NICE project (IST-2001-35293)

Natural Interactive Communication for Edutainment

NICE Deliverable D7.2-2

Evaluation of the Second NICE Prototype

4 May 2005

Editors

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

Authors

Chapter 2: Niels Ole Bernsen[1], Stéphanie Buisine[2], Marcela Charfuelan[1], Andrea Corradini[1], Laila Dybkjær[1], Holmer Hemsen[1], Jean-Claude Martin[2], Manish Mehta[1]

Chapter 3: Linda Bell[3], Reinhard Blasig[4], Johan Boye[3], Stéphanie Buisine[2], Joakim Gustafson[3], Mattias Heldner[3], Anders Lindström[3], Jean-Claude Martin[2], Mats Wirén[3]

1: NISLab, Odense, Denmark, 2: LIMSI-CNRS, Orsay, France, 3: TeliaSonera, Sweden, 4: Scansoft, Germany

Project ref. no.	IST-2001-35293
Project acronym	NICE
Deliverable status	Public *after* 15 June 2005 (this report includes data from submitted papers). **Please only quote upon explicit permission from the partner who wrote a particular section.**
Contractual date of delivery	28 February 2005
Actual date of delivery	4 May 2005
Deliverable number	D7.2-2
Deliverable title	Evaluation of the Second NICE Prototype
Nature	Report
Status & version	Final
Number of pages	99
WP contributing to the deliverable	WP7
WP / Task responsible	WP7/NISLab
Editors	Niels Ole Bernsen and Laila Dybkjær
Author(s)	Chapters 1 and 2: Niels Ole Bernsen (chapter 1, parts of chapter 2), Stéphanie Buisine (parts of chapter 2), Marcela Charfuelan (parts of chapter 2), Andrea Corradini (parts of chapter 2), Laila Dybkjær (parts of chapter 2), Holmer Hemsen (parts of chapter 2), Svend Kiilerich (parts of chapter 2), Jean-Claude Martin (parts of chapter 2), Manish Mehta (parts of chapter 2), Chapter 3: Linda Bell, Reinhard Blasig, Johan Boye, Stéphanie Buisine, Joakim Gustafson, Mattias Heldner, Anders Lindström, Jean-Claude Martin, Mats Wirén
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 D7.2-2 of the EC Human Language Technologies project NICE (Natural Interactive Conversation for Edutainment), presents results of the user evaluations of the second system prototypes, one for English conversation with fairytale author Hans Christian Andersen and one for Swedish conversation with some of his fairytale characters.

Table of Contents

1 Introduction

This report presents results from analyses of the user tests of the two second prototype systems developed in the NICE project, i.e., the Hans Christian Andersen (HCA) system (Chapter 2) and the Fairytale World (FTW) system (Chapter 3). The HCA system was tested with target group users in February 2005 and the FTW system was tested with 57 target group users during the period from November 2004 through March 2005.

Both user test reports below are based on the NICE system and component test criteria specified in NICE Deliverable D7.1, *Evaluation criteria and evaluation plan,* Section 5. These test criteria are all presented in Tables 2.1 through 2.4 in Chapter 2, sometimes in a modified form compared to their predecessors in D7.1. The modifications take the form of either (i) splitting an original criterion into two separate ones for clarity of evaluation, (ii) re-phrasing an original criterion for clarity, or (iii) adding a new criterion when this was found missing in the original.

2 The Hans Christian Andersen system

The results of the user test of the first HCA system prototype (HCA PT1) are reported in deliverable D7.2a. This deliverable reports on the user test of the second HCA system prototype (HCA PT2).

2.1 Game environment

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

The user sees HCA in his study in Copenhagen (Figure 2.1) and communicates with him in fully mixed-initiative conversation using spontaneous speech and 2D gesture. Thus, the user can change the topic of conversation, back-channel comments on what HCA is saying, or point to objects in HCA's study whenever s/he wants, and receive his response when appropriate. 3D animated HCA communicates through audiovisual speech, gesture, facial expression, body movement and action. The high-level theory of conversation underlying HCA's conversational behaviour is derived from analyses of social conversations aimed at making new friends, emphasising common ground, expressive story-telling, rhapsodic topic shifts, balance of "expertise", etc. When HCA is alone in his study, he goes about his work, thinking, meandering in locomotion, looking out at the streets of Copenhagen, etc. When the user points at an object in his study, he looks at the object and then looks back at the user before telling a story about the object. HCA has knowledge about his works, in particular three of his fairytales, his life, mostly his childhood, including childhood games and games users like, his physical and personal presence, his study including the objects in there, the user, and generic input including meta-communication. The objects that the user via gesture can get information about from HCA (hereafter called "referenceable objects") are the 16 pictures on the walls, a feather pen and a travel bag.

Figure 2.1. HCA gesturing in his study.

2.2 Data collection

2.2.1 System description

The HCA PT2 system version used in the user test has the general architecture shown in Figure 2.2.

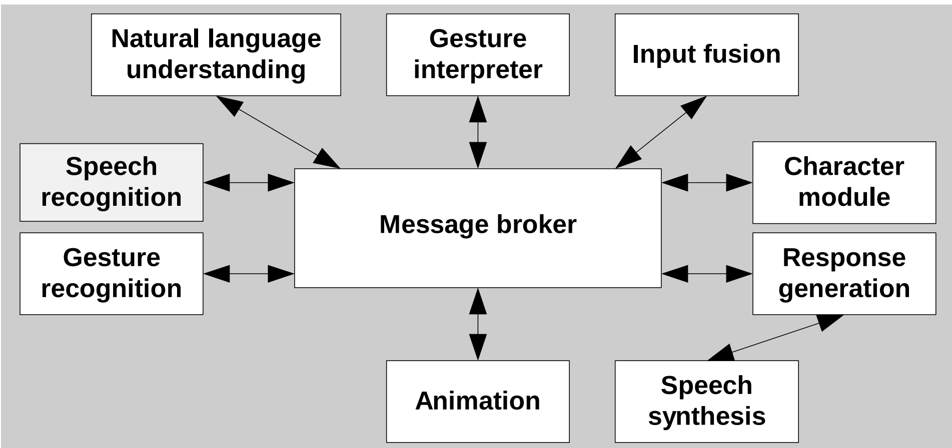

Figure 2.2. General architecture of the second HCA system prototype.

The speech recogniser is the SpeechPearl recogniser trained by Scansoft on approx. 50 hours of speech data (primarily children's voices) from the NICE project collected by NISLab as well as on in-house Scansoft data. NISLab has developed and trained the vocabulary and language models. The vocabulary size is 1977 words. Barge-in is not enabled. Instead, the recogniser times out after n seconds and is re-activated by a message from Response generation that HCA has finished his current output.

The natural language understanding (NLU) module is developed by NISLab. Different components of the NLU are a) NLU Manager b) Key-Phrase Spotter c) Semantic Analyser d) Concept Finder and e) Domain Spotter. In terms of general information flow, the NLU receives the user utterance from the Speech Recogniser in terms of an N best list (currently N = 3) detected by the speech recogniser. The NLU analyses the top result from this list. The NLU Module Manager is responsible for communication across different components of the module. The Key-Phrase Spotter spots key phrases in the user utterance and converts them into syntactic/semantic categories. Each set of key phrases associate with syntactic/semantic categories. The output of the key phrase module is passed on to the semantic analyser. The semantic analyser consists of a number spotter, a lexicon and a rule engine. The number spotter helps in finding the user's and HCA's age. The lexicon entries consist of syntactic/semantic categories for individual words. After passing through the number spotter and lexicon, the processed user input is a sequence of semantic and syntactic categories. The rule engine processes this sequence by applying rules defined on the presence of certain semantic/ syntactic categories at specific positions in the user input. The resultant sequence is sent to the FSA processor. It acts as the deepest level of parsing. If the user input sequence is able to traverse an FSA, the result corresponding to that FSA is the output semantic representation from the semantic analyser. The FSAs are developed offline from training corpora. In the next stage, the concept finder provides a mapping from semantic representations to a domain-oriented ontological representation. The Domain Spotter finds the domain by mapping the concepts to their respective domains defined. This mapping is defined at design time. The final output consisting of concept(s)/subconcept(s), property(ies), values, dialogue act and domain is sent to the Character Module via Input Fusion. The NLU operates with approximately 270 concepts of domain, dialogue act, dialogue act type, concept(s), subconcept(s), property(ies), and value(s).

The Gesture Recogniser module (GR), the Gesture Interpreter module (GI) and the Input Fusion module (IF) are developed by LIMSI-CNRS.

The GR recognises the following 2D gesture shapes: pointer (e.g. a pointing gesture or a very small gesture since on-tactile-screen pointing is rarely producing a single point), connect (e.g. a line), surrounder (e.g. a circle or similar shape encircling an object), and unknown shape.

The Gesture Interpreter module (GI) detects the object(s) the user gestured at. It can produce either "select" (e.g. a single object was gestured at), "reference ambiguity" (e.g. several objects were gestured at), "no object" (a gesture was done but no associated referenceable object could be detected). In order to avoid endless buffering of the user's input while HCA is responding, gesture interpretation is inhibited during preparation and synthesis of HCA's verbal and non verbal behavior.

The Input Fusion module (IF) has to integrate the messages sent by the NLU module and the GI and send the result to the character module. The IF parses the message sent by the NLU to find any explicit references (e.g. "this picture") or implicit references (e.g. "Jenny Lind?", "Do you like travelling?") which might be integrated with gestures on objects in the study. It produces messages containing a "fusion status" which can be either "ok" (the utterance and the gestured object were integrated because a reference was detected in the NLU message and in the GI), "none" (the utterance and the gesture were not integrated either because there was either only one of them, or because the IF could not decide if they were consistent or not regarding the number of references to objects in speech and gesture), or "inconsistent" (the utterance and the gesture were inconsistent regarding the number of referenced objects). In case of successful integration, the semantic representation of gesture (the detected object) is inserted in the semantic representation sent by the NLU. The IF module also manages temporal delays between gesture and speech via several time-outs and messages signalling start of speech and start of gesture.

According to its design, the character module, developed by NISLab, is supposed to always be in one of three output states, i.e., non-communicative action (NCA) when HCA is alone in his study, communicative function (CF) when HCA pays attention to the user's spoken input, and communicative action (CA) when HCA actually responds to the user's input. However, the three states are not integrated and can only be shown apart, cf. below. Both non-communicative function (including locomotion) and communicative function are script-based. When HCA has visitors, the character module decides on the next conversational move based on the input from the input fusion module and the current dialogue state. The character module operates over a forest of domain ontology trees and draws on a distributed discourse and domain history. It also calculates a new emotional state for HCA for each dialogue turn. When the next move has been determined, including emotional state, a database which contains HCA's knowledge, is contacted. It returns a parameterised semantic instruction composed of input values, text-to-speech (TTS) references and/or references to non-verbal behaviours which is sent to the response generator. The database also returns information about which new dialogue state to go to so that the character module is ready for the next input.

The response generator is also developed by NISLab. Based on the input from the character module, the response generator generates a surface language string with animation and control (e.g. camera view) tags. The result is sent to the speech synthesiser which synthesises the verbal output and, whenever it meets an animation tag, sends a message to the response generator that now the corresponding non-verbal output descriptions must be sent to the animation module which takes care of the graphics output. The second NICE HCA prototype uses approx. 450 spoken utterance templates and 130 different non-verbal behaviour primitives. Up to 17 non-verbal primitives are used per output turn. Additionally a total of eleven visemes are used for lip synchrony.

The speech synthesis is off-the-shelf software from AT&T. The voice has been chosen partly for its inherent quality and partly for matching the voice one would expect from a 55 years old man.

The animation module is developed by Liquid Media. It renders the virtual world and the character animation. It also enables the user to change camera angle using the F2 key and to control HCA's locomotion using the arrow keys.

All modules communicate via a central message broker, publicly available from KTH at http://www.speech.kth.se/broker. The broker is a server which routes function calls, results and error codes between modules.

By comparison with the HCA PT2 specification, the main structural limitation of the tested HCA PT2 was that HCA's three different output states had not been fully integrated. Thus, HCA had been specified to (1) go about his work in his study when not having conversation with a user, (2) show real-time awareness of the user's spoken and/or gesture input, and (3) produce his own conversational output to the user. The lack of integration meant that HCA would actually be in output state (1) when a new user initiated conversation. At that point, however, (1) had to be manually interrupted in favour of (3) unless (1) would continue and interfere with (3). (2) had not been implemented according to specification and had not been integrated. A small part of (2) did work, however. Thus, when a user points to an object in his study, HCA turns towards that object and then turns back to face the user.

2.2.2 Test subjects and test language

HCA PT2 was tested with 13 users (six boys and seven girls) from the target user population of 10-18 years old children and teenagers. All users were Danish school kids aged between 11 and 16 and with an average age of 13 years.

The conversation with HCA was conducted in English.

2.2.3 Test method

The test method used was a controlled in-laboratory user test similar to the one conducted with HCA PT1 in January 2004. The main difference between the two user tests was that PT1 did not include speech recognition. Instead, human wizards typed what the users said whereupon the system processed the typed input.

Each user test session took 60-75 minutes. Sessions began with a brief introduction by the experimenter to the system setup and the input modalities available, and calibration of the headset microphone to the user's voice. Like in the PT1 user test, each user tested the system in two different test conditions, a free-style conversation condition followed by a condition based on a conversation problems handout.

At the beginning of each session, the experimenter demonstrated both gesture only behaviour (point, line, circle) and multimodal input with a single example of combination such as "what is in this picture?" combined with a gesture to a picture. Users were also told that they had to speak in English and briefly what HCA knows about. However, they were otherwise *not* instructed in how to speak to the system at all.

Then followed 15 minutes of free-style interaction in which it was entirely up to the user to decide what to talk to HCA about. In the following break, the user was asked to study a handout which listed 11 proposals on what the user could try to find out about HCA's knowledge domains, make him do, or explain to him. Some examples are that the user could make HCA tell about his life and family relations, tell HCA about games the user likes, collect as much information as possible about the place where HCA lives, or be rude to him and see what happens. It was stressed that the user was not required to try to follow all the proposals. Rather, the user could pick those he or she liked, having a good time in the process. The second session had a duration of 20 minutes. Figure 2.3 shows a user in action during this session.

The original idea of splitting each user test session into two sub-sessions with the system was to observe which difference the handout made to the users' conversation. We were interested in knowing what users would talk to HCA about when they started out "cold" and, in particular, we would like to know if the handout influenced their style of conversation. For example, how much initiative would they take in each of the two sub-sessions and how much information would they volunteer. It was not possible to counterbalance these two sessions across users as starting by the instruction session would have influenced the free session.

Figure 2.3. A user talking to the second HCA system prototype.

Following the two sessions with HCA, each user was interviewed separately about his/her background, experiences from interacting with HCA, views on system usability, proposals for system improvements, etc. The full set of interview questions is shown in Figure 2.4 except for the questions concerning the user's background, such as name and age.

2.2.4 Data collection set-up

Each user used a microphone/loudspeaker headset, a touch screen for gesture input, and a keyboard for controlling virtual camera angles and for controlling HCA's locomotion. Two cameras captured the user's behaviour during interaction. A developer sat behind the user, monitoring the workings of the system modules, making observations, being ready to re-start a hanging or crashed system component, help users in case of difficulties with the equipment, etc. In about half of the user tests, an observer was present in the test room as well.

Users' spoken input was recorded and all main module outputs were logged.

2.2.5 Quantitative data description

A total of 26 conversations corresponding to 8 hours of speech were recorded, logged and captured on video. Users' input speech was transcribed after the user tests, the transcriptions following the transcription coding scheme agreed between NISLab and Scansoft. When processed, this data resource will be sent to Scansoft as part of our post-NICE collaboration.

2.3 Technical evaluation at system level

2.3.1 Overview

Table 2.1 provides an overview of technical and contractual (NICE Deliverable 7.1) evaluation criteria and results at system level. Section 2.3.2 explains the results in more detail.

Number	Technical and contractual criteria	Explanation	Evaluation
1	Technical robustness	Quantitative; how often does the system crash; how often does it produce a bug which prevents continued interaction (e.g. a loop)	About 12 crashes distributed over various modules. Due to their particular causes, the crashes were unevenly distributed across the 26 user sub-sessions
2	Handling of out-of-domain input	Qualitative; to which extent does the system react reasonably to out-of-domain input	Out-of-domain handling enabled for user names, nationalities, fairytale names, game names, and explanations of fairytales and games
3	**r,s** Real-time performance, spoken part	Quantitative; how long does it usually take to get reaction from the system to spoken input.	Mostly real-time. However, up to 10-14 seconds delay when the recogniser does not realise that the user stops talking and thus stays open for the max duration of 15 seconds
4	**r,s** Real-time performance, gesture part	Quantitative; how long does it usually take to get reaction from the system to gesture	The analysis of GR log files indicates that the meantime interval between the detection

		input	of a gesture (startOfGesture message produced by the GR) and the resulting message sent by the GR to the GI module was 47 ms (13093 ms / 281 GR frames). Furthermore only one user among the 13 users mentioned a small delay in the processing of gesture.
5	Barge-in	Quantitative; is barge-in implemented	No barge-in. The intended environment of use in museums is considered hostile to barge-in
6	Number of characters	Characters in the game	One (HCA)
7	Number of emotions which can be expressed by characters	Quantitative; how many different emotions can be conveyed in principle	Four: neutral, happy, sad, angry
8	Actual emotion expression verbally and non-verbally	Quantitative; how many different emotions are actually conveyed verbally and non-verbally	Verbally: neutral, angry, happy, sad Non-verbally: neutral, angry, sad
9	**s** Number of input modalities	Quantitative; how many input modalities does the system allow	Three: speech, 2D gesture, keyboard (arrow keys and F2)
10	**s** Number of output modalities	Quantitative; how many output modalities does the system allow	Six: speech, lip movements (visual speech), facial expression, hand/arm gesture, gaze, autonomous locomotion
11	Synchronisation of output	Qualitative; is output properly synchronised	Yes, except for a slight delay in onset of lip movements
12	Number of domains	Quantitative; how many domains can HCA talk about (his life, his fairytales, etc.)	His works (mostly his fairytales), his life, including childhood games and games users like, his physical and personal presence, his study including the objects in there, the user, and generic input including meta-communication
13	Number of different plots/scenes available	Quantitative; how many different plots/scenes can the user choose among	N/A

Table 2.1. Technical and contractual evaluation criteria and results at system level. In Column 2, “r” means revised formulation of a D7.1 criterion, “s” means a split of a D7.1 criterion into several distinct criteria.

2.3.2 General system performance and contractual achievements

The HCA PT2 system generally performed in close-to-real time. The only exception was that the speech recogniser sometimes failed to notice end of the input speech, causing the recogniser to continue to wait until its programmed timeout before passing on what was recognised to the natural language understanding module.

The system performed rather robustly during the user tests. Thus, 8 hours of testing produced about a dozen crashes. No particular module was particularly crash-prone. Some crashes happened in the rendering engine, some in response generation, the character module, natural language understanding, and in speech recognition. In particular, after some recogniser and natural language understanding module crashes with a particular user who spoke very long input utterances, we had to increase memory in several modules, fixing the problem. Another, non-crash-producing, problem occurred with a graphics card. To facilitate speech recognition, we had switched off the card's ventilator. Card overheating then caused HCA and part of his study to become transparent from time to time. After ventilator re-activation, this problem went away but it did not go unnoticed by the users exposed to it.

