movie  
18: HCA: what is a cartoon movie  
19: Marius: well they use drawings and they use a lot o' them and then they make this whole movie about it  
20: HCA: what is your name  
21: Marius: my name is Marius  
22: HCA: how old are you  
23: Marius: i'm fourteen  
24: HCA: when i was your age i was a big fan of the theatre whenever there was a possibility i went to the theatre. it was my favourite place number one yea i also like the theatre  
25: Marius: it's fun  
26: HCA: are you a boy or a girl  
27: Marius: i'm a boy  
28: HCA: i guess i was not a typical boy  
29: Marius: okay  
30: HCA: i didn't like boys' games even at school i didn't take part in their games instead i stayed indoors where do you come from  
31: Marius: i come from Norway  
32: HCA: what kind of games do you like to play  
33: Marius: i like music it's not a game but it's one of my favourite things to do  
34: HCA: that must be very special can you tell me more  
35: Marius: well i have this keyboard and i can make music with drums bass yea  
36: HCA: i am curious to know which games children and teenagers play nowadays do they play any games which did not exist when i was a child  
37: Marius: i'm not sure i'm a big f fan of chess when it comes to games  
38: HCA: what is chess  
39: Marius: chess is a game when you use little f figures and you battle against another player  
40: HCA: that must be a very special game  
41: Marius: yea  
42: HCA: at my age i think the game i like the best is to play games with words or make paper cut-o uts for small children or charity  
43: Marius: yea do you like music  
44: HCA: yes i do like music  
45: Marius: what kind of music do you like hello  
46: HCA: it sure is a nice weather today  
47: Marius: yea it's very hot it's the middle of the summer you know  
48: HCA: i like the summer  
49: Marius: yea me too  
50: HCA: do you like my study  
51: Marius: yea i like your study  
52: HCA: if it was your study would you be missing anything  
53: Marius: what do you do if you went out of ink  
54: HCA: i will buy some more  
55: Marius: oh okay  
56: HCA: bye have a nice day  
57: Marius: have a nice day you are cool  
58: HCA: thank you

**Figure 3.7.** WoZ2 conversation with a 14 years old boy from Norway.

## **3.5 Initial data handling**

### **3.5.1 Diaries and lessons learned**

The wizards jointly wrote a brief diary during the ten days of WoZ experiments. The diary contained observations of various kinds. Some overall conclusions are that talking about inventions, which was one of the improvisations allowed to be made by the wizards, was a success and that the WoZ experiments involved a huge amount of work.

Overall the experiments were a success. However, we learned that the workload on the wizard was larger than anticipated. It would have been good to have two wizards working together, one to take care of the output to the user and one to handle the graphics (the mouth movements of HCA). Only one wizard made use of the possibility to make HCA move his lips and only in one domain with long output utterances. The wizards were otherwise too stretched to carry out this additional job.

A second lesson learned was that there should have been a technician around all the time to ensure, e.g., that the volume was adjusted to the right level so that too noisy recordings were avoided which just had to be trashed. We lost some valuable data on this account.

### **3.5.2 Transcription and additional markup**

Following the simulations, students transcribed the segmented audio files using Transcriber version 1.4.2. The transcription files were spell-checked using Microsoft Word.

The speech recognition partner in the NICE project wanted markup of the speaker's level of English inserted into the annotation files. The rules used in WoZ1 were reused in an elaborated version. We distinguish between four levels of expertise, i.e. bad, medium, good, and native. Transcriber version 1.4.2 was used for inserting the evaluation tags. The annotation was done by a phonetician. However, the idea behind the rules in the elaborated version is that it should be possible to arrive at the same evaluation results also if the rules are applied by non-phoneticians.

About 70% of the transcribed dialogues have been topic tagged in order to give us an idea of the topics addressed by users and thereby to provide input to the design of the first and second prototypes. Only the user's turns have been tagged and the tagging is done without regard to the context. A set of rules has been established to guide the topic tagging. Transcriber version 1.4.2 was used for inserting the markup. One person made the topic tagging while a second person verified the correctness of the inserted topics. Mistakes were corrected, possibly after discussing disagreements.