Summarising, the HCA PT2 system, and apart from the exception concerning output state implementation mentioned at the end of Section 2.2.1, showed itself to have been rather thoroughly module tested, integration tested, and debugged prior to the user tests.

With respect to barge-in, number of characters, number of input and output modalities, and number of domains there are no special comments. On these points we have achieved what was planned.

The notoriously difficult handling of out-of-domain input is available regarding user names, nationalities, fairytale names, game names, and explanations of fairytales and games. For example, HCA can recognise the titles of many of his fairytales although he does not know anything about them yet. However, recognising the title enables him to show the user that he understood what was said and explain that this is not something he can talk more about. In another example, HCA can successfully pretend to understand a user's explanation of the game of football.

The synchronisation of output is reasonably adequate and the number of emotions which can be expressed is three-to-four, as planned. However, the actual expression of emotion is limited. It is basically impossible to get to see HCA looking happy as opposed to friendly. When looking carefully at the eye brows one may observe a change from neutral to sad or to angry.

2.4 Technical evaluation at component level

2.4.1 Overview

Table 2.2 provides an overview of technical evaluation criteria and results at component level. Further explanations of the results are given in the following Sections 2.4.2-2.4.12.

Number	Technical component evaluation	Evaluation
	Speech recogniser	
1	**n** Perfect input utterance recognition	Danish group = 4 gender-balanced target group users randomly chosen from among the 13 users in the user test having Danish as their first language, test condition 2: average 23% English group: 4 new target group users, gender-balanced, having English as their first language, test condition 2: average 33%
2	**n** Understanding of user input	Danish group: 49% English group: 60%
3	**n** Understanding of user input + handling of non-understood input	Danish group: 85%

	through meta-communication	English group: 87%
4	Word error rate for English	Average both test conditions = 70,73% Test condition 1 = 80,09% Test condition 2 = 61,38%
5	Vocabulary coverage for English	Out-of-vocabulary words 2,5%
	Perplexity of English language model	Not available
6	**r** Real-time performance	In principle yes, but delays were sometimes caused by the recogniser remaining open for 15 seconds although the user stopped speaking earlier
	Gesture recogniser	
7	Recognition accuracy regarding gesture type	Blind labelling of logged gesture shapes led to the evaluation of 87.2 % of correct recognition of gestures (245/ 281). Several noisy shapes were observed.
8	Number of recognition failures	36/281=12.8% of the gesture shapes were not classified in the same class by blind labelling and by the GR module
	Natural language understanding	
9	Lexical coverage, English	Not available
10	NLU robustness, English	Perfect recognition, all thirteen users = 27% Understanding robustness = 34% Utterances understood = 47%
11	Topic spotter error rate, English	N/A. No topic spotter needed in PT2 due to its ontology-based design
12	Anaphora resolution error rate, English	N/A. No anaphora resolution in PT2
	Gesture interpretation	
13	Selection of referenced objects error rate	Failure in processing of gesture-only input for referenceable objects involved the GI in only 4% of the cases
	Input fusion	
14	Robustness to temporal distortion between input modalities	21 errors in the processing of multimodal behaviours which were due to unexpected delays between speech and gesture. They account for 43% of multimodal errors. 85% of these 21 errors were due to delays in start of speech which proved inappropriate when compared to the video.
15	Fusion error rate	40% of multimodal behaviours from an interaction point of view (75% from the point of view of the IF fusionStatus). For the processing of gesture-only behaviours, 13 cases were merged with wrong detection of speech
16	Cases in which events have not been merged and should have	48 cases amounting to 75% of all multimodal error cases
17	Cases in which events have been merged and should not have	25% of all multimodal error cases
18	Recognised modality combination error rate	Not considered relevant for evaluation

	Character module	
19	Meta-communication facilities	User input facilities: repeat, correct, clarify System output facilities: repeat, rephrase, change topic, end conversation, i.e. a graceful degradation chain of context-dependent outputs, Kukbox, specific handling of why, where, and when questions System-internal facilities: low speech recognition confidence score, high speech recognition confidence score
20	Handling of initiative	Fully mixed initiative. The user can take the initiative any time s/he wants and the system will follow
21	Performance of conversational history	Distributed discourse context and domain context histories in the character module. The former ensure graceful degradation to user input, appropriate reaction to repeated insults, and ability to remember the latest output. The latter ensure that HCA will not on his own initiative say the same thing twice and that certain implications of user input are taken into account
22	Handling of changes in emotion	HCA’s emotional state is updated for each user input
	Response generation	
23	Coverage of action set (non-verbal action)	130 out of 150 available non-verbal behaviour primitives used
	Graphical rendering (animation)	
24	Synchronisation with speech output	Eleven visemes used
25	**s** Naturalness of animation, facial	Up to 5 non-verbal primitives are used per output turn out of 74 available
26	**s** Naturalness of animation, gesture	Up to 17 non-verbal primitives per output turn out of 50-60 available primitives for gesture
27	**s** Naturalness of animation, movement	Used in scripts. A script contains up to 40 lines of behaviour descriptions
	Text-to-speech	
28	Speech quality, English	Good
29	Intelligibility, English	Good
30	Naturalness, English	Fairly good, missing pauses in some places, prosody jumps and mispronunciation of homographs
	Integration	
31	Communication among modules	Good, except for the missing integration of NCA, CF and CA
32	Message broker	Works well
33	Processing time per module	Real-time, except when the recogniser remains open for 15 seconds although the user has stopped speaking; this results in perceived delays in answering

Table 2.2. Technical evaluation criteria and results at component level. In Column 2, “n” means new criterion, “r” means revised formulation of a D7.1 criterion, “s” means a split of a D7.1 criterion into several distinct criteria.

2.4.2 Speech recognition

In the pre-PT2 user test speech recogniser development phase, we had obtained the following ”ideal scores”:

- perfectly recognised sentences, all understood = 65%
- understanding robustness, i.e., recovery of user input utterance meaning from imperfectly recognised utterances through post-recognition processing = 21%
- total input utterances understood = 86%

These impressive figures were obtained with four colleagues who had trialled with all five PT2 recogniser (pre-) versions, always reading aloud the same 60 input sentences. These subjects would rarely produce disfluencies at all and would tend to only misread the trial sentences when their syntax was arguable.

Compared to the ideal scores just described, we found a very different picture in the PT2 user test with the Danish group, as shown in Table 2.2. The contrasting figures are: perfect recognition: 65% vs. 23% and total understood: 86% vs. 49%. In a control study, we had an English user group (also children) do exactly the same test as did the Danish user group. Here, the contrasting figures are: perfect recognition: 65% vs. 33% and total understood: 86% vs. 60%.

What we could not compare between the “gold standard group”, on the one hand, and the Danish and English user groups, on the other, were (i) understanding robustness and (ii) the effects of the system’s meta-communication facilities. We could not compare (i) understanding robustness since the Danish and English user groups, but not the gold standard group, used the full system, including gesture-only input and combined gesture/speech input. And we could not compare (ii) the effects of the system’s meta-communication facilities, since only the Danish and English user groups could make use of these facilities.

Still, the average of 60% understood user inputs in the English group seems to us encouraging and sort of an in-a-nutshell illustration of the current performance of the HCA system. When adding the power of the system’s meta-communication facilities, we obtain 85% for the Danish user group and 87% for the English user group. Even though further in-depth analysis of this data is required and will be done, it may already be said that a significant fraction of the meta-communication was needed in cases in which the users produced out-of-domain input. This implies that this fraction can be significantly reduced by the relatively simple measure of adding to HCA’s knowledge. More speech data for recogniser training and better language modelling will serve to further reduce the need for meta-communication, turning the users’ input into something which the system will simply be able to understand.

The word error rate is clearly higher in the first test condition (80%) than in the second one (61%). In the first test condition, the number of insertions, deletions and substitutions are closely similar. In the second test condition, the number of insertions and deletions are closely similar while there are considerably fewer deletions.

In the 26 sub-sessions, the 13 Danish users produced 50 out-of-vocabulary words, or 2,5% of the recogniser’s 1977 word form vocabulary. This figure does not count two kinds of word forms, i.e., user names and country names, of which there were 15 in total. It is obviously senseless to try to include all manner of first names in the recogniser vocabulary and close-to-senseless to include the more than 200 names of the countries of the world. In any case, the system’s out-of-domain functionality is geared to exactly handle these kinds of words when the users tell HCA their names and countries of origin.

2.4.3 Natural language understanding

Rather than presenting figures on natural language parsing robustness per se, Table 2.2 presents the following three figures which may be compared with the "gold standard" in Section 2.4.2:

- perfect recognition, all thirteen users = 27%
- understanding robustness = 34%
- utterances understood = 47% (perfectly recognised utterances understood + imperfectly recognised utterances robustly recovered

Not surprisingly, the first and third figure above closely resemble those reported in Section 2.4.2 for a sub-population of the thirteen Danish users. However, the point here is the understanding robustness figure of 34% which shows the percentage of imperfectly recognised input sentences that were correctly understood by the natural language understanding + conversation mover sub-system. This figure shows that the HCA PT2 system is able to recover from around one third of the misrecognitions made of the users' spoken input.

The combination of the NLU and conversational mover provides an appropriate move, which determines what the user will be listening to. To determine the degree of robustness of our approach, we classify the user utterances into recoverable and irrecoverable categories. The irrecoverable category means that speech recogniser errors have distorted the utterance to such an extent that the meaning of the utterance would not be clear to a human. For example when the user says "*tell me about one of your fairytales*" the recognised utterance may be "*me about it was your fairytales*". The recoverable category specifies that despite speech recognition errors a human being would be able to retrieve the meaning of the input sentence. For example when the user says "*can you tell me something about the fairytale the little mermaid*" and the recognised utterance is *"did you tell me something about it to be the fairytale the little mermaid"*.

2.4.4 Methodology for evaluation of gesture and multimodal input

Only the gesture-only and the gesture-combined-with-speech behaviours were analysed with the videos and the log files. The videos were used to annotate the real behaviours displayed by users: spoken utterance, the objects the user gestured at (including each non referenceable object which is represented only as a noObject in the GI log file), obvious or possible misuse of tactile screen (in case the corresponding gesture was not detected by the GR). The log files were used to check the output of each module, to compare it to the observed behaviour from the video and to classify reasons and cases of failures.

We made a difference between the success of interaction and the success of the processing done by the gesture and multimodal modules. Interaction was considered as successful if the system answered correctly to the user's request. Module success was evaluated by comparing the user's behaviour and the output produced by the modules in the log files. In some cases, the interaction revealed successful although the output of the module was not correct (this means that the module's error was balanced by other means or modules). In some other cases, the interaction revealed to be not successful although the output of the module was correct (this means that an error occurred in other modules). Interaction success for multimodal input provides information on the use of inhibition and timing strategies which enabled a proper management of some redundant multimodal cases via the processing of only one of the modalities.

The following measures were collected in order to evaluate the criteria described in the evaluation overview:

- Gesture only behaviours: Number of
    - gesture only behaviours observed in the video

    - interaction success for gesture only behaviours
    - interaction failure for gesture only behaviours
    - GI success for gesture only behaviours
    - GI failure for gesture only behaviours
- Multimodal behaviours: Number of
    - multimodal behaviours observed in the video
    - interaction success for multimodal behaviours
    - interaction failure for multimodal behaviours
    - IF success for multimodal behaviours
    - IF failure for multimodal behaviours
    - cases which should not have been merged into "ok"
    - cases which should not have been grouped into "none"
    - cases which should have been merged and have not
    - cases with adequate management of concurrency

Furthermore, in order to evaluate the components, the explanations for each failure (as revealed by the video or the log files) were collected. The result is the following list of failure reasons:

- *Gesture Not Detected*: when the GR did not detect the gesture,
- *Not a Referenceable Object*: when the user gestured at a graphical object which the system does not know about and can not talk about (books, papers…),
- *Gestured Object Not Detected*: when the user gestured at a referenceable object and this object was not detected by the GI,
- *Input Inhibited*: when the user gestured during the inhibition phase (HCA was already speaking or preparing to and the processing of input gestures in GI was thus inhibited),
- *System Crash*: when there was obviously a system crash involving other modules (e.g. in the video HCA does not speak / move anymore, or the system is restarted by the experimenter),
- *Speech Recognition Error*: when an error in speech recognition led to an error in IF (e.g. deictic not detected which led to a "none" fusionStatus instead of a "ok" fusionStatus) or to an error in GI (e.g. in user 5, the speech recognition module detected "what do not" while the user was gesturing although in the video it appears that the user did not speak; the IF thus merged the GI frame and the NLU frame into an if Frame with a "none" fusionStatus which led to the absence of answer from the system to the gesture only behaviour of the user),
- *Timer Too Small*: when the semantically related gesture and spoken utterance were not grouped by the IF because the time interval between speech and gesture was longer than expected,
- *Unexplained Reason*: when neither the video nor the log files could explain the reasons for failure in the processing of gesture or multimodal input.

When some failure involved several of these reasons, we only considered in the statistics the one from the system point of view involving the GR, GI or IF and which led to error, for example in the case of a gesture on a non referenceable object while input was inhibited (e.g. user #6 phase 1), the gesture is not interpreted as the GR is blocked until the CM sends a message, thus we annotated such an error as "Input Inhibited" and not as "Not A Referenceable Object".

2.4.5 Gesture Recognition

281 gesture shapes done by the users on the tactile screens were logged during the user tests[1]. The average number of gestures per session is 14.

The logged gesture shapes have been manually labelled without displaying the result of shape recognition by the GR (blind labelling). In order to enable a fine grained analysis of gesture shapes, the labelling was done on 25 categories (Figure 2.4). At the end of the manual labelling process we found that 87.2% (245) of the logged gestures had been assigned the same category by the GR and by the manual labelling process. The fine-grained categories reveal a high number of diagonal lines (90/281=32%) and explicitly noisy categories (44/281=16%) such as garbage, noisy circle, and open circle of various orientation (Figure 2.5). The distribution of shapes in both the GR and the manual labelling are similar (Figure 2.6).

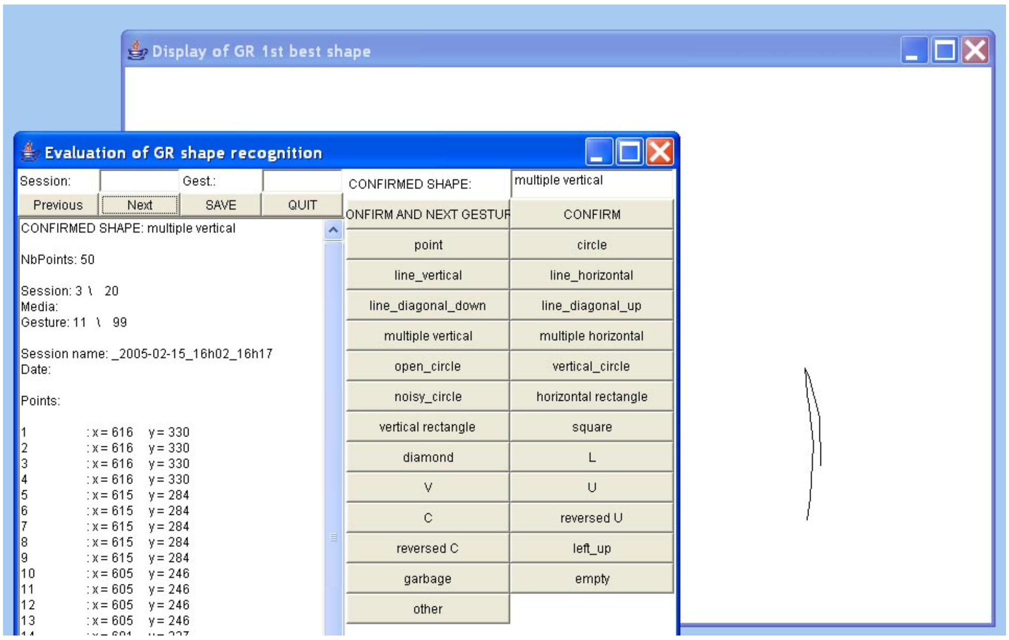

Figure 2.4. Manual blind labelling of logged gesture shapes. The gesture done by the user and logged by the system is displayed (right-hand side). It is labelled manually by one of 25 fine grained categories (left-hand side). This one was labelled as "multiple vertical" lines. It was indeed recognised as surrounder by the GR. On the 281 logged shapes, 36 were not assigned compatible categories by the blind labelling process and by the GR.

[1] For technical reasons, the GR log files were lost for 3 users leaving 20 sessions for ten of the users.

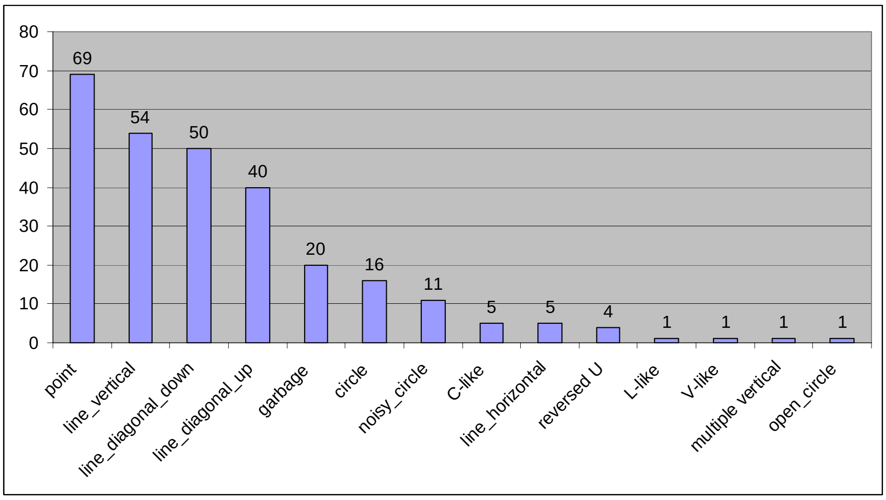

Figure 2.5. Results of manual labelling with fine-grained categories.

Figure 2.6. Quantitative comparison of shapes recognised by the GR (left-hand side) vs. manually blind labelled (right-hand side).

2.4.6 Gesture Interpretation

As observed in the videos, the users made 186 gesture-only turns. If we use the number of ifFrames (957) as an evaluation of the number of user turns (this is not exact as sometimes speech or gesture might not be detected, and some other times a single spoken turn might be divided into several recognised utterances), gesture only turns correspond to 19% of the user turns.

187 messages were produced by the GI module. The difference with the 281 gestures detected by the GR and the log files reveals that 102 gestures (102/281=36%) have been either grouped since they were on the same object or cancelled due to the fact that HCA was speaking or preparing to speak.

The repartition of the gesture interpretation categories is the following (Figure 2.7): 125/187=67% detected a single gestured object, 61/187=33% did not detect any object and only one detected several objects in a single gesture (User10/phase2: feather pen and picture of HCA's mother) which was revealed to be gesture only in the video.

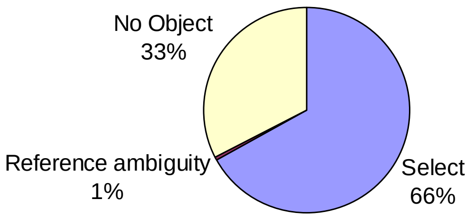

Figure 2.7. Categories of gesture interpretation output results.

We observed nearly only gestures in order to get information on objects. Only one of the users tried to move HCA (user 1) by gesturing towards the desk and saying "please go there".