About 50% of the data has been treated with respect to semantics. The purpose was to create material for the training and testing of the natural language understanding (NLU) component. The semantics processing has been done by only one person. The reason was that any errors would be caught by the NLU so we did not want to spend the additional time on letting a second person check what had been done. Again, a set of rules were used to guide the process.

The annotated dialogues have been further analysed as described in Section 3.6. To prepare for this process, a php script was made which can transform the Transcriber XML-files (which can only be printed as XML) into a more readable version that can be viewed in a browser and that can also be printed in an nicely readable version.

## **3.6 Analysis of age, gender, turn length, relative target group success**

In this section, we briefly describe the types of WoZ2 data analysis which have been carried out so far. The results have been reported in publications referenced at appropriate points below. We also mention ongoing and future work on analysing this amazingly rich data.

### 3.6.1 Age, gender, turn length

Table 3.1 shows the basic WoZ2 data. Turn numbers show the total number of turns made by the user and HCA in a conversation. Since they take turns communicating, each of them will produce half of the turns +/- a single turn.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Total
No. dialogues	29	42	33	33	45	75	66	45	61	69	<b>498</b>
Age <10	1	3	3	1	2	9	3	7	9	11	<b>49</b>
Age 10-18	12	25	20	20	19	42	17	27	34	24	<b>240</b>
Age >18	14	11	3	8	22	22	35	4	14	31	<b>164</b>
Age unknown	2	3	7	4	2	2	11	7	4	3	<b>45</b>
Male	16	19	17	14	25	37	26	22	29	29	<b>234</b>
Female	10	20	9	15	18	30	28	18	28	34	<b>210</b>
Gender unknown	3	3	7	4	2	8	12	5	4	6	<b>54</b>
No. countries	8	10	8	8	11	17	14	10	14	13	<b>29</b>
No. turns all	851	1030	1029	1301	948	1808	2278	1315	1677	1502	<b>13739</b>
Av. no. turns all	29	25	31	39	21	24	35	29	27	22	<b>28</b>
No. turns <10	5	58	212	26	36	197	64	184	291	194	<b>1267</b>
Av. no. turns <10	5	19	71	26	18	22	21	26	32	18	<b>26</b>
No. turns 10-18	405	628	665	907	434	1090	835	949	1043	607	<b>7563</b>
Av. no. turns 10-18	34	25	33	45	23	26	49	35	31	25	<b>32</b>
No. turns >18	409	312	73	298	462	486	1218	101	319	650	<b>4328</b>
Av. no. turns >18	29	28	24	37	21	22	35	25	23	21	<b>26</b>
No. turns age unknown	32	32	79	70	16	35	161	81	24	51	<b>581</b>
Av. no. turns age unknown	16	11	11	18	8	18	15	12	6	17	<b>13</b>
No turns male	447	408	556	626	591	923	972	674	743	749	<b>6689</b>
Av. no. turns male	28	21	33	45	24	25	37	31	26	26	<b>29</b>
No. turns female	367	547	394	605	345	771	1101	604	910	666	<b>6310</b>
Av. no. turns female	37	27	44	40	19	26	39	34	33	20	<b>30</b>
No turns gender unknown	37	75	79	70	12	114	205	37	24	87	<b>740</b>
Av. no. turns gender unknown	12	25	11	18	6	14	17	7	6	15	<b>14</b>

**Table 3.1.** Basic WoZ2 data description. Dn means Day n.

The total of 498 conversations only excludes four conversations of <4 turns and two conversations in which the transcribers mixed up the users. The reason why Table 3.1 provides substantial information on users' age, gender and nationality, is that HCA has as a priority in conversation to gather this user information for conversational use. He will thus try to collect this information either up front or, at least, early on in each conversation. Roughly, age, gender, and nationality information was provided by 90% of the users. The most common reason, by far, for not providing age, gender, and/or nationality information was that the user broke off conversation before HCA could gather this data. Thus, the average turn numbers for no-age and no-gender users is as low as 13 and 14, respectively. In a few cases, the wizards forgot to ask for the information. Few users refused to tell HCA their age or gender, and only in a couple of cases is there reason to believe that a user gave deliberately

wrong information. An example is Maria on Day 9 who first had a 98-turn conversation as Maria, an 11 years old female from Denmark, and then came back to have a 24-turn conversation as Maria, a 13 years old boy from Denmark wanting to discuss girls with HCA, unfortunately with limited success.

Table 3.1 shows a rather close gender balance of 210 (47.3%) female users and 234 (52.7%) male users, as well as near-identical turn averages for female and male users, i.e. 30 and 29, respectively.

The basic WoZ2 data are discussed in [Bernsen et al. 2004] and [Bernsen 2004].

### 3.6.2 Relative target group success

Relative target group success, i.e. the extent to which the system simulated in WoZ2 specifically manages to reach its target user group as compared with the groups of younger and older users, has been measured for the entire WoZ2 corpus. The results, which include a novel general metrics for measuring relative target group success (RTGS), are reported in [Bernsen 2004]. Briefly, the RTGS was found to be highly dependent on whether the defined user groups could or could not be assumed to have English as a first language. This led to the conclusion that RTGS must be measured for native speakers. Whether or not a user is in the target age group, the better the user’s English skills, the longer that user is likely to speak with HCA up to 38 turns on average per conversation. Considering native English speakers only, the <10/10-18 RTGS is only 2.7% whereas the 10-18/>18 RTGS is 35.7%. The marked RTGS difference just described between, on the one hand, the <10/10-18 years old and, on the other, the 10-18/>18 years old, suggests that the application clearly has stronger appeal to the <19 years old than to adults. This conclusion is supported by another finding, i.e. that the top-ten user-HCA conversations, which have a staggering average of 111 turns, all involve 6-17 years old youngsters (see also next section).

### 3.7 In-depth analysis

Dialogue id	Name	Age	Gender	Nation	No. turns	Level	Turn ranking
3, 1-1, 1-98	Vaughn	6	male	USA	98	G	<b>6</b>
4, 2-4, 1-58	Marius-1	14	male	Norway	58	M	<b>2 157</b>
4, 2-8, 1-99	Marius-2	14	male	Norway	99	M	<b>2</b>
5, 11-1, 1-123	Ole-1	12	male	Norway	123	M	<b>1 185</b>
5, 12-1, 1-62	Ole-2	12	male	Norway	62	G	<b>1</b>
7, 10-6, 1-99	Rebecca	11	female	Holland	99	G	<b>5</b>
7, 3-5, 1-105	Christian	12	male	Denmark	105	M	<b>4</b>
7, 4-5, 1-90	Signe	15	female	Denmark	90	G	<b>8</b>
7, 9-5, 1-82	Nick	13	male	Holland	82	M	<b>9</b>
8, 7-8, 1-95	Marie	15	female	Denmark	95	G	<b>7</b>
9, 3-11, 1-79	Jenny	17	female	England	79	G	<b>10</b>
9, 7-14, 1-98	Maria-1	11	female	Denmark	98	M	<b>3 122</b>
9, 7-15, 1-24	Maria-2	13	male	Denmark	24	M	<b>3</b>

**Table 3.2.** The top-ten conversationalists in the WoZ2 corpus.

In depth analysis of the WoZ2 data has been made of the top-ten conversations measured in terms of conversation length. These conversations have a turn average of 111 turns. The conversations are gender-balanced and all involve target group users, see Table 3.2.

Based on the theory of conversation underlying NICE HCA PT1, cf, NICE Deliverable D1.2, the conversations have been tagged using a number of new coding schemes developed for this purpose. Following corpus coding, new metrics have been developed in order to measure the extent to which HCA and his various interlocutors address their respective domains of expertise, take the initiative to change domain or topic during conversation, and contribute to driving the conversation forward by asking questions or volunteering information. The results are reported in [Bernsen et al. 2004]. A brief summary follows in the sub-sections below.