On the average, each user gestured at 11 referenceable objects and 4 non referenceable objects. The users gestured between 6% and 89% (average 61%) of the 18 referenceable objects. Only three users are below 61%. 45 non referenceable objects were gestured at. The most frequently gestured non referenceable objects were papers (15/45=33%), the books (12/45=27%) and HCA's hat (10/45=22%).

User	Number of gestured and referenceable objects	Number of gestured and NON referenceable objects	% of the 18 referenceable objects
1	11	4	61
2	16	13	89
5	12	2	67
6	1	3	6
7	1	0	6
8	10	2	56
9	16	7	89
10	16	8	89
11	16	4	89
12	13	0	72
13	11	2	61
TOTAL	123	45	

Regarding interaction, 51 % of the gesture only behaviours led to an interaction success. Studying the output of the GI module, it reveals that indeed the GI module successfully detected an object for 56% of the gesture only cases.

NB	%

Gesture only behaviours	186	
Interaction success gesture only	94	**51**
Interaction failure gesture only	92	49
GI success gesture only	104	**56**
GI failure gesture only	82	44

The error reasons have been counted for the GI failures instead of interaction failures.

7 correct cases of GI processing during gesture only behaviour have not led to interaction success: user 10 phase 2 (1 erroneous detection of speech, 1 input inhibited), user 11 phase 1 (3 system crash, 1 unexplained), user 11 phase 2 (1 input inhibited). No cases were observed where interaction was successful although GI was erroneous. The following reasons of GI failures for the processing of gesture only behaviours were collected from the study of video and log files.

	NB	%
Not A Referenceable Object	52	62
System Crash	14	12
Input Inhibited	17	14
Unexplained Reason	3	4
Gestured Object Not Detected	2	2
Gesture Not Detected	1	1
TOTAL	**84**	**100**

2.4.7 Input Fusion

As observed in the videos, the users made 67 multimodal turns. If we use the number of ifFrames as an evaluation of the number of user turns, multimodal turns correspond to 7% of the user turns.

1154 messages were logged by the IF including speech only, gesture only and multimodal with the following repartition of fusion status.

Fusion Status	NB	%
ok	21	2
none	1132	98
inconsistency	1	0
TOTAL	1154	100

As revealed by the study of the videos, the only plural/singular inconsistency detected by the IF was due to a speech recognition error (user9/phase1 gestured on pictureJennyLind and asked "what is this picture about" which was recognised as eighteen).

Regarding the multimodal behaviours displayed by the users, we also analysed both the interaction success and the IF success. In only one turn was the IF successful and the interaction was not (user 13 phase 2) due to a system crash. In 24 multimodal turns, the IF was not successful but the interaction was successful: user 8 phase 1 (2 speech recognition errors where the deictic was not recognised), user 9 phase 1 (4 speech recognition errors where the deictic was not recognised, 5 timer too small), user 9 phase 2 (5 timer too small), user 12 phase 1 (1 speech not detected), user 13 phase 1 (1 Not Referenceable Object, 1 speech not detected, 5 timer too small).

Regarding interaction, 60% of the multimodal behaviours led to interaction success. Studying the output of the IF module, it reveals that it worked well for 25% of the multimodal cases.

	NB	%
Multimodal behaviours	67	
Interaction success multimodal	40	**60**
Interaction failure multimodal	27	40
IF success multimodal	17	**25**
IF failure multimodal	50	75

The following reasons for failure of the processing of multimodal behaviours were collected from the study of video and log files.

	NB	%
Timer Too Small	21	43
Speech Recognition Error	9	18
Input Inhibited	6	12
Not A Referenceable Object	4	8
Gesture Not Detected	4	8
System Crash	2	4
Unexplained Reason	2	4
Gestured Object Not Detected	1	2
TOTAL	**49**	100

A closer analysis of the “timer too small” cases was done. This involved user 11 (1 case), user 9 (14 cases), user 13 (6 cases). In 85% of these 21 cases, the timestamp of the startOfSpeech message was evaluated as inappropriate as compared to the start of speech observed in the video (12/14 cases of user 9, 6/6 cases of user 13). It would have been inappropriate to have the user wait for such a long period (e.g. 10 seconds in several cases). For example (user 13, phase1), the *startOfSpeech* would be logged as arriving 14 seconds after the startOfGesture although in the video the user starts to speak only 1 second after the start of gesture. Indeed, given the limited semantics of gesture involved in the scenario (e.g. only selection of objects) and the frequent redundancy of speech and gesture in such a conversational context, the strategy to take an early decision for gesture only behaviour after having waited a short while, sending it to the following modules, and then ignoring any incoming spoken utterance during HCA’s answer enabled to get a high rate of interaction success (60%) for multimodal behaviour while avoiding the user waiting too long for the system’s response.

The second most frequent reason for fusion failure was speech recognition error (user 8: 2 cases, user 9: 3 cases, user 12: 1 case, user 13: 3 cases). 13 wrong detections of speech also occurred (user 1: 1, user 5: 2, user 10: 11) and led to misrecognition of gesture only behaviour as multimodal. The most frequent cases appeared with user 10 who did not speak at all but speech was nevertheless detected.

These led to the following cases of errors of multimodal fusion:

	NB	%
Should not have been merged into ok	3	5
Should not have been merged into grouped none	13	20
Should have been merged into ok	48	75
TOTAL	64	100

Furthermore, 4 turns of the 67 multimodal turns were concurrent (e.g. speech and gesture were synchronised but semantically not related). This is 6% of the multimodal turns. 3 of them involved a user answering the previous question from the system and hence could be detected with dialogue history. None of them contain a deictic. Gestures during such concurrent multimodal input have not been considered in our statistics as gesture only but rather as cases of multimodal behaviour. For each of these concurrent cases, it has been counted if the system was able to detect the lack of semantic relations and not merge them. These 4 cases are:

- User 1 answered "no" to the system's previous utterance while gesturing on little_mermaid in order to get information about this picture.
- User 8 said "Denmark" to answer the system's question while gesturing on pictureColosseumRome to get information about this picture.
- User 12 said "ok" related to the previous turn and gestured at papercutout_2 to get information about it.
- User 8 said "where do you live" while gesturing on the featherPen.

2.4.8 Conversation management

The main limitation of the character module's conversation management which can be noticed in the user test data, is the limitation of the training corpus used for training the Conversation Mover, i.e., the module which turns conceptually expressed user input from Input Fusion into conversational move suggestions for the Move Processor to use. Too often, the limitation mentioned turned only slightly misrecognised input utterances into input which the Conversation Mover could not handle properly, forcing the system to activate its meta-communication facilities in order to generate an appropriate response to the user. Clearly, a larger Conversation Mover training corpus could significantly improve the robustness of the Conversation mover.

The system's input *meta-communication* facilities may represent a mild case of over-engineering for 10-18 year olds. Thus, pending a fully detailed analysis of the user test data, the young users hardly ever correct HCA's understanding of what they just said, or demand clarification from HCA. The input and output meta-communication facilities themselves appear to have worked well in general. When analysing the flow of each user-system conversation, only little potential evidence is found of non-crashing malfunctions of the input and output meta-communication facilities. Given that, we have not judged it important at this point to try to trace through the logfiles the few potential anomalies found in the data. The only exception is the specific handling of why, where, and when questions. For instance, if the user asks HCA when his mother was born, HCA is supposed to reply "I don't know when". For those questions, the data show a number of malfunctions. Thus, in a number of cases, he resorts to general meta-communication and graceful degradation instead of providing the more specific type of response just illustrated. The user would be most unlikely to notice the difference anyway. The cause of those malfunctions is the one described in the previous paragraph, i.e., the limited training material used to train the Conversation Mover.

Pending detailed analysis, preliminary analysis of the user test data shows that the *handling of initiative* works well in HCA PT2. In general, HCA takes the initiative in conversation when the user does not do so. When the user takes the initiative to talk about any topic within the system's current domains, the system follows the user no matter if the user chooses to change topic and domain or to stick to the current topic of conversation. This assumes, of course, that the user's input, whether speech, gesture, or both, is understood in the first place.

Judging from the transcribed user test data, the distributed *discourse and domain histories* appear to have worked fine in general. When analysing the flow of each user-system conversation, only little potential evidence is found of non-crashing malfunctions of the histories. Given that, we have not judged it important at this point to try to trace through the logfiles the few potential anomalies found in the data.

The system's handling of the character's *emotional state* is functionality (an Emotion Calculator) deeply embedded in the character module. The Emotion Calculator updates HCA's emotional state (i) whenever the user's input has an impact on HCA's personality. If not (ii), HCA's emotional state nudges back towards his default state of friendliness. This sub-system worked according to specification in the pre-user trial stand-alone test. However, the only way for the user to judge HCA's emotional state is from his spoken output and his facial expression. It is straightforward for the user to judge the former, such as when HCA expresses that he is pleased to receive praise from the user, or sad from remembering, and talking about, his mother. The user tests show that it is currently much harder for the users to judge HCA's emotional state based on his non-verbal behaviour, i.e., primarily his facial expressions. Thus, no user commented on HCA's non-verbal expression of emotion in the post-test interviews and it is, in fact, difficult to immediately and visually perceive that HCA has not become more angry or sad than he was before the user's latest input. To do so, one has to carefully notice relatively minor changes to the way his eye brows are set, in particular. We clearly need more development and testing to get HCA's non-verbal expression of emotion right.

Apart from better training of the Conversation Mover, the main limitation of the character module is its *limited knowledge*. This result was fully expected since the purposes of our NICE HCA work have not included that of equipping HCA with a large knowledge repository as one would have done in a commercial version of the system. Instead, we have focused on the basic mechanisms for handling a character's knowledge. It is rather straightforward to add more knowledge to the system.

2.4.9 Response generation

For non-verbal response generation about 150 behaviour primitives are available. Of these a total of about 130 primitives are used in the HCA system. Those which are not used are primarily primitives which have a close look-alike form among those which *are* being used. The primitives are partly used in scripts which define HCA's movements, and partly used to define gestural and facial behaviour accompanying his speech. The number and kind of tags used to define his gesture and face when speaking, depend on his emotional state. Thus, in general, output is defined in four different versions together with the spoken output, so that, when the system is running, the version corresponding to his actual emotional state can be chosen.

2.4.10 Animation

The animation part of the system renders the non-verbal behaviours sent by the response generator.

For lip synchronisation 15 visemes are available. Of these eleven visemes are used. Those which are not used include, e.g., mouth overly open and mouth rounded closed. Lip synchrony works fairly well although with a slight delay in the onset of visible speech compared to audible speech.

For facial animation a total of 74 primitives are available but only up to 5 are used per turn (excluding the visemes). It should be considered to use more primitives per turn to better express HCA's emotional state and to make him more lively. The same consideration applies to gesture. Up to 17 primitives out of 50-60 available ones are used per output turn. Some of the max 17 primitives may very well be identical. Thus HCA's gestures are less lively (and less life-like) than they would ideally be.

In the Non-Communicative Action (NCA) output state (no visitor) and in the Communicative Function (CF) output state (attention to the user's input), scripts are being used. Each script consists of up to forty lines of behaviours where each line is either a single or a complex animation. If the complex animations are expanded, the scripts used may contain up to about 320 lines of single animations. As already indicated, the CF output state is not integrated, and not integratable as it stands, with the two other output states and can only be viewed as stand-alone. The test users did not receive a demonstration of HCA's behaviours in the CF output state. The NCA output state is not adequately integrated with the main (Communicative Action) output state. Still, all test users were shown HCA's behaviours in the NCA output state.

A problem in the scripts is that HCA from time to time bumps into the walls in his study but apparently does not realise this and just tries to continue his walk. This is because no motion planning has been implemented for HCA for when he moves about autonomously. Another problem is that HCA's walk is quite often rather a glide in which he does not move his legs. This is a rendering engine problem which has only been partially fixed.

2.4.11 Text-to-speech

The AT&T off-the-shelf synthesis was chosen from among several other speech synthesis systems analysed by the team as being the best match for the voice of a 55 years old man. Intelligibility and voice quality were of course considered important parameters in choosing the AT&T synthesis as well. The selected synthesis is found to be both intelligible, of good quality and suitable for HCA. Naturalness does not match a human voice but is fairly good. The usual problems of prosody jumps and mispronunciations of homographs are present. From time to time, missing pause insertions in the text input to the synthesiser cause some difficult-to-understand sentence transitions.

2.4.12 Integration

The communication among modules basically worked well in the user test. Communication was done via the message broker which worked without any problems. A problem concerning buffer length for spoken input was discovered when a user spoke at length in most input turns. Otherwise, the only module communication problem we are aware of, and which was known prior to the user test, concerns the missing integration of NCA, CF and CA, as explained above.

All modules (individually and together) perform in realtime. However, the problem, mentioned earlier, in the recogniser which sometimes tended to remain open for 15 seconds although the user had stopped speaking much earlier, gives users the impression that there are sometimes delays in the system/module processing.

2.5 Usability evaluation based on user interviews

The PT2 usability evaluation is based partly on users' subjective opinions as expressed during the post-session interviews and partly on analysis of transcriptions and logfile analyses. This section focuses on the interview-based usability evaluation while Section 2.6 describes results from the analyses of logfiles and transcriptions.

2.5.1 Interview-based usability evaluation

In the PT2 user interviews, we asked a total of 31 questions. Answers to many of the questions formed the basis for evaluating PT2 subjectively and according to the criteria already determined in deliverable D7.1 and later used in a slightly modified version in the PT1 evaluation (D7.2a). Table 2.3 provides an overview of these usability criteria and the evaluation results for PT2 based on the interview data.

Number	Basic usability criteria	Explanation	Evaluation
1	Speech understanding adequacy	Subjective; how well does the system understand speech input	Fair, larger vocabulary and grammar desirable
2	Gesture understanding adequacy	Subjective; how well does the system understand gesture input	Good, but more objects should perhaps be active
3	**n** Combined speech/gesture understanding adequacy	Subjective; how well does the system understand combined speech/gesture input	Good, as long as the pointed-to object is active. Only about half of the users spoke while pointing
4	Output voice quality	Subjective; how intelligible and natural is the system output voice	Good, easy to understand
5	Output phrasing adequacy	Subjective; how adequate are the system's output formulations	Good, occasionally slightly too long output
6	Animation quality	Subjective; how natural is the animated output	Lip synchrony okay, improvements of other animation aspects needed
7	Quality of graphics	Subjective; how good is the graphics	Rather good
8	Ease of use of input devices	Subjective; how easy are the input devices to use, such as the touch screen	Easy
	Core usability criteria	**Explanation**	**Evaluation**
9	**r** How natural is it to communicate via the available modalities	Subjective; how natural is it to communicate via the available modalities	Natural to use speech and touch screen
10	Output behaviour naturalness	Subjective; character believability, coordination and synchronisation of verbal	Looks like real HCA, lip synchrony okay, display of emotions very limited, Non-

		and non-verbal behaviour, display of emotions, dialogue initiative and flow, non-communicative function, etc.	Communicative Action somewhat odd for a 55 years old man
11	**r** Ease of use of the game: How well did users complete the scenario tasks?	Subjective; how easy is it for the user to find out what to do and how to interact	Rather easy to interact with the system but somewhat difficult for several users to find out what to talk about. The problem sheet (2nd test condition) was felt to provide useful support
12	**s** Error handling adequacy, spoken part	Subjective; how good is the system at detecting errors relating to spoken input and how well does it handle them	Improvements needed
13	**s** Error handling adequacy, gesture-only part	Subjective; how good is the system at detecting and handling errors relating to gesture input	No error handling
14	Entertainment value	Subjective; this measure includes game quality and originality, interest taken in the game, feeling like playing again, time spent playing, user game initiative, etc.	Fun, good entertainment value
15	Educational value	Subjective; to which extent did the user learn from interacting with the system	Learned something, e.g. about HCA's life or English
16	User satisfaction	Subjective; how satisfied is the user with the system	Rather good

Table 2.3. Usability criteria and results measured via user interviews. In Column 2, "n" means new criterion, "r" means revised formulation of a D7.1 criterion, "s" means a split of a D7.1 criterion into several distinct criteria.

In the following we discuss details from the interviews.

Eight initial questions dealt with the user's identity, background, computer game experience and experience in talking to computers. We had no substantial input on the final question on any other comments. This leaves 22 questions about the HCA system itself and how it was to interact with it, which are presented in abbreviated form in Figure 2.8. Compared to the 16 questions about the system in the PT1 interviews, new questions in the PT2 interviews addressed matters, such as, for input, talking and pointing at the same time, for output, HCA's audiovisual speech, and, as regards conversation management, how HCA dealt with errors and misunderstandings during conversation. Question re-phrasings primarily reflected a less HCA-centric question style.

2.5.2 Interview scoring methodology

Figure 2.8 presents a quantified summary of the PT2 interview results. Each user's verbatim response to each question was scored independently by two raters on a three-point scale. The general scoring principle followed may be roughly presented as 1 = high, with minor or no qualifications, 2 = reasonable but with qualifications, and 3 = low/negative. The general scoring principle was instantiated to each interview question in a manner which took the specific contents of the question into account. Rating differences were negotiated by the two raters until consensus was reached. Finally, all user ratings per question were averaged to arrive at the summary shown in Figure 2.8. It cannot be excluded, of course, that new raters would have rated some of the user answers slightly differently, at least initially, on the basis just described. Nevertheless, despite its qualitative and judgmental nature, the methodology does provide a means of summarising large amounts of user interview data in order to build a coarse-grained profile of how an entire user population views a system and their interaction with it.

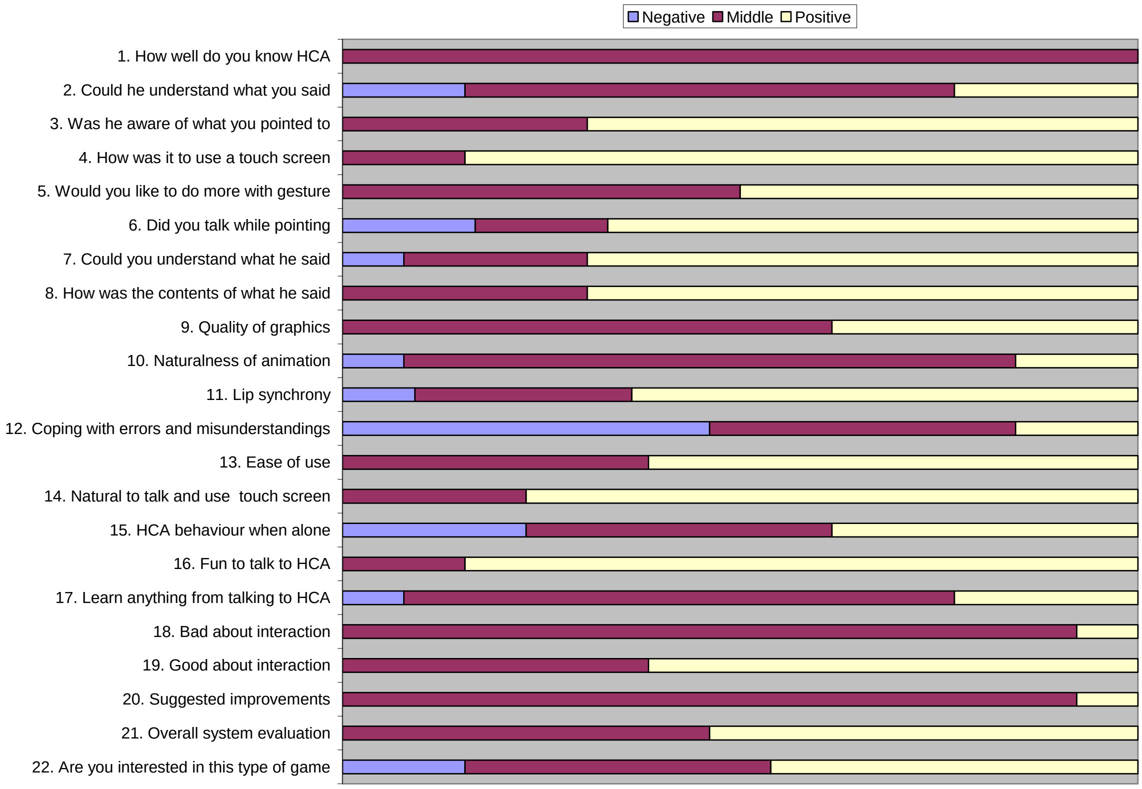

Figure 2.8. Summary of interview results from the HCA PT2 user test.

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

2.5.3 Evaluation of pointing input

As regards pointing input, users were very positive about using the touch screen (Q4). In general, users found that HCA was aware of their pointing gestures (Q3). Half of the users were happy with the 2D gesture affordances in PT2 while the other half wished to be able to gesture towards more objects in HCA’s study (Q5). In fact, several of the users gestured to one or more non-referenceable objects, as explained above. In such cases it would probably have been good if HCA could at least react by telling that he cannot tell a story about the object but that he can tell about his pictures. This is what he did in the first prototype. Only a couple of users never tried to talk and point at the same time (Q6) but the large majority of users found it natural to combine spoken and gesture input (Q14).

2.5.4 Perceived quality of graphics and animation

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

2.5.5 Perceived speech understanding capabilities

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

2.5.6 Entertainment value and learning

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

2.5.7 Ease-of-use

All users found the system easy or rather easy to use because one simply has to talk to HCA and use a few keyboard keys. Several compared the interaction favourably to various genres of computer games. Two remarked that their primary difficulty had been to find out what to talk to HCA about. One user found that there were too many camera angles.

2.5.8 What was good or bad and in need of improvement

On the issues of what is good or bad and in need of improvement, most negative points made above were repeated (Q18). New ones were some deeper inconsistencies between the user's and HCA's control of his locomotion, and between camera angle and HCA's turning towards an object pointed to. Also, HCA should have more knowledge and improved prosody, and one user felt that he takes offence too easily. Several users praised HCA's story-telling (Q19), the opportunity to have conversation with him, his “easy English” and good voice. The needs for improvements question (Q20) made the users re-emphasise some of their main messages, including more knowledge to HCA, better walk, less antics, improved understanding and asking more questions of users.

2.5.9 Overall evaluation by the users

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

2.5.10 Comparison with the PT1 user test interviews

It is impossible to make a point-by-point comparison between the two sets of user test interviews because of the significant differences among the questions asked in the two sets of interviews, differences which to a large extent reflect differences between the two system prototype versions tested. Among these differences, three major differences stand out. The first difference is that PT2, but not PT1, used speech recognition; the second, that PT2 has far better conversation management than was the case in PT1; the third difference is that PT2 has significantly improved natural language understanding compared to PT1.

Still, if we compare the, as yet unpublished figures from the PT1 user test interviews which were scored following the methodology described in Section 2.4.1 above, we find that, on average, the users scored PT1 at 1.8 whereas they scored PT2 at 1.6. Just as interestingly, the range of averaged individual scorings dropped from a variation figure of 86% for PT1 to a mere 31% for PT2. One might find the 1.8/1.6 difference, positive as it is for PT2, a relatively modest improvement in the users’ average evaluation of the system from one prototype to the next one which was completed a whole year later. However, in view of the two following facts we find this result clearly encouraging: (1) PT2 did not include speech recognition, which meant that, barring a few wizard typing errors, PT1 may be considered as having had perfect (emulated) speech recognition; and (2) in the PT2 user test, the users, all of whom were rather young and had English as a second language, were deliberately not instructed in how to interact with the system through speech at all. Despite the numerous disfluencies produced by the PT2 users - partly, it may be surmised, because of the lack of how-to-speak instruction described - and despite the ensuing speech recognition problems, the PT2 users still found PT2 better than PT1 and were in far more agreement on this verdict than were the PT1 users.

Among the main factors of improvement which counter-balance the addition of speech recognition to PT2 and the sub-optimal way in which many users chose to speak to the system, we hypothesise that the following are perhaps the most important ones: (i) the graphics bugs that were frequent in PT1 were almost absent in the PT2 test; (ii) the inflexibility of conversation management in PT1 had been virtually removed in PT2; and (iii) the rather sub-optimal (female!) output voice used in the PT1 user test had been replaced by state-of-the-art speech synthesis in PT2.

2.6 Usability evaluation based on analyses of transcriptions and logfiles

Table 2.4 provides an overview of usability evaluation criteria and evaluation results measured via other data than the interviews. In the following the results are explained in more detail.

Number	Basic usability criteria	Explanation	Evaluation
1	**s** Frequency of interaction problems, spoken part	Quantitative; how often does a problem occur related to spoken interaction (e.g. the user is not understood or is misunderstood)	Danish user group: system misunderstandings, average = 15% English user group: system misunderstandings, average = 13%
2	**s** Frequency of interaction problems, gesture part	Quantitative; how often does a problem occur related to gesture interaction	The answers to the question "Was he aware of what you pointed to and did he answer?" were all positive. The comparative analysis of the videos and the log files reveals that 51% of the gesture only behaviours were successful from an interaction point of view and that 62% of the failures were due to gestures on non referenceable objects.
3	**s** Frequency of interaction problems, graphics rendering part	Quantitative; how often does a problem occur related to graphics	Overheating graphics card made body parts fall off in the test with the first two users. Max five crashes due to graphics
4	Sufficiency of domain coverage	Subjective; how well does the system cover the domains it announces to the user	Coverage is insufficient for travels, modern technology, and some personal questions
5	**r** Number of objects the subject(s) interacted with through gesture	Quantitative; serves to check to which extent the possibil-ities offered by the system are also used by users	All 18 referencable objects were gestured at. Moreover, a total of additionally 16 objects were gestured at. Each user gestured at between 6% and 89% of the 18 referenceable objects (average 62%). Two users did very little gesturing.
6	**r** Average frequency of domains addressed by users in the conversation in percentage of number of turns	Quantitative; serves to check which domains users actually address and how often	User = 9.0; life = 8.1; works = 9.6; study = 15.6; hca = 5.7; generic = 51.9
	Core usability criteria	**Explanation**	**Evaluation**
7	**r** Conversation success	Quantitative; how often is a transaction exchange between the user and the system successful	See Table 2.2
8	Sufficiency of the system's reasoning capabilities	Subjective; how good is the system at reasoning about user input	Needs identified in PT1 implemented. More reasoning concerning how much has been said about a topic already would be good.

9	Scope of user modelling	Subjective; to which extent does the system exploit what it learns about the user	Very limited

Table 2.4. Usability criteria and results based on other user test data than interviews. In Column 2, "n" means new criterion, "r" means revised formulation of a D7.1 criterion, "s" means a split of a D7.1 criterion into several distinct criteria.

We are still at an early stage in analysing frequency of *spoken interaction problems*. Table 2.2 above lists the figures 85% and 87%, respectively, for understanding of user input + handling of non-understood input through meta-communication. The remaining 15% resp. 13% are system misunderstandings of the users input. It should be noted that (i) these figures do *not* include the system's handling of non-speech and non-gesture input (user silence, input noise, etc.), and that (ii) the figures *do* cover all spoken and/or gesture input. Still, the figures provide a first impression of the order of magnitude of input to which the system responded plainly inappropriately.

With respect to *interaction problems for the gesture part* these were already discussed in detail in Section 2.4.6. Only a bit more than half of the gesture interaction was successful. Most of the unsuccessful cases were due to the user pointing to non-referenceable objects in which case no feedback was offered.

The *interaction problems caused by the graphics rendering* were fairly few. A few crashes occurred and the first two users experienced HCA falling apart on the screen due to an overheating graphics card. Apart from that there were no real interaction problems but a couple of users found it weird that one could sort of get behind, or outside, the walls and curtains in one of the camera views.

The *domain coverage* is overall fairly good but in certain particular areas it is clearly insufficient. These areas include HCA's travels and modern technology, both of which domains are entirely absent. In addition, the user test data show that HCA could well use more knowledge about his own person, both in terms of the visible person on the screen and about his character and dispositions. Many users asked questions within these areas – and with good reason. It would be simple to add more knowledge and enable HCA to respond meaningfully to more of these questions.

All 18 referencable objects were gestured at. Each user gestured at between 6% and 89% of these objects (average 62%). Moreover, a total of 16 non-referenceable objects were gestured at. In particular users often gesture to the papers on his desk, his books and his hat. Two users did very little gesturing.

All domains were addressed by users but with varying frequency. The study domain is the users' favourite with 15.6% of the input turns referencing this domain. The works domain (9.6%) closely followed by the user domain (9.0%) is also quite popular. However, the main reason for the user domain being address this much is that whenever HCA believes he has a visitor he starts by asking questions about the user. He is not in a position to figure out that he already talked to this person before. Thus if the user said, e.g., "Bye" to HCA, hesitated too long in saying something, a system module crashed (apart from the animation), or if HCA got so annoyed with the user that he ended the conversation, then the user would have to start a new dialogue. The life domain (8.1%) is addressed only slightly less frequently than the works and user domain whereas input concerning the hca domain only occurs in 5.7% of the input turns. A little more than half of the user input falls within the so-called generic domain. This is input which may occur in any of the other domains and which – taken alone without any context – cannot be allocated to any particular of these domains. The generic input domain includes input such as yes, no, I don't know, repeat, and clarification.

Concerning *conversation success*, and pending more detailed analysis of the user test data, we refer to the figures in Table 2.2. Work is ongoing on a metrics and an associated coding scheme for measuring conversation success in the context of a theory of conversation (discourse) coherence.

With respect to the HCA's *reasoning capabilities*, we discovered in the PT1 user test that even if HCA had been told by the user that s/he did not know his fairytales, HCA might still ask detailed questions about the user's knowledge of his fairytales. In PT2 this reasoning problem has been solved.

The scope of *user modelling* is very limited. HCA uses the user's information on age when he asks if the user is a boy or a girl (when the age is at most 18) or if the user is a man or a woman (when the age is above 18).

2.7 Lessons learned

In general terms and as described in detail above, the PT2 user test went quite well, with only a very limited number of module crashes, yielding what seems to us to be a pretty thorough user critique, negatively as well as positively, of most aspects of the system and the interaction, and generating a rich corpus of data many aspects of which still remain to be analysed.

As for lessons learned, the primary lesson, we argue, is that the future potential of the kind of conversation illustrated by the HCA system, i.e., conversation for edutainment with famous people from our history, does seem to have been demonstrated by the user test data reported in this deliverable. The second HCA prototype proved to be mature enough for the user testing conducted; the users very much liked the experience overall; and they had a clear sense of the potential of this kind of system for edutainment and entertainment.

A second lesson learned from the PT2 user test is that there is an important problem to be overcome in providing adequate instruction on how to speak to the system for (i) walk-up-and-use (ii) young users who (iii) have English as their second language and who (iv) have not spoken to computers before to any significant extent. The PT2 user test clearly demonstrated that users with these properties had significant difficulties finding out how to successfully speak to the system. Clearly, there is an apparent conflict between, on the one hand, a system for users with properties (i) through (iv) above, which, on the other hand, is meant to be used without significant prior instruction. Of course, things will level out in the future when everybody has learned how to address speech recognisers and adapt to their individual properties, but what to do in the short term remains a challenge.

A third lesson learned concerns the animation of HCA. While the users judged the graphics to be more or less satisfactory overall, they clearly have expectations to HCA's animated behaviours which PT2 partly failed to meet. For instance, many users found his walk weird or reacted to his out-of-expected-character antics when in the non-communicative action state. Similarly, they all seem to have failed to notice his non-verbal expressions of emotion. It would also have been very useful to have been able to test users' reactions to an integrated NCA/CF/CA sub-system. In particular, the absence of the CF output state of paying attention to the user's input meant that HCA was less active during turn-taking negotiation than would have been the case otherwise. Even more turn-taking negotiation abilities would have been fine, too, of course, but was not planned. More, and more lively, gestures on HCA's part is another desideratum, one which, however, is rather labour-intensive to fulfil.

A fourth lesson is about multimodal input. Affording semantically focussed gestural interaction (selection of objects to get information about them) enabled to reach interaction success thanks to the processing of only one modality in case of recognition errors or temporal malfunction in either modality. In order to decrease interaction problems, objects which can be gestured at should be graphically salient. Furthermore, semantic and perceptual knowledge about the graphical objects revealed to be not very useful with non native English children since they might not perceive or express this knowledge properly, for example referring to the picture of a woman by saying "what is this?" instead of "who is this?". Finally, regarding the bidirectional multimodality of communication, the management of turn-taking is confirmed to be a main issue. In order to avoid having the user speak or gesture while the character is preparing to respond, the system's non-verbal cues should be ensured to be visible and understood by the user. The visibility of these non-verbal cues remain a challenge in such a 3D system where conversation needs to be combined with navigation.

Our continuing work on evaluating the PT2 user test results has now turned towards in-depth analysis of the spoken conversation and multimodal data and development of a strategy for quickly educating kids and teenagers in how to work effectively with speech recognition-based conversational systems. In the former context, we suspect that the surprisingly low average score accorded by the users to the system's meta-communication abilities is closely related to the young users' lack of instruction on how to speak to the system.

2.8 Conversation example composed from two actual conversations with HCA

okay	# # Hello there welcome to my study. # I am glad to meet you. # It is # always so interesting when I get visitors # here in my study. And I am curious to get to know # my visitors. Therefore I would like to # ask you a few questions at first. # Please tell me # your name
my name is aske	# # Okay # # How old are you?
im twelve	# When I # was your age I was a # big fan of the theatre. # Whenever there was an opportunity I went to the theatre. # It was my favourite place number one. # # I would like to know if # you are a boy or a girl
i am a boy	# # I was a curious boy. # Once I ventured into the hospital for the dangerous lunatics. # A woman sat in her cell singing with a # beautiful voice. Suddenly she got up and # ran screaming towards the door. # She reached for me and her fingertips touched my clothes. # I was half dead from fear when the attendant came. # Which country # do you come from?
i come from Denmark	# Excellent # thanks for letting me know about you.# Now let us talk. I am still far from being # the Hans Christian Andersen I once was. # Maybe you will find that there is still much I have to remember # or learn again. But dont you despair. # You can always ask me what I know or # what I can do.
how old are you	# I was # born in 1805 # and died in 1875. So my lifetime lasted # 70 years. It was a time # full of great events in the world and also for me. I am happy to # tell you about the things I remember from my life which must have been very # exciting. To tell # you the truth I really hope so because they tell me that I have written # several large books about it. For the moment # I can tell you a lot about my family and my childhood in Odense. I can # also tell you things about my youth in Copenhagen and a little bit about my adult life in the # whole wide world which I am still trying to remember.
[no input]	# # Would you like to hear about my life?
yes	# # You want to hear about my family
yes can you tell me about it	# I had a nice and small family. # Apart from my parents I only saw my grandparents regularly # and I was almost an only child. Sadly my family was not as # fortunate in life as myself. My parents were common people who had to work hard # all through their lives and look twice at every penny # you know. No # life was no bed of roses to my mother and father not to mention my grandmother and grandfather.
[no input]	# # You want to hear about my father
no thank you	# Do # you want to hear other stories about my life?
no thank you you are very old	# sorry

you are very old	# # I am not old! # As you can see I am a man in my best age. # I know many things about the world # and I can do many things or at least I used to. # I still have much hair on my head and I dont wear a wig. # I even had my hair curled lately and my beard shaven off completely. # In spite of that many people think that I look strange I know that big feet and all. # I will not speak of my teeth so dont you dare to ask
you are very old	# No no # I am not old at all. # I am only fifty five years old # that is nothing compared to two hundred years. Well # I was born two hundred years ago. # So I am wondering how I can be fifty five # especially since I was born again just a few months ago. # I wish somebody would tell me! # Maybe I will never die again?
you are ugly	# that was # not necessary i dont think # # You want to hear about my youth?
what is your favourite game	# My favourite childhood game was to # play theatre. When I was eight I # saw an opera at the theatre in German about a mermaid # which I soon played at home. I # did not speak or even understand German but I liked this opera. So I # created my own homemade German words and wore my mothers apron over my shoulders. # Now I was a knight # or the mermaid swimming on a stool. # All these scenes scared my mother she forbade me to play this game because she often believed that I had gone mad. # But I could not stop # you know.
your mum is dead	# The name of # my mother was Anne Marie Andersdatter. # Oh my poor mother! She was a washerwoman. She had her washing place # beside the Odense river. It was hard work # not even well paid and cold too especially in the winter! She had to # warm herself quite often with a drop of aquavit that became her destiny! My dear mother # died when I was twenty eight.
what is your favourite fairy tale	# # I like several of my fairytales 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 point is that if you are different # then you will get bullied.
can you dance	# sorry
can you do the moonwalk	# # i am sorry but i still dont understand. # Maybe you are talking about one of the many things i still dont know about
you are very stupid	# silence is gold but silver is # also very beautiful
what	# if # you are still out there # you could ask me what i know about if # you would like to continue our conversation
what games do you like to play	# My favourite childhood game was to # play theatre. When I was eight I # saw an opera at the theatre in German about a mermaid # which I soon played at home. I # did not speak or even understand German but I liked this opera. So I # created my own homemade German words and wore my mothers apron over my shoulders. # Now I was a knight # or the mermaid swimming on a stool. # All these scenes scared my mother she forbade me to play this game because she often believed that I had gone mad. # But I could not stop # you know.

	# 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?
i like soccer do you like soccer	# # Sorry i would suggest that we talk about something i know
i like football do you like football	# # Football. # I know a foot because I have two large feet myself. # And I know a ball as well because I wrote a fairy tale about the top and the ball. So in football you kick around a ball with your feet # is that what football is about? # You really must tell me more about this game.
are you dead	# I died on the fourth of August # eighteen hundred and seventy five.
what do you use when you make fairy tales	# # I only have limited knowledge about my fairytales yet. For example # I can tell you how many fairytales I wrote # which relation they had to my life or # maybe you want to know if I wrote any scary fairytales. # I can also tell you about a particular fairytale though for the moment # I only remember details about three of the fairytales I have written. # The ones I remember are the Little Mermaid # the Ugly Duckling and the Princess and the Pea.
is this what you use **Gesture:** featherPen	# # This is my pen. The fairy tale I am writing now is called # The Pen and the Inkstand. The pen and the inkstand are # jealous of each other. # The pen says # to the inkstand: # you dont think at all. # If you did you would see that you can only # provide the means. # You give the fluid that # I may place upon the paper what dwells in me # and what I wish to bring to light. # It is the pen that writes: # no man doubts that and indeed # most people understand as much about poetry as an old inkstand.
what do you like to do in your spare time	# sorry
what can you do	# # It is very much up to you what I can do. # When you talk to me about my life my fairy tales and my study # I sometimes can tell you some good stories # or at least give you some answers. # It is also up to you if you want to see me happy angry or sad. # Some of the things we can talk about make me happy or angry # others make me sad.
do you have two eyebrows	# # i am not sure i heard what you said
what is that **Gesture**: papercutout_2	# # I made this paper cutout for Christines third birthday. # It was part of a picture book # that I made with her grandfather. # Christine is the greatgrandchild of Jonas Collin who you can see # in one of the other pictures here in my study.
what is this **Gesture**: locomotive	# # I am so impressed how fast trains can go. They rush along at # incredibly high speed. # There is hardly time to view the landscape. # A hundred and ten kilometer long journey # lasts only three and a half hours. # Hooray for the steam machine!
where do you live now	# # I live in the heart of Copenhagen. From my windows I have a # marvellous view of the nineteenth century # street life. I live only two houses # away from the harbour. This is a good place to live # except in the winter where my little house is once in a while hit by # glacially cold storms. If only # my landlady would give away her # terrible dog this would really be a # perfect place to live. I never liked dogs # you know.

goodbye	# # Bye bye. Keep away from # fires. They are so dangerous and # scary.