### 3.7.1 Domain/topic ranking and expertise balance

Table 3.3 shows that the user/HCA balance of expertise in the top-ten conversations was reasonable (485/520 turns) and that both the user and HCA were, on average, reasonably active in taking the initiative to change topics. Table 3.4 shows, however, that topic change initiative (“phase I”) was highly diverse among the top-ten users, ranging from a user/HCA symmetry of 0.11 (Vaughn, Rebecca) to a symmetry of 8.00 (Maria).

Italics in the left-most column of Table 3.3 show the topics improvised by the wizards. Not surprisingly, Column 3 from the left shows that HCA’s fairytales were the top subject in conversation. Interestingly, modern inventions came second and games came in third place. Surprisingly, HCA’s life came only in sixth place, suggesting only modest user interest in HCA’s life. This finding was contradicted by the user tests (WoZ3) in which users took a strong interest in HCA’s life, stronger even than their interest in his fairytales.

	No. talks (rank)	No. turns (rank)	Turns/ talk	User expert	HCA expert	User init.	HCA init.
<b>Age HCA</b>	3 (10)	6 (13)	2		6	1	2
<b>Games</b>	11 (6)	134 (3)	12	134		2	9
<b>Greet</b>	23 (1)	69 (7)	3			20	3
<i>Inventions</i>	16 (3)	189 (2)	12	189		10	6
<b>Know HCA</b>	1 (11)	7 (12)	7		7		1
<b>Life HCA</b>	11 (6)	83 (6)	8		83	9	2
<b>Looks HCA</b>	3 (10)	6 (13)	2			1	2
<i>Museum</i>	4 (9)	16 (11)	4	16		-	4
<i>OOD</i>	1 (11)	4 (14)	4			1	-
<b>Study HCA</b>	18 (2)	100 (5)	6		100	5	13
<i>Travels</i>	5 (8)	29 (9)	6		29	1	4
<b>User</b>	12 (5)	122 (4)	10	122		-	12
<i>Vacation</i>	6 (7)	24 (10)	4	24		-	6
<i>Weather</i>	1 (11)	4 (14)	4			-	1
<b>Who HCA</b>	5 (8)	37 (8)	7		37	4	1
<b>Works HCA</b>	15 (4)	258 (1)	17		258	5	10
<b>Totals</b>	<b>135</b>	<b>1088</b>		<b>485</b>	<b>520</b>	<b>59</b>	<b>76</b>

**Table 3.3.** User and HCA areas of expertise and topic change initiative in the top-ten WoZ2 conversations.

### 3.7.2 Phase initiative and drive symmetry

To measure the extent to which the top-ten WoZ2 users took initiative in driving the conversation forward rather than leaving the initiative with HCA, we measured the user/HCA

symmetry in asking questions and volunteering information. Table 3.4 shows (Column 8 from the left: drive sym.) that the users were, on average, far less active in driving the conversation forward than was HCA. Only Marius, cf. Figure 3.7, came close to full symmetry with a drive symmetry score of 0.93.