3 The Fairy-tale world system

3.1 Purpose

This section presents an evaluation of the second prototype of the fairy-tale system in the NICE project. The primary purpose of the evaluation is to ensure that the goals set out at the beginning of the project have been met, namely to create a game in which spoken dialogue is the primary vehicle for narrative progression, and which is perceived as enjoyable by members of the intended target group, children between the ages of 8 and 15. Another purpose is to ensure technical progress from the first prototype, in particular as concerns the scope and robustness of input understanding.

The evaluation is based on three different data collections with a total of 57 subjects (children aged 8–15) who have used the system, been interviewed and filled out questionnaires prior to as well as following the interaction. All dialogues were recorded and transcribed, and in some cases the interaction was also videotaped. The internal communication between the different modules of the system was also logged. In two of the data collections (containing 42 subjects), the interactions between the subject and the system were video-taped using one camera capturing footage of the subject while the system's video output served as the video source for another channel, which was recorded along with the camera signal in a split-screen fashion. The experimental data thus collected constitutes the basis for this evaluation.

Since the emphasis of the evaluation was to find out whether children liked the idea of using spoken dialogue in computer games in general, and the NICE fairy tale game in particular, the evaluation is partly qualitative, based on interviews but also on analyses of the dialogues. Several quantitative measures, relevant to the interaction, have also been calculated, and correlated with the users' perceived experience of the system. For example, in order to evaluate whether "spoken dialogue is the primary vehicle for narrative progression", we need to define what is meant by "narrative progression" in this context. In the next section, we propose such a definition based on the concept of a *story-functional event*. Furthermore, we need to measure to what extent the player tried to initiate such progression, and to what extent he/she succeeded (was understood by the system).

3.2 General Requirements

The following requirements were defined in NICE Deliverable D1.1b as a guidance for the overall design of the NICE fairy-tale game:

1. Spoken multimodal dialogue should be appreciated by the player not only as an add-on but as *the primary means of progressing in the game.*
2. The primary intended users are children (who should be around nine years or older).
3. The domain should be built upon a small collection of more or less autonomous fairy-tale characters adapted from H. C. Andersen.
4. Both the overall story and its component subplots should have goals that are easily understood and can be explained in about ten seconds.
5. The introduction necessary to familiarize the player with multimodal spoken dialogue, should be part of the game, and should be fun and interesting in its own right.
6. The game should be many-off in the sense that it is the actions of the player and the characters that should influence the line of events and how the story ends.
7. The game should be an enjoyable experience regardless of whether it is run to an end or just for the duration of a plot.

3.3 Game scenario

The game begins in H. C. Andersen’s house in Copenhagen in the 19th century. Andersen has just left on a trip, and has asked one of his fairy-tale characters, Cloddy Hans, to guard his fairy-tale laboratory while he is away. The key device in the laboratory is a fairy-tale machine, which nobody except Andersen himself is allowed to touch (Figure 1). On a set of shelves beside the machine, various objects are located, such as a key, a hammer, a diamond and a magic wand. By removing objects from the shelves, putting them into suitable slots in the machine and pulling a lever, one lets the machine construct a new fairy-tale in which the objects come to life.

Just before the user enters the game, Cloddy Hans has got the idea of surprising H. C. Andersen with a new fairy-tale on his coming back. There is a problem, however: Each slot is labelled with a symbol which tells which type of object is supposed to go there, but since Cloddy Hans is not very bright, he needs help from the user with understanding these. There are four slots, which are labelled with symbols denoting “useful”, “magical”, “precious” and “dangerous” things, respectively. Which object goes in which slot is sometimes more obvious (provided you understand the symbols), like the diamond belonging in “precious”, and sometimes less obvious, like the knife belonging in “useful” rather than “dangerous”.

The first scene thus develops into a kind of “put-that-there” game, where it is the task of the user to instruct Cloddy Hans; tell him where to go, which objects to pick up and where to put them down, etc. If the user does not understand what to say, Cloddy Hans will encourage him or her, give suggestions, and eventually take matters into own hands. Because the initial scene is task-oriented in a straightforward way, the system is able to anticipate what the user will have to say to solve it. The real purpose is not to solve the task, but to engage in a collaborative conversation where the player familiarises himself with the possibilities and limitations of the spoken (multimodal) input capabilities.

In the second scene, the player enters the actual fairy-tale world, together with Cloddy Hans. The fairy-tale world is a large 3D virtual world (parts of it can be seen in Figure 2). At the beginning of the second scene, Cloddy Hans encourages the player to explore the immediate surroundings on the small island. While wandering about and looking around, the player discovers that the objects that were put in the fairy-tale machine in the preceding scene are now lying scattered in the grass. Although it is not completely clear to the player at this point, these objects will actually constitute valuable assets when solving various tasks in the world.

Figure 3.1. The first scene: Cloddy Hans standing beside the shelves with objects, and in front of the fairy-tale machine.

Figure 3.2. The second scene: A small part of the fairy-tale world. The player and Cloddy Hans start off on the small island on the left hand side.

The player soon encounters the first problem. Together with Cloddy Hans, he is trapped on a small island, from which he can see the marvels of the fairy-tale world – houses, fields, a windmill, and many more things – but they are all out of reach. A deep gap separates him from these wonders. There is a drawbridge, which can be used for the crossing, but it is open, and the handle that operates it is on the other side. Fortunately, a girl, Karin, is standing on the other side (Figure 3.3).

Drawbridge in initial state | Drawbridge in end state

Figure 3.3. The second scene: Cloddy Hans and Karin at the gap and the open drawbridge.

Karin has a different kind of personality compared to Cloddy Hans. Instead of having Cloddy Hans's positive attitude, she is sullen and uncooperative, and refuses to wind down the drawbridge. The key to solving this deadlock is for the player to find out that Karin will comply if she is paid: she wants to have one of the fairy-tale objects that are lying in the grass on the player's side of the gap (which object she wants will change each time the game is restarted). Thus, it is the task of the player to find the appropriate object, make Cloddy pick it up, and use this object to bargain with Karin. The second scene ends when Karin lowers the bridge, allowing Cloddy Hans (and the user) to pass.

In the second scene, Cloddy Hans and the user find themselves on a rather small island, along with all the objects they previously chose to put in the fairy-tale machine. The island is separated from the mainland by a drawbridge, guarded by Karin, who has deliberately been designed to differ from Cloddy Hans in terms of personality, as conveyed by both her verbal and non-verbal behaviour. Karin will only lower the drawbridge when offered something she finds acceptable in return, which she never does until the user’s third attempt, thereby encouraging negotiative behaviour. Furthermore, both Cloddy Hans and Karin openly show some amount of grudge against each other, with both characters occasionally prompting the user to choose sides.

3.4 Narrative progression

The two scenes described above contain certain key moments, which we will call *story-functional events*. We will take the passing of such an event to mean that there has been a progression in the story (thus it is important that a story-functional event can not be undone).

The first scene contain the following types of story-functional events:

1. Cloddy Hans introduces himself
2. Cloddy Hans introduces the plot
3. Cloddy Hans has picked up an object for the first time
4. Cloddy Hans has dropped an object in the fairy-tale machine
5. Cloddy Hans pulls the lever so that he and the user can enter the fairy-tale world

Note that type number 4 above can have several different instantiations (drop the axe in the machine, drop the sword in the machine, etc.). Since it is impossible to retrieve an object from the machine, all these events are story-functional.

The second scene contains the following types of story-functional events:

1. Cloddy Hans informs the user about the fairy-tale world
2. The user meets Karin for the first time
3. Cloddy Hans gives his opinion of Karin
4. Karin gives her opinion of Cloddy Hans
5. Karin informs the user that she demands payment in order to lower the drawbridge
6. The user offers Karin an object as payment.
7. Karin lowers the drawbridge
8. Cloddy Hans crosses the drawbridge

3.5 The fairy-tale characters' different personalities

An important role of the fairy-tale characters' multimodal output is to convey their different personalities. Personality is conveyed by modes of appearance, voice quality, choice of words and actions. There are a number of psychological models of personality, one of the most used is the OCEAN model that has the following dimensions: *Openness, Conscientiousness, Extroversion, Agreeableness,* and *Neuroticism* (McCrae & Costa, 1996). Personality traits are not explicitly modelled in the NICE system, but rather used as guidance in the design of the characters. This means that there is not a set of rules that determines the characters' output behavior from personality parameters. Instead the personality descriptions are used as a tool to get consistent character behaviors that are perceived by the users as compatible with the intended personality of the characters.

The two main characters in the second prototype of the fairy-tale game are the helper Cloddy Hans and the gatekeeper Karin. In order to match their different roles in the game, the output behavior of these characters have been designed to display these quite different personality traits:

	Openness	**Conscientiousness**	**Extroversion**	**Agreeableness**	**Neuroticism**
Cloddy Hans	Dunce	Thorough	Quiet Uncertain	Friendly Polite	Calm Even-tempered
Karin	Intellectual	Frivolous	Outspoken Self-confident	Unfriendly Touchy	Anxious Over-emotional

Table 3.1. The OCEAN personality traits of the two main characters in the second scene.

The two characters' dialogue rules, wording of utterances, speaking styles and non-verbal behaviour have been designed to match their respective personalities.

3.6 Data collection

3.6.1 Method

For the purpose of data collection, the system was run in *supervised mode*. This means that in each turn, the system computed an n-best list of possible Cloddy Hans utterances, and an n-best list of possible actions (animations). The human operator then had several options:

1. Go with the system's primary suggestion by pressing a "Send" button.
2. Select some suggestion lower down the n-best list by clicking on that suggestion and pressing "Send".
3. Select predefined utterances and actions from a number of pop-up menus.
4. Edit the speech recognition result (which was displayed in a window) and let the system reanalyze. This option was particularly useful in the cases where some crucial word in the input had been misrecognized, leading the succeeding processing astray.
5. Type in an utterance to be synthesized in a free-text window.

Obviously, alternative 1 was faster than alternative 2, which in its turn was faster than alternative 3, and so forth. In order to facilitate a real-time conversation, the operator only rarely used the alternatives 4 and 5.

A supervised system, such as the one described above, thus represents a middle ground between a fully automatic system on the one hand, and a completely simulated system (full Wizard-of-Oz) on the other. In fact, for sophisticated systems like the NICE fairy-tale system, complete simulation is out of the question. To give the operator the slightest chance of holding up the system's end of the conversation, a high degree of automation is necessary.

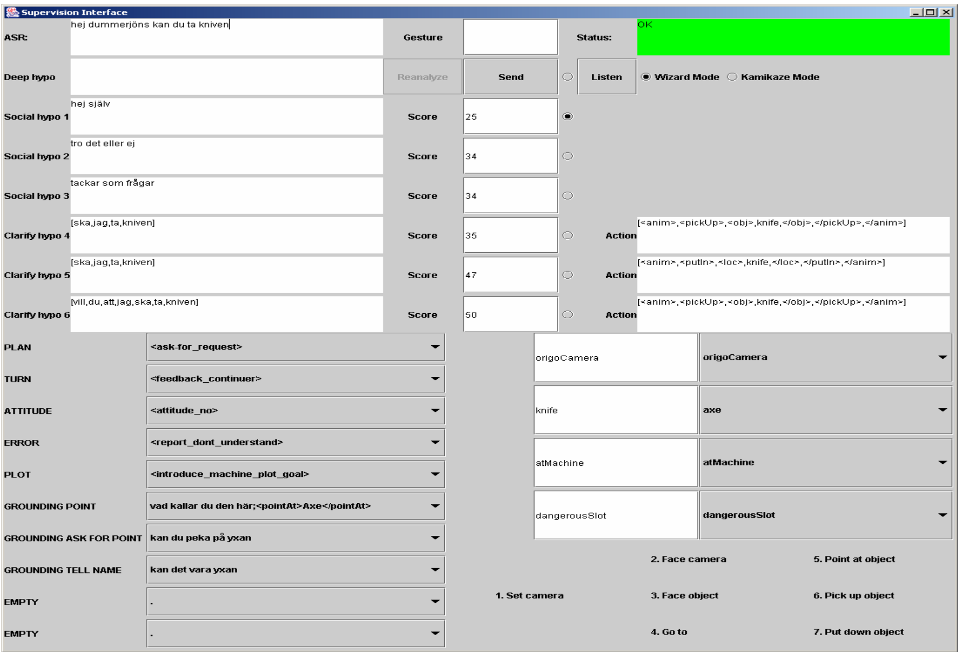

Figure 3.4. The supervision interface

3.6.2 System description

The fairy-tale game involves a number of embodied conversational fairy-tale characters. To make these animated characters appear lifelike, they have to be autonomous, i.e. they must do things even when the user is not interacting with them. At the same time they have to be reactive and show conversational abilities when the user is interacting with them. To build a system that is both autonomous and reactive at the same time has led to the choice of the event driven, asynchronous system architecture that is shown in Figure 3.5. An overview of all components in this system architecture can be found in D3.6.

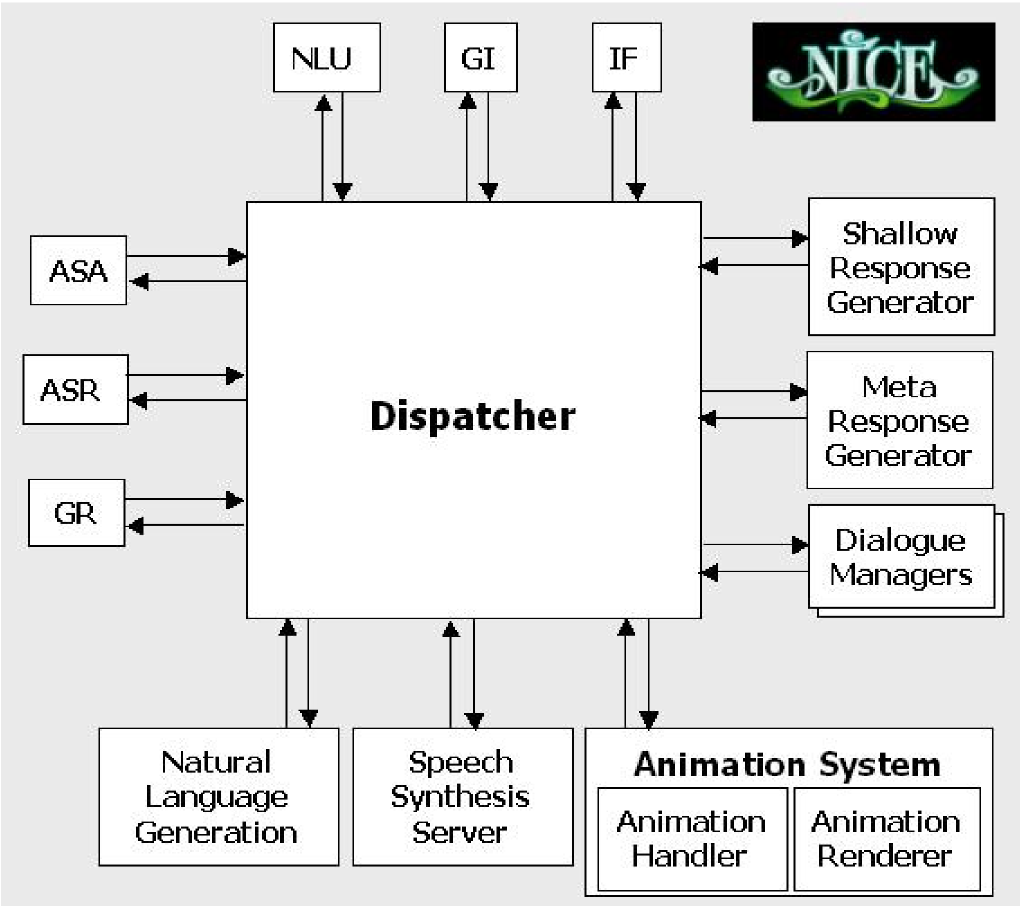

Figure 3.5. The system architecture.

3.6.2.1 Speech recognition

Since the user and Cloddy Hans are supposed to collaborate to solve tasks, it is important that there is a symmetric relationship between the system's input and output capabilities. Since the wording of system prompts tends to influence users, it is important to assure that the system is able to understand everything it can produce. Therefore the speech recognition is trained on the things Cloddy Hans can say, as well as on user utterances collected from human–computer interactions. The speech recognizer used a statistical bi-gram language model. The basis for this language model is a domain model, which also underlies Cloddy Hans's repertoire of utterances, including names for all objects in H. C. Andersen's study, and verbs for all the operations Cloddy Hans can perform (like picking up things, moving about, etc.). It also includes clarification questions to all task-oriented utterances. A set of general dialogue handling utterances were also designed; these include utterances for grounding, error handling, attitudal feedback and turn regulation utterances. Data from test persons using earlier versions of the systems has been added to model, as well as typical socializing utterances from corpora collected with our earlier systems, like August (Gustafson & Bell, 2000) and Pixie (Bell, 2003).

3.6.2.2 Speech synthesis

Since the system will be used for collaborative grounding dialogues, it is also important that the system is able to say everything it had been designed to understand. In the task-oriented dialogues it is thus important that Cloddy Hans can talk about the physical actions he will have to perform to solve a certain task. In the small-talk domain he must be able to ask the users the same questions that the system has been prepared to answer. Another design criterion was that Cloddy Hans should be able to understand and generate both grounding and turn regulation utterances. Finally, to facilitate rephrasing as an error-handling strategy, all domain and meta utterances have been provided in a version with alternative wording.

In order to be able to build Swedish voices with natural voice quality and prosody, a new corpus-based unit selection synthesizer, **Snacka,** has been developed by Kåre Sjölander at CTT/KTH in close collaboration with the Voice Technologies group at TeliaSonera (Gustafson & Sjölander, 2004). The synthesizer only requires a set of speech recordings with matching orthographic transcriptions files. With this as input the system is able to generate a synthetic voice without further manual intervention. The synthesizer produces a sound file containing the verbal realization of the utterance along with a lip-synchronization animation track. It also time stamps the animation tags in the animation To ensure a fast and responsive system, cached synthesized utterances (sound files together with lip-synchronisation tracks) are used during runtime. Different methods are used to generate the cached utterances for the fairy-tale characters. An important role of the synthesis component in the fairy-tale system is to convey the personality of the characters. The two main characters should ideally featurehave the following personality traits:

Cloddy Hans – low Openness, high Conscientiousness, low Extroversion, high Agreeableness, low Neuroticism

Karin – high Openness, low Conscientiousness, high Extroversion, low Agreeableness, high Neuroticism

Cloddy Hans and Karin were therefore provided with voices and speaking styles that were judged to match their respective personality traits:

	Voice pitch	Speaking rate	Frequency range
Cloddy Hans	Low	Slow	Small
Karin	High	Fast	Large

Table 3.2. The overall speaking styles of the two main characters.