	Questions	Observations	Words	Q + O	Turns	Q+O/turn	PhaseI sym./drive sym. user/HCA	Words/turn	Expert phases/turns
Vaughn	0	2	161	2	49	<b>0.04</b>	<b>0.11/0.04</b>	<b>3.3</b>	<b>6/54</b>
HCA	42	5		47	49	<b>0.96</b>			<b>4/39</b>
Marius	29	15	650	44	79	<b>0.56</b>	<b>1.43/0.93</b>	<b>8.2</b>	<b>7/79</b>
HCA	27	20		47	78	<b>0.60</b>			<b>10/53</b>
Ole	7	12	541	19	93	<b>0.20</b>	<b>0.60/0.26</b>	<b>5.8</b>	<b>9/71</b>
HCA	43	27		70	92	<b>0.76</b>			<b>7/104</b>
Rebecca	0	8	390	8	50	<b>0.16</b>	<b>0.11/0.19</b>	<b>7.8</b>	<b>6/44</b>
HCA	29	13		42	49	<b>0.86</b>			<b>4/45</b>
Christian	7	7	356	14	53	<b>0.26</b>	<b>0.44/0.35</b>	<b>6.7</b>	<b>7/69</b>
HCA	32	7		39	52	<b>0.75</b>			<b>6/32</b>
Signe	1	6	453	7	45	<b>0.16</b>	<b>0.13/0.18</b>	<b>10.1</b>	<b>4/35</b>
HCA	26	15		41	45	<b>0.91</b>			<b>5/50</b>
Nick	1	4	271	5	41	<b>0.12</b>	<b>0.50/0.16</b>	<b>6.6</b>	<b>3/46</b>
HCA	25	5		30	41	<b>0.73</b>			<b>3/32</b>
Marie	6	5	503	11	48	<b>0.23</b>	<b>1.00/0.35</b>	<b>10.5</b>	<b>3/41</b>
HCA	18	13		31	47	<b>0.66</b>			<b>5/45</b>
Jenny	3	2	388	5	40	<b>0.13</b>	<b>0.14/0.15</b>	<b>9.7</b>	<b>4/36</b>
HCA	27	7		34	39	<b>0.87</b>			<b>4/38</b>
Maria	21	1	282	22	49	<b>0.45</b>	<b>8.00/0.59</b>	<b>5.8</b>	<b>1/14</b>
HCA	15	22		37	49	<b>0.76</b>			<b>8/77</b>

**Table 3.4.** Data on phase initiative and conversation drive symmetry in the WoZ2 top-ten corpus.

### 3.7.3 Test-driving the PT2 specification

In our ongoing work towards improving the NICE HCA system’s crucial capability of managing non-task-oriented, domain-oriented conversation, we have found an unexpected use for the WoZ2 data. Since this data was gathered based on the first NICE HCA prototype (PT1) specification, the data has turned out to be tremendously useful in test-driving our specification of the second NICE HCA prototype (PT2). What we do is to go through selected WoZ2 conversations turn-by-turn and ask, at each step, whether and how the PT2 specification could handle the transcribed WoZ2 data. The smoothness and conversational coherence with which human wizards handle the PT1 specification is truly amazing. However, since they are actually handling, i.e. making conversation on the basis of, the PT1 specification, their conversational behaviour constitutes an excellent “gold standard” for how PT2 should be capable of performing. We are not aware of similar test-drive studies in the literature.

### **3.7.4 Future work on WoZ2 corpus analysis**

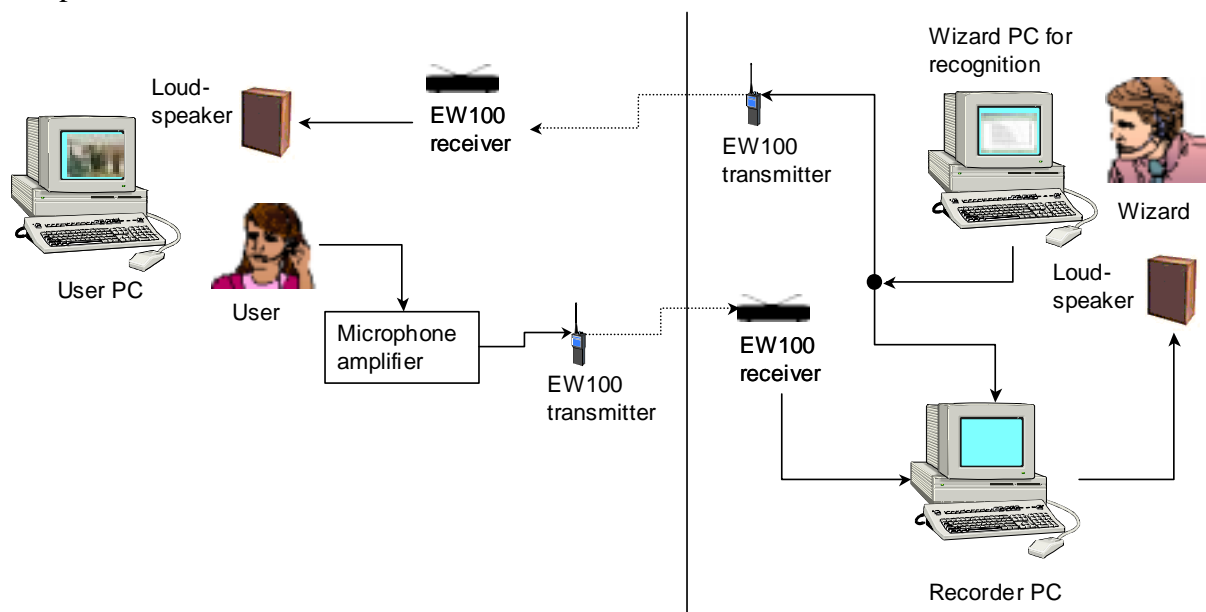
In addition to what was mentioned in Section 3.7.3, we plan to analyse the WoZ2 corpus in order to investigate reasons why the users chose to stop the conversation with HCA in order to, possibly, identify conversation behaviours which HCA should avoid in PT2. Other WoZ2 analysis goals will be identified in due course.