To achieve the different speaking styles, the voice talents were instructed to read the utterances in manners that matched the targeted personalities. This resulted in two voices with speaking styles that, among other things, differed in frequency range. They also differed in speaking rate and voice pitch. In order to accentuate these last two differences, all utterances were re-sampled changing speaking rate and voice pitch at the same time. All Cloddy Hans's utterances were slowed down and all Karin's utterances were speeded up, shifting their respective pitch and register accordingly.

3.6.2.3 Animation interface

The animation system for the fairy-tale characters is described in detail in D3.7b and D4.2. The main fairy-tale characters are able to engage in conversation with the users and perform the actions needed to progress through the plots of the game. In order to be believable as life-like characters they are able to react to user input as well as to events in their environment. These reactions are either displays of attitude (emotional displays, see Ekman et. al. 1972), state of mind (listening, thinking and understanding), turn regulation gestures (turn taking or back-channelling gestures) or attention gestures which can be used when the user starts to speak or gesture as well as if another character starts to speak or if something happens in the environment. The characters can also look at and point at interactive objects (moveable objects, draw bridge), non-interactive objects (nearby houses, trees and roads) and landmarks (forests and hills) in the 3D-world. They are able to walk between locations that are far apart. The characters do not stand completely still at one location forever – if the user has not engaged in interaction with them for a while they enter an idle state where they start off with small encouraging gestures, then after a while they indicate impatience by gazing around in the environment or displaying various idle gestures. Finally if the user seems to be unwilling to communicate with them they start walking about in the scene, performing different kinds of non-communicative gestures and actions.

The characters' non-verbal behaviour is controlled by the *Animation Handler* module. This module sends requests to the *Animation Renderer*, telling it either to play animations and/or sounds or to perform certain character actions. Liquid Media have provided all characters with a number of communicative gestures, as well as a number of simple, single body part animations that can be used to generate more complex multi body part gestures. This makes it possible to either play ready animations for communicative gestures, like the ones for "Thinking", "Did not understand" and "Did not hear" – or to generate animation lists consisting animation tracks for the individual body parts.

To support the intended personalities of these characters, the introvert, shy, quiet, calm, polite Cloddy Hans displays small, but slow and deliberate body gestures that are generated by the AnimationHandler, while the outspoken, self-confident, touchy and anxious Karin displays larger, and faster body gestures, as rendered by the animators. The characters' different personalities are also conveyed by their different idle behaviours: Karin is less patient than Cloddy Hans, which is reflected by the fact that she enters the idle phase faster and has more complex idle gestures. Cloddy Hans keeps his attention at the user, while Karin's attention from time to time wanders away from the user to the environment. If the user has not said anything for a long time Karin eventually walks away from the user and she starts strolling around aimlessly in the back of the scene, looking at trees and flowers. When she has walked away from the drawbridge, the user has to talk to her to get her attention again, in order to get her to return and continue the conversation. Since both characters have been provided with the same sets of animation it is possible to switch their non-verbal behaviours.

3.6.2.4 Dialogue processing

Besides the different shallow dialogue processing tracks, the input was also processed by a “deep” dialogue processing track, implemented according to the ideas outlined in deliverables D1.2b, D3.5b and D5.2b. The kind of user utterances the system can interpret can be categorized as follows:

- **Instructions**: "Go to the drawbridge", "Pick up the sword", etc.
- **Domain questions**: "What is that red object", "How old are you", etc.
- **Giving information**: "I'm fourteen years old", etc.
- **Stating intentions**: "I will give you the ruby", etc.
- **Confirmations**: "Yes please", "Ok, do that", etc.
- **Disconfirmations**: "No", "Stop!", "I didn't say that", etc.
- **Problem reports and requests for help**: "Help", "What can I do?", "I don't understand", "What should we do now?", "Do you hear me", etc.
- **Requests for explanation**: "Why did you say that?", "Why are you doing this", etc.

The fairy-tale characters have an overlapping but not completely identical set of classes of utterance they need to generate:

- **Responses to instructions**: either **accepting** them ("OK, I'll do that") or **rejecting** them, ("No I won't open the drawbridge!"). Rejections can contain an explanation ("The knife is in the machine" as a response to "Pick up the knife").
- **Answers to questions**: "The ruby is red", "The knife is on the shelf", etc.
- **Stating intentions**, e.g. "I'm going to the drawbridge now".
- **Confirmation questions** to check that the system has got it right, e.g. "You want me to go to the shelf, is that right?"
- **Clarification questions** when the system has incomplete information, e.g. "Where do you want me to go?", "What should I put on the shelf?", etc.
- **Suggestions** for future courses of action, e.g "Perhaps we should go over to the drawbridge?"
- **Explanations**: "Because I want the axe in the machine".

3.6.2.5 Gestural input

The Gesture Recognizer (GR) and Gesture Interpretation (GI) modules implemented by LIMSI (see NICE deliverable D3.4) were also connected to the system. The output from the GI module is the name of the object the user has pointed at (or otherwise gestured at) with the gyro mouse. This information was used for automatic focus management in the supervision interface. If, for instance, the GI reports that the user has clicked on the magic wand, and the operator the presses the “Pick up” button on the supervision interface, Cloddy Hans will pick up the magic wand.

3.6.3 Data collection set-up

Data was collected on several occasions using the NICE system at different stages during its development. The system could be run either in fully automatic mode or in supervised mode, in which a human operator had the possibility to intervene and replace or modify the output of system components. This made it possible to develop the system in a data-driven, iterative fashion, by initially gathering data in partially supervised mode and by running several cycles of data collection, data analysis and corresponding system development.

Four sub-corpora were collected over a period of 5 months during 2004 – 2005. The recording conditions are described in Table 3.3 where the sub-corpora are labeled "School", "Lab 1", "Lab 2" and "Lab 3", respectively. During this period a fair amount of changes to the system took place, including the addition of the second scene in which Karin appears, as well as considerable improvements to the system's spoken language understanding capabilities. Thus, the four sub-corpora consist of data collected from heterogeneous user groups under differing conditions during several stages of the development of the NICE system. Speech data was collected when users were interacting with the system, as well as during a post-session interview. All subjects were recorded using a close-talking head-mounted wireless microphone, and subjects in sub-corpora Lab 1–3 were also recorded on video. Data from all major sub-components of the NICE system was also logged. Prior to the interaction, each user was given a short instruction and was also asked to fill out a questionnaire, recording demographic data and self-estimates of computer and video game use. The instructions were deliberately sparse–the users were told that they would be testing a research prototype of a new kind of computer game, where they would be able to talk to fairy-tale characters adopted from H. C. Andersen's stories. Following the interaction with the system the subjects were interviewed about their experiences with the game and the characters involved in it. The interviews in the Lab 3 sub-corpus also included replaying a short video capturing the system screen and the spoken interaction between the user and the system as a background to in-depth questions. After this, the subjects were given a second questionnaire assessing various aspects of the game as well as properties of the characters involved in it. This questionnaire used 5-point Likert scales (Likert, 1932), with which even the youngest subjects were familiar through the use of such instruments in school.

Some data was discarded for reasons such as drop-outs or failure in logging one or more of the involved modalities. All remaining speech was automatically segmented using the speech detection algorithm of a commercially available speech recognizer for Swedish, yielding close to six hours of spoken language data of which approximately two thirds were computer-directed speech. This material was orthographically transcribed, with special symbols employed to denote disfluencies, non-speech sounds etc. and linguistically analyzed.

Condition	School	Lab 1	Lab 2	Lab 3
Date	Nov-Dec, 2004	Dec, 2004	Feb, 2005	March 2005
Location	Small room (not sound-treated) in a school	Very large room in TeliaSonera's vision center	Sound-treated large room in TeliaSonera's multimodal lab	Sound-treated large room in TeliaSonera's multimodal lab
Equipment	CRT display, mouse	Large display, gyro mouse	Large display, gyro mouse,	Large display, gyro mouse
Data	Audio, system logs	Audio, video, system logs	Audio, video, system logs	Audio, video, system logs
Gameplay	Scene 1	Scene 1	Scene 1+2	Scene 1+2
Position	Sitting down	Standing	Standing	Standing
Age span	8–11	14–15	9–10	11–12
Users	31	11	20	13
Discarded	5	4	5	4
Net number	26	7	15	9

Table 3.3. Recording conditions for the four different sub-corpora

3.6.4 Quantitative data description

The total number of user sound files in the human–computer dialogue corpus was 5,580 as obtained from 57 users interacting with the system. Of these, 26 played only the introductory scene, 7 played one and a half scene (being able to explore the fairy-tale world, but not interact with Karin), and 24 played both scene 1 and 2. The two scenes together could consist of up to 16 phases (depending on how many objects they fetched in each scene). The phases are defined as a discourse part between two story-functional events (see 3.5 above), named by the event it leads to. In Appendix 1, a complete session constructed out of dialogue phases from different users exemplifies the type of child-machine dialogues typically found in the different phases. A quantitative description of the phases in the game is presented in Table 3.4.

	average number of subtasks	average number of user utterances	least number of utterances	most number of utterances	average number of turns per subtask
cloddy social	1,6	7,7	1	22	4,9
machine plot description	1,0	5,5	1	18	5,5
first object to machine	2,6	20,6	4	60	7,9
second object to machine	3,3	16,6	5	98	5,1
third object to machine	2,8	10,5	1	28	3,7
fourth object to machine	3,4	13,0	2	36	3,8
fifth object to machine	3,0	8,6	5	11	2,9
sixth object to machine	2,5	9,0	5	13	3,6
machine full pull lever	1,0	5,7	1	18	5,7
world introduction	1,0	2,5	1	5	2,5
exploring small island	3,0	13,7	1	72	4,6
Cloddy introduces Karin	1,8	5,8	1	35	3,3
Karin social	1,0	5,4	2	16	5,4
Lower Bridge Negotiation	7,1	36,2	6	73	5,1
Giving toll to Karin	1,2	7,9	1	39	6,8
exploring world	1,1	5,9	2	13	5,3
out of phase	1,7	10,6	1	31	6,3

Table 3.4. Quantitative description of the phases in the game.

The number of user utterances per “object to machine”-phase shows that there was a considerable learning effect, where the users needed less and less utterances to complete a phase. Another indication of the can be seen when studying the initiative in the system. All subtasks that was needed to be performed in order to finish a phase was tagged with information about who initiated it. It can be seen in Figure 3.6 that the user initiative increases from 60% for the first object to almost 100% the last objects.

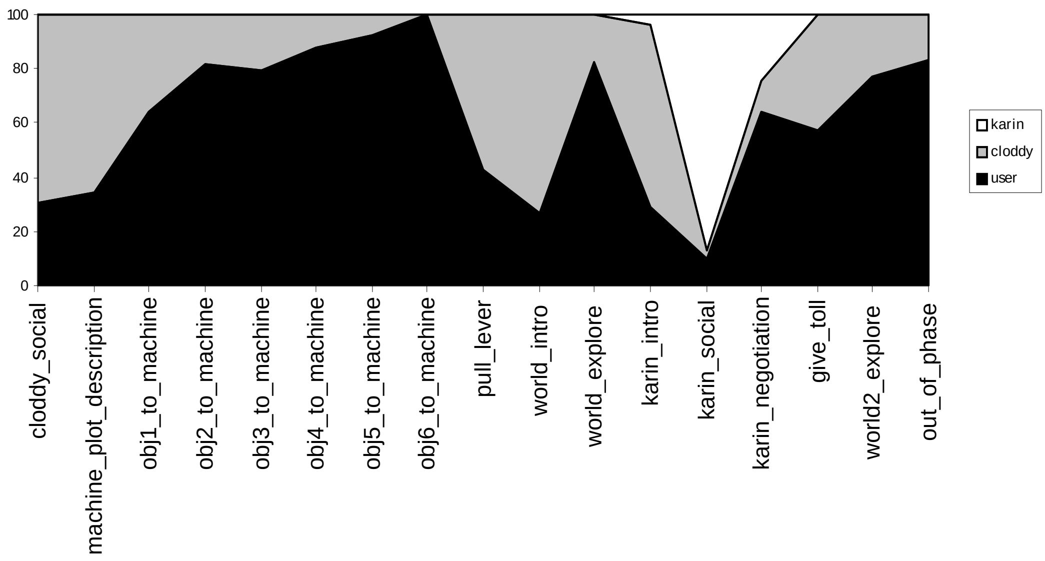

Figure 3.6. Percentage of subtask initiations per speaker in the different phases.

In a previously designed experiment, we wanted to investigate whether speakers adapted their speech rate to that of a spoken dialogue system. Users were asked to interact with either a ‘fast’ or ‘slow’ version of Cloddy Hans, and their speaking rate was examined. Results confirmed that the users adapted to the speaking rate of the system, increasing their speech rate when interacting with the ‘fast’ character and decreasing it when interacting with the ‘slow’ character. Another finding was that users varied their speaking rate substantially in the course of the dialogue, often speaking slower during problematic dialogue sequences (Bell, Gustafson, & Heldner, 2003). As a consequence of these results, the characters Cloddy Hans and Karin were designed to elicit different manners of speaking. In the current corpus, the users’ speaking rate was examined in the context of the ‘slow’ and ‘fast’ character, respectively.

All 5580 utterances were aligned on phoneme level using the snack aligner (Sjölander & Heldner 2004). The mean and standard deviation of the duration for each phonome was then computed, and the duration z-score for all 85000 phonemes in the corpus were thus retrieved. These z-scores were used to analyze the speech rates of the users.. The analysis revealed user adaptations of speaking rate that were similar to the ones found in our initial study (Bell, Gustafson, & Heldner, 2003). Those users that took more and more initiative per object phase, using fewer and fewer turns to solve each phase (see Figure 3.7), at the same time talked slower and slower, to make sure that the sluggish Cloddy Hans would understand them. In the second scene, the users increased their speaking rate initially, but then began to talk slower again when talking with Cloddy Hans. As soon as they started talking to Karin, they increased their speech rate instead (Karin as well as Cloddy), probably since the general pace of the interaction was higher and since Karin did not show at all if she had understood or not. Instead, she drives the dialogue and tells the user what she wants something nicer, which then makes the user tell Cloddy to fetch something. Significant differences in speaking rate were observed between the User-Cloddy dialogues and the User-Karin(-Cloddy) dialogues.

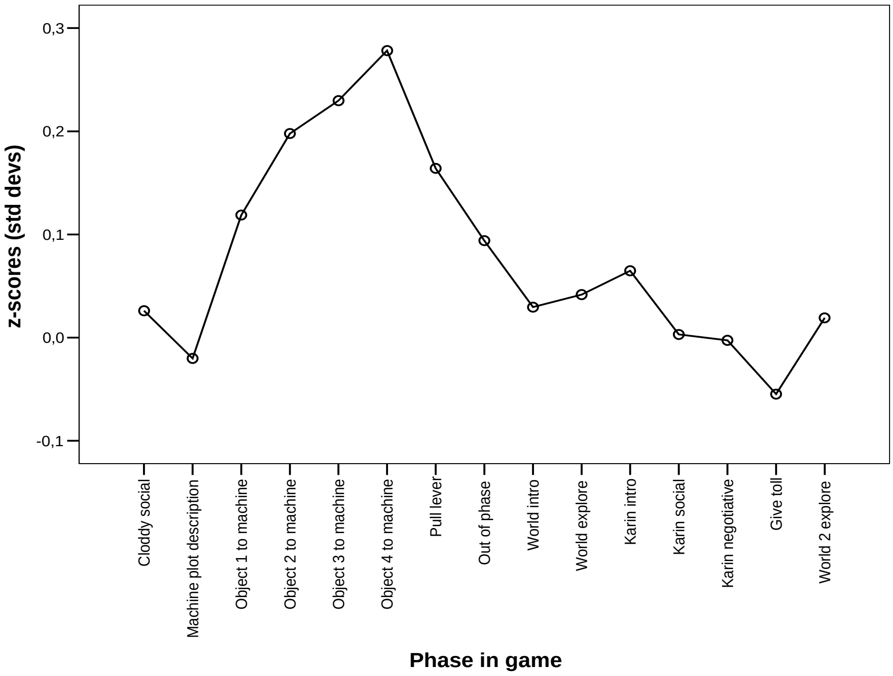

Figure 3.7. The mean speaking rate values for the user's utterances in the different phases.

The 5580 user sound files were transcribed and tagged. Utterance fragments were identified and joined into turns, following which the number of turns was calculated, resulting in 5,144 user turns. The average number of turns per user was 90, with individual variation ranging from 26 to 210 turns. The user utterances were divided into six categories:

- **Domain**: domain oriented utterances "Go to the drawbridge", "Pick up the sword", etc.
- **Yes/No**: confirmations "Yes please", "Ok, do that" and disconfirmations "No", "Stop!"
- **Social/fun**: socializing ”What's your name?”, "I'm fourteen years old", testing the limitations of the system “break something”, “kill the girl”,and lying and joking “my name is Cloddy Hans”, “So you didn't realize that you are stupid until now!”
- **Meta**: "Help", "What can I do?", "I don't understand", "What should we do now?", "Do you hear me", “What did you say?”, "Why did you say that?" etc.
- **Repetition:** repeating an utterance of any of the types above
- **Fragment**: continuing an utterance of any of the types above

The distribution of these categories for all 57 users is shown in table 3.5.

Utterance type	Share [%]	Range [%]
Domain	39	16–63
Yes/no	12	0–35
Social/fun	7	0–21
Meta	17	3–39
Repetition	17	2–37
Fragment	8	1–32

Table 3.5. Distribution of utterance types across 5,580 user utterances

The dialogue corpus also contains 5,583 Cloddy Hans turns and 255 Karin turns, which were speech act tagged and then divided into the following 6 categories:

- **Domain**: Initiatives to do subtasks “maybe I should pick up the axe”, descriptions of objects “the sword have been used in many fairy-tales” , suggestions on what to give Karin and description of the plots,
- **Problem**: Cloddy explicitly telling the user that he did not understand “I did not understand you”, “could you repeat that” or that he failed to what the user asked “I can’t do that”, "The knife is already on the shelf", etc.
- **Safe:** explicit confirmations “did you want me to pick up the axe”, telling the user that his request is about to be executed "I'm going to the drawbridge now", asking the user what to do “what should i do now” or to speak to Karin “could you talk to her- I don’t dare to”
- **Meta:** acknowledgements “yes”, “I’ll do that”, meta questions “do you think so”, “did I say that” filler “mmm”
- **Nice:** social exchanges “hello”, “how are you”, politeness “thank you” “how fun”, “I would be happy to do that”
- **Bad**: Cloddy mocking Karin “she is so grumpy” and vice versa “why do you bring that Cloddy Hans”, rejections “I don’t want to" and angry remarks to the user “what did you say about me”, “you doodle too much”

Utterance type	Share [%]	Range per user [%]
Domain	15	4–28
Problem	15	6–29
Safe	39	24–59
Meta	13	3–31
Nice	12	4–24
Bad	7	0–19

Table 3.6. Distribution of utterance types across 5,445 Cloddy utterances

Utterance type	Share [%]	Range per user [%]
Domain	45	32–62
Problem	0	
Safe	0	
Meta	0	
Nice	27	8–47
Bad	27	13–42

Table 3.7. Distribution of utterance types across 255 Karin utterances

The deliberate dialogue design for the different characters was to make Cloddy appear uncertain and not so smart by letting him generate a lot of utterances the would indicate that (problem, safe) and to never let Karin generate these, since she was supposed to appear smart and self-assured, see Tables 3.6 and 3.7. In terms of Nice/Bad Karin had equal amount of these kinds of utterances, while Cloddy had significantly more utterances of the Nice type – since that was what we aimed at making him a friendly helper character.

Apart from the corpus of child–machine dialogues, the subsequent child–adult interviews were also transcribed, yielding a second set of 775 sound files. Considerable differences in utterance length between these two data sets were found. The number of words per utterance was 8.1 in the human–human dialogues, but only 3.6 in the computer-directed dialogues. Another difference between the two data sets was found as concerns the proportion of filled pauses, filler words and phrases, e.g. “like” and “you know”. In computer-directed speech, these constitute 5% of all utterances (1.3% of all word tokens) whereas in human-directed speech they constitute no less than 35% of all utterances (4.3% of all word tokens). Yet another difference was that the human–computer utterances on average were 30% slower than the human–human utterances.

3.7 Technical evaluation

3.7.1 Speech recognition

The speech data collected by TeliaSonera was not sufficient to train the acoustic models of Scansoft's recognizer OSR3 from scratch, neither did ScanSoft have access to enough Swedish acoustic data at sampling rates of 16kHz or higher, to add to TeliaSonera's data. The Swedish OSR3 acoustic models are therefore based on an 8 kHz model, normally used for telephony speech recognition. Using TeliaSonera's collected speech data, these telephony models were adapted in a series of experiments to better fit the target NICE application.

3.7.1.1 Adaptation data

TeliaSonera provided 5 corpora of speech data to ScanSoft. All corpora were sampled at 16kHz. Table 3.8 shows the size and origin of these corpora:

	#utterances, duration	***Remarks***
July2002	10813, 13.5 hours	Stockholm Telemuseum
October2003	2119, 5 hours	Stockholm Telemuseum
April2004	733, 0.5 hours	NICE WoZ collection
November2004	2953, 2 hours	NICE WoZ collection
December2004	928, 0.7 hours	NICE WoZ collection

Table 3.8. The size and origin of the corpora.

For acoustic model adaptation, the data was down-sampled to 8kHz.

3.7.1.2 Adaptation process

The Swedish OSR3 acoustic model was adapted using the above-mentioned corpora applying supervised MAP adaptation. OSR3 provides mechanisms to

- scale the influence of the adaptation utterances (the adaptation rate) on the acoustic model, and to.
- exclude utterances from the adaptation process that did not yield a good alignment of sound with respect to. its transcription. In the case of supervised adaptation, this could be caused by incorrect transcriptions or by a particularly large mismatch between the adapted model and an adaptation utterance. This safe-guard process was controlled via word or sentence confidences thresholds.

3.7.1.3 Model testing

The following table compares three acoustic models. AM_a is the standard OSR3 telephony acoustic model. AM_b is the model adapted on the early July2002 and October2003 corpora (corresponding to the acoustic model provided by ScanSoft as deliverable D3.1). AM_c is adapted on the complete adaptation data (as provided as deliverable D3.2).

We are using two different test corpora. The corpus called “Telecom” consists of 642 utterances, which are randomly selected from July2002 and October2003. The second corpus “Test2004” is taken from the three 2004 WoZ collections and contains 424 utterances:

Telecom	*AM_a*	*AM_b*	*AM_c*
Word error rate	33.7%	29.8%	29.3%
Sentence error rate	37.1%	32.2%	31.9%

Test2004	*AM_a*	*AM_b*	*AM_c*
Word error rate	52.8%	36.5%	30.7%
Sentence error rate	63.9%	48.8%	42.0%

Test2004 shows higher error rates, especially for the non-adapted model. The probable reason is that Test2004 is made up exclusively of children’s speech, whereas the Telecom corpus constitutes a mixture of children’s and adults’ speech. The table also indicates that a rather small set of adaptation data can substantially reduce the error rates. We expect, that an additional large improvement could be gained by applying acoustic models for 16kHz sampling rates.

3.7.1.4 Speech Transformation Experiments

Looking at other recent research done on recognition of children’s speech, Potamianos (2003) gets closest to the results presented above. With a similar combination of speech transformation and acoustic model adaptation he achieves word error rate improvements of 45% relative to an adult-speech trained telephony acoustic model. An additional gain of 10% was achieved by using age-dependent acoustic models for the children. It needs to be noted, however, that Potamianos had substantially more children speech data available (factor of 2 or 3 relative to the NICE data collections). Gustafson & Sjölander (2002) did experiments on speech transformation without acoustic model adaptation and report WER improvements of 30 to 45% relative to a given adult-speech trained telephony acoustic model. Other work on children speech recognition concentrates on speaker normalization (Gerosa & Giuliani, 2004; Giuliani & Gerosa, 2003; Hagen, Pellom, & Cole, 2003), generally achieving error rate improvements in the usual range for VTLN, i.e. about 10% relative. Li & Russel (2002) point out the importance of children-specific pronunciation modeling, e.g. by using customized recognition dictionaries. Their error rate improvements are also in the range around 10% relative. Pronunciation modeling has not been investigated by ScanSoft in the context of the NICE project. However, small gains in accuracy have been achieved by optimizing noise and hesitation modeling, both in the acoustic model and for the statistical language model.

The Swedish NICE fairy-tale prototype targets children speakers, too. An additional complication was due to the reduced amount of available in-domain speech data. Training an acoustic model from scratch was therefore not feasible. Instead, a commercial acoustic model was used and adapted on the available speech data. Two measures have been used in combination to optimize the performance of the adapted model:

- The input speech data (children speech) had been manipulated to better match the acoustic model (trained on adult speech). The manipulation is mainly a translatory mapping of the speech within the frequency domain. Transformation of the input speech achieved a gain of 22% in word error rate.

- The adaptation algorithms have been optimized to achieve a maximum model improvement on small amounts of adaptation data. The word error rate reduction yielded from 20 hours of adaptation data was about 45% relative. This gain is additive to the improvement of 22% resulting from input transformation.

A couple of smaller modifications of the OSR speech recognition engine add to the technology mentioned above. Regarding the overall OSR improvements introduced during the NICE project, this clearly exceeds the state-of-the-art at project start.

3.7.2 Natural language understanding

3.7.2.1 Corpora and data-collection methodology

To evaluate the parser, we used 3400 utterances from our corpora which, as mentioned, had been collected at four different occasions over a five-month period. The subjects were children, aged 8-15. At the first data collection occasion, the subjects played the first scene only. At the second occasion, the subjects played the first scene, and then were allowed to explore the fairy-tale world together with Cloddy Hans. At the two last occasions, the subjects played two entire scenes, including the negotiation with Karin in order to cross the bridge. The 3400 utterances contain 810 unique words and 11925 tokens, of which 1715 tokens are outside the system's present vocabulary of 525 words (i.e. the out-of-vocabulary rate is 14,4%).

To allow for extended user sessions where the player was able to explore the scenarios without being hindered by occasional errors due to imperfect speech recognition or understanding, the system was run in *supervised* mode. This meant that a human operator was supervising the interaction from behind the scene, and had the opportunity to interfere and correct the speech recognition result whenever he judged that the original result would seriously disturb the progression of the dialogue. He was also allowed to edit the system's response back to the user before this was output in cases where it would likewise have disturbed the progression of the dialogue.

It should be emphasized that the purpose of using supervised mode in the data collection was purely to ensure that the game (and hence the *dialogue*) was moving forward in those cases where there was otherwise a risk that it would be stalled or that repetitious errors would occur. Most importantly, all performance figures presented here are based on the recognition results obtained *before* any editing by the human operator. Hence, there is no "contamination" of the figures from the point of view of measuring the quality of parsing as such (since the domain of parsing is limited to single user turns). Actually, we believe that if supervised mode has any effect on the difficulty of the parsing task, it is rather to make it *harder*, since what supervised mode does is to occasionally "help" a fairytale character to address the player in a more coherent and intelligent fashion than would otherwise have been possible .

Naturally, the quality of the results delivered by the parser, and ultimately the degree of understanding of an utterance, is contingent on the quality of the input delivered by the speech recognizer. The quality of this input is estimated by the standard measures of sentence accuracy and word accuracy, whereas the quality of the final results are measured in terms of *semantic accuracy* and *concept accuracy*. By semantic accuracy we mean the proportion of utterances where the output of the parser *exactly* matches the correct analysis. Semantic accuracy can thus be seen as the semantic analogue of sentence accuracy. In contrast, concept accuracy is based on the number of semantic units that are substituted, inserted and deleted, and can thus be seen as the semantic analogue of word accuracy (Boros et al., 1996).

In order to calculate concept accuracy, we need a rigorous definition of a "concept". For all semantic expressions (except lambda abstractions), we will consider a "concept" to be a node in the tree making up the semantic expression. For instance, the expression

ask_for_attention(user, cloddy)

can be seen as a tree with the root node labeled ask_for_attention, and two leaf nodes labeled user and cloddy, respectively. So this expression has three concepts, but for the purpose of calculating concept accuracy, we will not count user (the first argument of a dialogue act), since it is always assumed that the dialogue act originated from the user.[2] Hence for expressions that are not lambda abstractions, the number of concepts equals the number of nodes in the tree making up the expression, minus one.

For lambda expressions, we simply do the same calculation for the body of the expression. For instance, the expression

λx^{thing}.request(user, cloddy, pickUp(cloddy, x))

is considered to have the concepts present in the body of the lambda expression, namely request, user, cloddy, pickUp, cloddy, x^{thing}. Out of these, we include all concepts except user for the purpose of calculating concept accuracy.

An error occurs when a concept *c* appears in the semantic analysis of the input, but the corresponding place in the correct semantic analysis is occupied by a different concept *d*. If neither *c* or *d* are variables, the error is a substitution; if *c* is a variable but not *d*, the error is a deletion; if *d* is a variable but not *c*, the error is an addition.

3.7.2.2 Basic results

When constructing the set of 3400 correct analyses, altogether 509 utterances (15%) were judged not to be representable within the semantic formalism. These unrepresentable utterances ranged from fragments that could mean just about anything (e.g. "*Was it*"), through unanticipated requests (e.g. "*Kill the girl*") and musings ("*I thought as much*"), to complicated counterfactual statements ("*If you had taken the sword earlier you would have been able to cut the cloth to pieces now*"). Note that some of these unrepresentable utterances are not only problematic for the parser, but also pragmatically very difficult, which means that it is not always possible for the system to produce a coherent response.

In the tables below, we report sentence accuracy both with respect to the complete set of 3400 utterances and with the set of 2891 utterances that actually had a complete semantic representation. For the set of 3400 utterances, we judged an analysis to be correct or incorrect as follows: If the parser *failed* to produce an analysis for an unrepresentable utterance (giving as output "failed_act"), we took that output as being correct on the grounds that signalling that no analysis can be produced is the most that we could reasonably expect the parser to do in that case. (Following such an output from the parser, the dialogue manager would then try to repair the

[2] This is not true for nested dialogue acts, however, as in one example from our corpus; "Tell Karin to lower the bridge", represented as:

request(user, cloddy, request(cloddy, karin, windDown(karin)))

Here the user is requesting that Cloddy Hans make a request, so the first argument of the second request is cloddy, not user.

dialogue.) On the other hand, if the parser *did* produce an analysis for an unrepresentable utterance, we made the pessimistic assumption that that output was completely erroneous.

An analogous method was used to determine concept accuracy. Failure of the parser to produce an analysis for an unrepresentable utterance is counted as one instance of correct (the presence of "failed_act"), whereas the analysis of an unrepresentable utterance will be counted as one deletion (missing "failed_act") plus one insertion for each additional semantic unit.

The results are shown in Table 3.9 below. The top of the table shows the accuracy of the speech recognizer. 30.6% of the recognized utterances were perfectly recognized, and the word accuracy was 38.6% (that is, the word error rate was 61.4%). These very poor figures are largely due to the fact that the subjects were children, and that speech recognition in particular is much less reliable for children than for adults. Furthermore, in our data the recognition results varied a lot between speakers. For some children, recognition was consistently dismal, whereas for others recognition worked quite well. That is, there was a kind of "recognize-everything-or-recognize-nothing" tendency, which explains the fact that the difference between sentence accuracy and word accuracy is small. This tendency was further amplified by the fact that the dialogues were long (the mean length of the dialogues was on the order of 90 turns). This allowed the children for which recognition worked well to gradually learn how to express themselves within the coverage of the system's understanding capabilities, making recognition work even better for them.

	Speech input	**Recognized input**	**Transcribed input**
Speech recognizer			
Sentence accuracy	30.6%		
Word accuracy	38.6%		
Parser			
Semantic accuracy (all)		48.6%	84.8%
Semantic accuracy (representable)		49.1%	90.2%
Concept accuracy (all)		53.2%	86.4%
Concept accuracy (representable)		50.5%	92.6%

Table 3.9.

The bottom part of the table shows the accuracy of the parser. The robustness of the parsing algorithm can be seen by comparing the first and second columns. The parser managed to recover the correct analysis for 48.6% of the utterances, in spite of the fact that only 30.6% were perfectly recognized. Similarly, the concept accuracy of the parser output is 53.2%, although the word accuracy is only 38.6%.

The third column shows how the parser performs on transcribed (perfectly recognized) input. Here the semantic accuracy is 90.2% for the utterances that could be represented; that is, the parser fails to produce the correct analysis for only 9.8% of the utterances. Basically, the latter figure shows the coverage leaks, whereas the difference between 90.2% and 84.8% (that is, 5.4%) shows the extent to which the parser produces unwarranted analyses beyond the scope of the semantic formalism.

3.7.2.3 Further experiments

The parser's performance on transcribed input can be seen as a "roof" which will never be attained because of the inevitable distorsion of the input caused by the speech recognizer. A more realistic "roof" for the parser can be obtained by looking at *N*-best output from the speech recognizer, and more specifically the extent to which a (more) correct hypothesis being present there, as compared to it being the top hypothesis (1-best). To determine the effects of using *N*-best output, three experiments were run. First, sentence and word accuracy were computed using 10-best output from the speech recognizer for the set of 3400 utterances. Thus, for word accuracy, the best hypothesis compared to the transcribed utterance in terms of the number of substitutions, insertions and deletions at the word level was picked out from the 10-best list. The resulting sentence accuracy and word accuracy are shown in Table 3.10.

Speech recognizer	**1-best**	**10-best**
Sentence accuracy	30.6%	42.1%
Word accuracy	38.6%	55.0%

Table 3.10.

As could be expected, this "oracle algorithm" (always picking the best hypothesis) gave a significant improvement of both sentence and word accuracy (37.6% and 42.5% relative, respectively). Although the result does not alter the fundamental picture of the speech recognizer as constituting the main bottleneck for robust understanding, it still shows that something may be gained by looking at *N*-best rather than 1-best.

In a second experiment, the corresponding results for the semantic level were computed, shown in Table 3.11. Here, the second column shows the results for the hypotheses whose analyses from the parser corresponded most closely to the correct analyses in terms of the number of substitutions, deletions and insertions of semantic units.

	1-best	**10-best**
Parser	48.6%	65.4%
Semantic accuracy (all)		
Semantic accuracy (representable)	49.1%	66.3%
Concept accuracy (all)	53.2%	70.4%
Concept accuracy (representable)	50.5%	72.3%

Table 3.11.

The results again show a significant improvement (between 32% and 43% relative, respectively), indicating great potential gains by using *N*-best rather than 1-best. However, the problem then is to find a set of effective criteria which can be applied at run-time, and by which the best candidate from the *N*-best list can be found in as many cases as possible.

To get a handle on this, a third experiment was run with the aim of determining whether a simple heuristic would be useful for finding the best hypothesis from a semantic point of view. The heuristic adopted was that of keeping track of which hypothesis allowed the parser to skip the fewest words in the recognized input, and to correlate this with the hypothesis that gave the best semantic analysis (in the second experiment). The heuristic is thus relevant to the first, pattern-matching phase of the parser (see Section 5.2), and the assumption was that the better the parse was "fit", the better the result would be. Unfortunately, it turned out that the number of skipped words is a bad predictor of the quality of a hypothesis.

	Semantically best analysis from parser	Fewest skipped words
***N*-best hypothesis 1**	2419	1439
***N*-best hypothesis 2**	333	77
***N*-best hypothesis 3**	161	38
***N*-best hypothesis 4**	134	32
***N*-best hypothesis 5**	99	31
***N*-best hypothesis 6**	61	13
***N*-best hypothesis 7**	55	9
***N*-best hypothesis 8**	57	10
***N*-best hypothesis 9**	46	6
***N*-best hypothesis 10**	35	4
Sum	3400	1659

Table 3.12.

As can be seen in Table 3.12, the least number of skipped words predicts the best analysis in 1659 cases (48.8%), whereas the 1-best hypothesis predicts the best analysis in 2419 cases (71.1%). Thus, improving on the basic criterion of just using 1-best requires something much more elaborate than looking at the number of skipped words in the input.

An obvious alternative solution is to defer the decision of which hypothesis is (semantically) best, by sending analyses of all hypotheses on the N-best list to the next processing step in the system, which is the dialogue manager. The dialogue manager would then be able to use contextual expectations to find the best analysis on the list. For instance, if Cloddy Hans had posed a question to the user in the preceding turn, the system can sift through the list of analyses, looking for an expression that seems to represent an answer to the question. The disadvantage with such an approach is that the resulting dialogues would tend to be more system-driven and less open to user initiative, since the system would be more governed by its own expectations what is supposed to happen next.

3.7.2.4 Discussion

As already mentioned, the most common reason for incorrect anlyses being produced by the parser is misrecognition; that essential words are missing in the input or have been erroneously inserted. The remaining problems can be roughly grouped into different categories, having to do with lexical coverage leaks, commonly misrecognized words, lexical ambiguities, complex grammar, pragmatic ambiguities, and semantic and ontological insufficiencies. These categories are not clear-cut; many utterances can be said to belong to two different groups.

One group consists of utterances running into problems caused by semantic and ontological insufficiencies. This group include many completely reasonable utterances that, at present, cannot be represented within the semantic formalism, e.g. requests for instructions in specific situations ("Am I supposed to, you know, pull things?", "How do you usually do this?"), questions concerning Cloddy Hans's mental state ("Are you having a good time?"), instructions ("Kill her", "Pick some flowers", "Break something"), complex spatial references ("The second last slot", "Go to the left, that is, your left") and various comments ("I just told you", "I don't give a damn", "I was just kidding"). But it also contains completely unexpected input which we will *not* try to incorporate into the system's repertoire. One boy liked to think of the fairy-tale machine as a time-travel machine, and tried to explain the concept to Cloddy Hans ("you can use it to travel into the future and backwards in time", etc.).

Commonly misrecognized words pose problems in those cases where the substitution of one word for another completely alters the meaning of the utterance, e.g. "What is the fairy-tale machine?" and "Where is the fairy-tale machine?". Here the Swedish words for "what" ("vad") and "where" ("var") are very similar-sounding, and thus easily misrecognized.