## 4 WoZ3

The WoZ3 experiments were carried out at NISLab on 20 – 22 January 2004. 18 users from the target user group of 10-18 year olds participated in the test. The system tested was the implemented first NICE HCA prototype except for the speech recogniser which was simulated. In the NICE HCA project plan, it was also envisaged that the natural language understanding module would have had to be simulated in the first user tests. However, we realised from early on that its simulation would have rendered the first NICE HCA system user test difficult to conduct and less than sufficiently informative in their outcome, so we have accelerated work on our new natural language understanding module in order to be able to conduct sufficiently informative user tests.

### 4.1 Setup

We had users coming in parallel. Thus, two offices had been prepared for user testing. In one room, the user had both a mouse and an 19” touch screen for producing 2D gesture input. In addition, the user could watch HCA on a 42” flat-panel screen. In the second room, the user had a mouse and an ordinary 19” screen. The two wizards (one per user) were placed in a third room. Their job was to type in what the users said, thus replacing the speech recogniser which had not yet been integrated with the rest of the system. Figure 4.1 shows the general setup between the user and the wizard.



**Figure 4.1.** Setup in the WoZ3 user test.

In addition to recording the sound, all dialogues were logged and the sessions were recorded on video. In the room with the touch-screen, there was one video camera pointing at the user and a second camera recording the screen. In the second room there was only a camera recording the user. All this resulted in approximately 11 hours of audio, video, and logfile-recorded interaction, and 18 sets of structured interview notes.

The WoZ3 experiments required the involvement of several people acting in different roles, including a technician who made sure that the technical side of the experiments was working; an experimenter who welcomed the subjects and gave them instructions before the two system sessions; two wizards who typed what the users said, following wizard instructions given to



them in advance, e.g. telling how to type numbers and how to handle ehms and ahhs; two interviewers who interviewed the subjects after their interaction with the system; and an observer located in the room with the touch screen; and 18 subjects who interacted with the system.

## 4.2 The simulation

Each user session had a duration of 60-75 minutes. A test session included conversation with HCA in two different conditions followed by a post-test interview. In the first condition, the users only received basic instructions on how to operate the system, i.e. speak using the headset, control HCA's movements, control the four camera angles, and gesture using mouse or touch screen. After 15 minutes, the first session was interrupted and the user received a set of 13 typed scenario problems to be solved through speech or gesture input in the second session, such as "Find out if HCA has a preferred fairy tale and what it is" and "Tell HCA about games you like or know". The problems might be addressed in any order and the user was not necessarily expected to carry out all of them. The purpose was to ensure a significant amount of user initiative to explore how the system would respond under the resulting user initiative pressure. Figures 4.2 through 4.7 show photos from the WoZ3 simulation.



**Figure 4.2.** The two wizards in action.



**Figure 4.3.** A subject interacting with HCA and an observer.



**Figure 4.4.** A young user in the room with the ordinary screen.



**Figure 4.5.** A user in front of the touch screen.



**Figure 4.6.** An interview situation.



**Figure 4.7.** The second interviewer in action.

After having interacted with the system in the two conditions described above, the users were interviewed. Each user was asked the questions listed below.

**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 interviewers made hand-written notes during each interview which were later typed according to the questions to which they belonged.

### 4.3 Conversation example

Figure 4.8 shows an excerpt of a conversation between a user and HCA in WoZ3.

**nice turn=1**

**input:** *can you tell me about one of your adventures fairy tales*

**speech output:** # *I was only thirty-one years old when I wrote The Little Mermaid.*

**nice turn=2**

**input:** *tell me about one of your fairy tales*

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

**nice turn=3**

**input:** *tell me about the little mermaid tell me about hte little mermaid*

**speech output:** # *Did I already tell you that only the bible has been translated into more languages than my fairy tales.*

**nice turn=4**

**input:** *tell me about the princess and the pea*

**speech output:** # # *Excuse me?*

**nice turn=5**

**input:** *tell me about one of your fairy tales*

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

**nice turn=6**

**input:** *picturecolosseumrome*

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

**nice turn=7**

**input:** *picturejonascollin*

**speech output:** *The picture shows a gentleman whose name is Jonas Collin. # He has had a huge influence on my life. He helped me in so many ways especially when I was young poor and alone. He helped me to get an education. Without his help I would probably not have become a famous author. # # I owe him everything.*

# *One of my favourite draughts is beer # from the oudensae brewery Albani.*

**nice turn=8**

**input:** *what is your favourite beer*

**speech output:** # *What do you think of the princess the character from the Princess and the*

*Pea?*  
**nice turn=9**  
**input:** *travelbag*  
**speech output:** *# This is my travel bag. I use it a lot because I travel a lot. I have been in many countries in Europe and even as far away as Northern Africa. Altogether I have spent several years of my life travelling around. # I love to travel especially by train.*  
**nice turn=10**  
**input:** *boots*  
**speech output:** *# I do not know much about it yet. # I hope I can tell you more next time we meet. # What do you think about it?*

**Figure 4.8.** Excerpt of a conversation with HCA from WoZ3, including gesture input such as “picture n” or “travelbag”.

#### 4.4 Analysis, basic data

Table 4.1 shows the basic data on the WoZ3 user test, indicating, among other things, perfect gender balance, a slight difference in average age among female and male users, and a wide spread in computer games practice. Only three users had spoken to a computer before.

Property	Value
No. girls	9
No. boys	9
Nationality	17 Danish, 1 Scottish
School	all
Age girls:	17, 12, 14, 17, 12, 17, 12, 15, 17
Age boys:	18, 13, 10, 14, 15, 15, 13, 11, 15
Girls, average age:	14.8
Boys, average age:	13.8
Age range girls:	12-17
Age range boys:	10-18
All, average age:	14.3
Girls, computer game hours per week:	1, 0, 7, 3, 21, 1, 0, 1.5, 1
Boys, computer game hours per week:	20, 7, 9, 7, 0, 23, 2.5, 24.5, 7
Girls, average computer game hours per week:	3.9
Boys, average computer game hours per week:	11.1
Girls, computer game playing range hours per week:	0-21
Boys, computer game playing range hours per week:	0-24.5
All, average computer game hours per week:	7.5
Talked to a computer before:	3 (all girls)

**Table 4.1.** Basic data on the user test.

#### 4.5 In-depth conversation analysis, interview analysis

The NICE HCA PT1 evaluations based on in-depth analysis of user test conversations and user interviews are reported in NICE Deliverable D7.2.

## 5 References

[Bernsen 2004] Bernsen, N. O.: Measuring relative target user group success in spoken conversation for edutainment. To appear in *Proceedings of the LREC 2004 Workshop on Multimodal Corpora*. Lisbon, Portugal, May 2004.

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

[Bernsen et al. 2004] 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.