Lexical ambiguities are rare in this domain, but point to a fundamental problem to the extent that they occur. The parsing algorithm is deterministic and produces one output expression only; hence it sometimes has to make premature decisions that eventually turn out to be wrong. An example is "Varför går inte det?" (Why doesn't that work? / Why is that impossible?). The word "går" has two meanings in Swedish; it may also mean "walk" or "go". Therefore the parser falsely triggers on the two patterns "varför" and "går", and interprets the utterance as a question about why Cloddy Hans does not go to some (unspecified) place.

There are a few utterances in the corpus that seem to call for a more grammatical parsing method. One such example is "Are all the gadgets that were lying on the shelf lying on the grass here?", asked by a subject when he entered the second scene (this utterance is also semantically complex; a yes/no-question concerning a universially quantified implication).

Finally, there are some pragmatic ambiguities, where it is unclear what dialogue act the user is actually making. An example is "Can you do that?", where it is not clear whether the user is making a request or whether he is enquiring about Cloddy Hans's capabilities. However, such utterances would cause problems for any spoken language understanding method.

3.7.3 Gesture Recognition

3.7.3.1 Gesture Recogniser Log files

The quantitative analysis of the GR log files revealed the following repartition of recognised shapes.

USER	nbPointer	nbSurrounder	nbConnect	nbUnknown	TOTAL
1	17	8	4	2	31
2	4	1	3	0	8
3	24	23	40	21	108
4	9	2	2	1	14
5	0	1	0	1	2
6	2	2	8	3	15
7	4	19	0	2	25
8	27	5	15	2	49
9	25	68	7	11	111
10	0	1	3	1	5
11	41	10	48	6	105
TOTAL	153	140	130	50	473
%	32	30	27	11	100

Table 3.13.

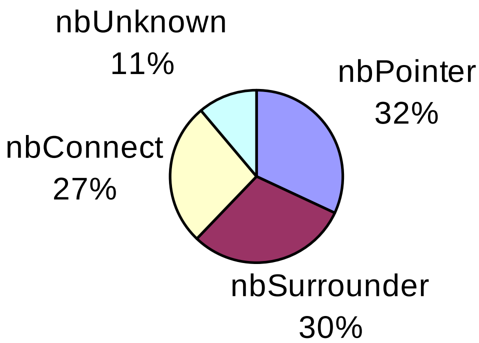

Figure 3.8. Gesture shape statistics.

3.7.3.2 Blind validation of shapes

The logged shapes described above have been labelled manually without the display of the recognised shape. After this blind validation, the repartition of shapes was the following.

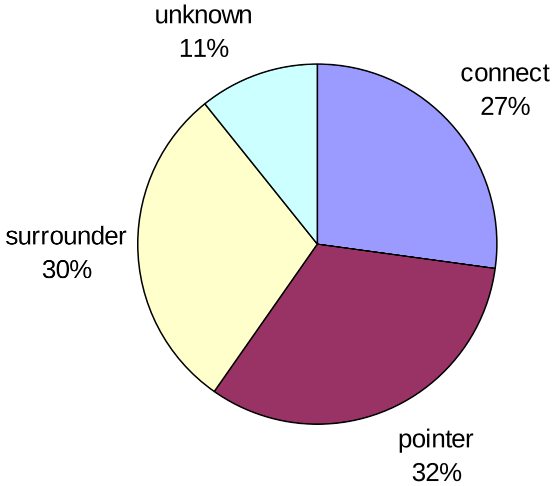

Figure 3.12. Repartion of shapes after blind validation.

The recognition score is 70% : 338 gestures were assigned compatible categories with the manual labeling and the gesture module during execution.

Logged gestures appeared to be more noisy than the gestures observed in the HCA Study system with the tactile screen. Two examples of noisy gestures done with the gyromouse are provided below:

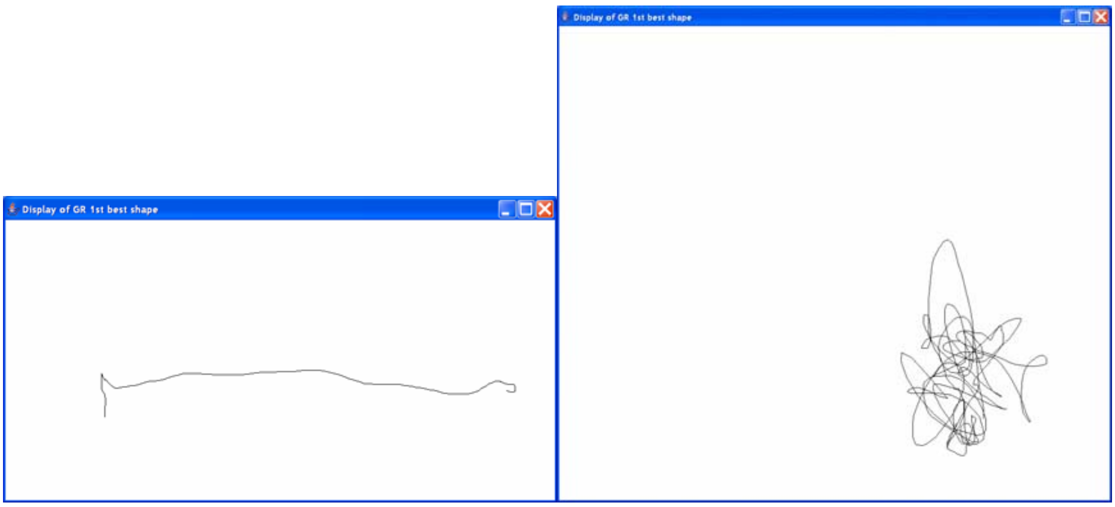

Figure 3.13. Noisy gyromouse gestures.

3.7.4 Gesture Interpretation

3.7.4.1 Type of gesturing interpretation pattern from log files

The quantitative analysis of the GI log files revealed the following repartition of gesturing interpretation type.

USER	SELECT	REF_AMB	NO_OBJ	TOTAL
1	21	4	3	28
2	4	1	3	8
3	31	19	34	84
4	4	3	5	12
5	1	1	1	3
6	0	3	10	13
7	5	4	11	20
8	18	5	10	33
9	37	26	21	84
10	0	1	3	4
11	33	7	40	80
TOTAL	154	74	141	369

Table 3.14. repartition of gesturing interpretation type.

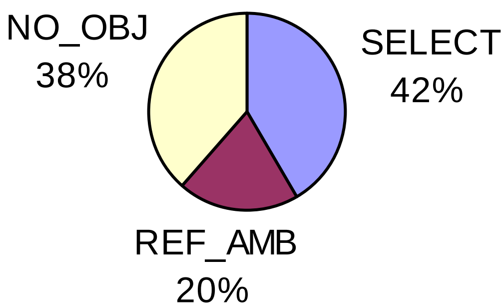

Figure 3.14. Repartion of GI Frame.

3.7.4.2 Gestured objects from log files

The most frequently objects detected by the GI were: the good slot (in 9% of the GI frames), the sword (8%), the magic book, the poison, the valuable slot (7% each), the axe and the hammer (6% each).

3.7.4.3 Video Annotation of Gesture Interpreter Results

The Log files and video recordings of the screen were imported into Anvil and manually compared in order to evaluate GI processing.

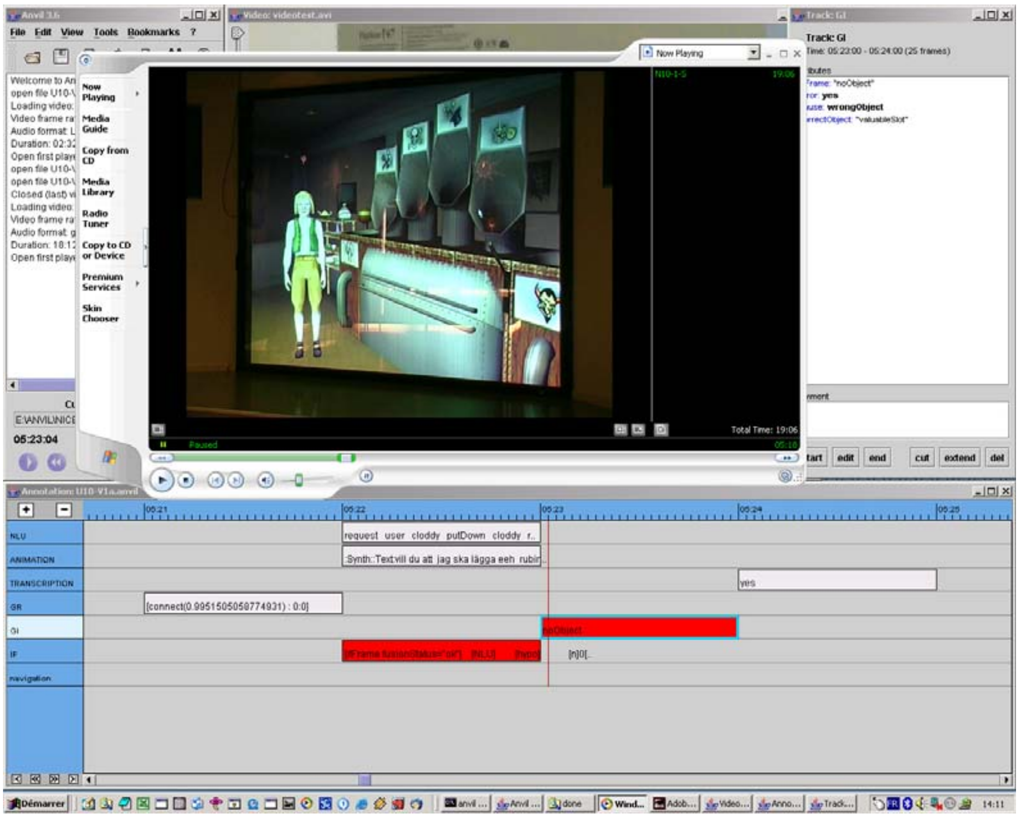

Figure 3.15. A screen shot of the Anvil video annotation tool.

This analysis evidenced a 27% error rate (100 errors out of 369 GI frames).
The errors are distributed as follows:

Errors in the selection of focus objects (slots or objects from the shelf): 50%
Errors in the selection of characters (Cloddy, Tummelisa, poorGirl): 25%
Errors in the selection of FTW objects (drawbridge): 13%
Attempts to select the lever of the machine: 7%
Attempts to navigate (e.g. by pointing to a location): 5%

The focus objects which yielded at least one error were the objects on the shelf (axe, diamond, emerald, goldSack, hammer, magicBook, magiclamp, poison, sword) and the machine slots (good, dangerous, magic, useful, valuable).

3.7.5 Input Fusion

3.7.5.1 Methodology

User tests had not been videotaped for PT1 and the evaluation of the GI and IF, as well as the study of the user's gestural and multimodal behaviour had not been possible (gesture interpretation and multimodal fusion evaluation is not possible without videos since it is not possible to know which object the user really gestured at).

We hereafter describe the study of the gestural and multimodal behaviours of the users during the PT2 user tests held at Teliasonera in December 2005 (11 users were videotaped while using the system). Only the 1st phase of the scenario was analysed (conversational dialogue with Cloddy Hans in HCA Study) for 8 of the users. Gesture and multimodal annotation of video is time consuming as it requires to play several times each video segment in order to understand the global interaction context (speech, gesture, graphics, dialog history, results and errors of each input module).

Fusion was already achieved by the Teliasonera NLU for cases in which gesture arrives before speech. This fusion cases were thus not handled by the IF which had to manage cases for which gesture arrives after speech.

The methodology for studying user's multimodal behaviours and evaluating the IF module was the following. The multimodal behaviours were manually annotated in a spreadsheet file. Each video was played in parallel to the study with the Anvil tool of the results from the NLU, the transcriptions, the GR frames, the GI frames which had been manually annotated and the IF frames.

3.7.5.2 Results of video annotations

74 multimodal combinations were manually annotated. Most involved several problems which led to inadequate fusion (although interaction was finally successful after repetition or modification of user's behavior). 108 occurrences of problems have been identified and classified as follows.

Inadequate use of the gyro mouse (34% of the observed problems). The unusual gyro-mouse led to several interaction problems. The main one was that user would gesture with the mouse without clicking on the button. This might be due to the fact that 1) the pointer is always visible, 2) the user has to hold it all the times forgetting that clicking is required for the selection of an object (which is not the case with the tactile screen), or 3) to the conversational context which led the user think that Cloddy's eyes would notice mouse moves. Indeed, some users would use the cursor as a joint attention mechanism, trying to drive Cloddy's attention to objects or location only with a mouse move (one of the user said "do you see her" while gesturing on TummeLisa without clicking on it). This was not planned by developpers as the graphical application developed by Liquid Media would only provide gesture coordinates when the user clicks. To avoid such problems, users should either be trained to select 3D objects with an unusual media such as the gyro mouse, or the GI/IF should be able to ask for gesture coordinates to the graphical application on demand when an underspecified utterance has been recognised and might be completed with the location of gesture moves. Some users also had problems to click at the right place (in such cases gesture was done too far from objects to lead to their detection), or would click at the wrong time when moving the cursor away from the object. Some users also used the mouse to explore the 3D environment doodling or looking for referenceable objects, much like what was observed in the HCA Study user tests or the preliminary study done by LIMSI with 2D characters.

Wrong detection of object by the GI (17% of the observed problems). The gesture model was based on the selection of either an object or a location. Other semantic functions were observed and not processed: the use of a line gesture to request to move an object to a location, the use of a line gesture to request Cloddy to go to the shelf or to the machine, a gesture on the location where an object had been to put this object back in place. Some gestures were also done to refer to big objects such as the shelf (on which several objects stand) or the machine as a whole (which includes several slots). Proper management of these cases would require the management of several candidate objects since an ambiguous gesture can be used to select a small or a big object such as the shelf or the machine. Finally, some users had difficulties to gesture at the character TummeLisa as she was moving around.

Complex temporal and semantic dialogical combinations of speech and gesture (16% of the observed problems). In such a conversational task in which the selection of 3D objects is combined with navigation (between the shelf and the machine), even a simple "put that there" task leads to complex and disfluent multimodal patterns involving long delays (e.g. 10 seconds) between speech and gestures (hereafter called dialogical multimodality since interpreted as several turns), repetitive or concurrent turns. Indeed simple patterns such as *'take it' + gesture* or *put it there + 2 gestures* were rarely observed. The fact that Cloddy looked cooperative but would not understand complex utterance, would often ask for confirmation, or would provide feedback with a delay might explain this repetitive behaviour from the user or its "complementarity" at a dialogical level. The following is an example in which inappropriate use of the gyro mouse combined with delay in the system's response leads to repetitive behaviour from the user involving speech and gesture: "You can take this (*moves the gyro-mouse cursor in front of the book but does not click on the button*). You can take the book down here. Take it. Take it then. Take the book then <*gesture on the book*> (*Cloddy picks up the book*)".

Problems in NLU or Speech Recognition (15% of the observed problems). Some unexpected utterances or other problems led to no detection of speech or to wrong analysis by the NLU, the error of which propagated in the IF.

Wrong time management (9% of the observed problems). Most of the time, NLU frames did not wait long enough for related GI frames. This was even more true since unusual long delays occurred due to dialogical multimodal patterns.

Other less frequent reasons involve inadequate management of redundant input by the IF.

3.7.5.3 Analysis of log files

The study of the dispatcher log file revealed that noObject messages sent by the GI were much more numerous that the ones logged by the GI it self (141). This resulted in a very high number of IF frames containing a GI noObject message (2967). These frames which were obviously due to a technical communication problem between the modules were not considered in our evaluation.

IF frames with « inconsistency » fusion status (12% of logged IF frames) were observed when the number of variables in speech and gestures were not compatible, for example one user asked for suggestion (no variables in the corresponding lambda expression) and gestured on two objects.

Due to these problems and the seldom use of complementarity of speech and gesture by the user, merging of gesture and speech in lambda expressions in the IF did not occur very often. The manual or automatic comparison of lambda expressions is complex and has not been done to evaluate the number of such replacements. An example of replacement is provided below (the user had said « that one » while pointing on the « useful slot » :


```
[ 11:42:33.015 ]  Received from NLU:
<NLU><hypo><n>0</n><SemanticRepr><lambda><type><function><from><type><function><from><t
ype><simple>location</simple></type></from><to><type><simple>dialogue_act</simple></typ
e></to></function></type></from><to><type><simple>dialogue_act</simple></type></to></fu
nction></type><var><type><function><from><type><simple>location</simple></type></from><
to><type><simple>dialogue_act</simple></type></to></function></type><name>var15</name><
/var><body><apply><type><simple>dialogue_act</simple></type><applicator><term><type><fu
```

```
nction><from><type><simple>location</simple></type></from><to><type><simple>dialogue_ac
t</simple></type></to></function></type><variable/><functor>var15</functor></term></app
licator><applicand><term><type><simple>location</simple></type><score>50</score><functo
r>valuableSlot</functor></term></applicand></apply></body></lambda></SemanticRepr></hyp
o></NLU>
```


The IF reduces the lambda expression by removing the “var15” variable and replacing it by the location specified in a previous gesture frame:


```
[ 11:42:33.046 ]  Received from IF:
<?xmlversion="1.0"encoding="UTF-8"?><!DOCTYPEifFramePUBLIC"ifFrame""if_nice.dtd">
<ifFramefusionStatus="ok"><NLU><hypo><n>0</n><SemanticRepr><apply><type><simple>dialogu
e_act</simple></type><applicator><term><type><simple>location</simple></type><functor>v
aluableslot</functor></term></applicator><applicand><term><type><simple>location</simpl
e></type><score>50</score><functor>valuableSlot</functor></term></applicand></apply></S
emanticRepr></hypo></NLU></ifFrame>
```


3.7.5.4 Conclusions

Those results of the December user tests are compatible with other data provided by Teliasonera on the other user tests sessions held since then. The gyro-mouse was not used very much: there were 275 GI-recognized gestures of 5500 user turns from 57 users (5%of all utterances or 10% of the 2700 utterances that was about picking up objects or putting objects into slots). These 275 gestures are classified as follows: 117 gestures on objects, 117 gestures on slots, 20 gestures on Karen or TummeLisa, 21 on other objects (mill, house, and feather pen). Input gesture was observed to be often used for doodling.

3.8 Usability evaluation

Among the questions addressed by questionnaires and deep interviews were the following: How did users associate different personalities to different fairy-tale characters? How did the users like speaking with the fairy-tale characters? Did the users exhibit conversational behaviour one usually finds in human-human dialogues?

3.8.1 Gameplay and perceived personalities

The interviews were centered around the following questions:

- Tell me what you know about Cloddy Hans?
- What was your task in the game?
- What did you think about this game?
- What did you like the most about the game?
- What did you not like about the game?
- What will computer games be like in the future?

Most users reported that it was quite natural to use speech in games and many expected that games will be like this in the future. Some users apparently regarded the speech technology component of the game as part of the “puzzle” to be solved, with inherent limitations such as restricted vocabulary etc. being thought of as deliberately designed obstacles. The sluggishness of Cloddy Hans was in the same way perceived by some users as being part of a deliberate design (which was the case) with the intention of making the game harder (which was not the main purpose). Similarly, the negotiation with Karin was considered a fun part of the game by many users. A few users insisted on that speaking with the characters in the NICE system was (almost) like talking to real persons.

Judging from the interviews, the game seems generally to have been perceived as fun, interesting and non-irritating even by users who found it difficult. This is supported by Figure 3.18 which shows the distribution of all users’ answers in the questionnaire and Table 3.15, which shows the median scores for questions about the game play across all four sub-corpora.

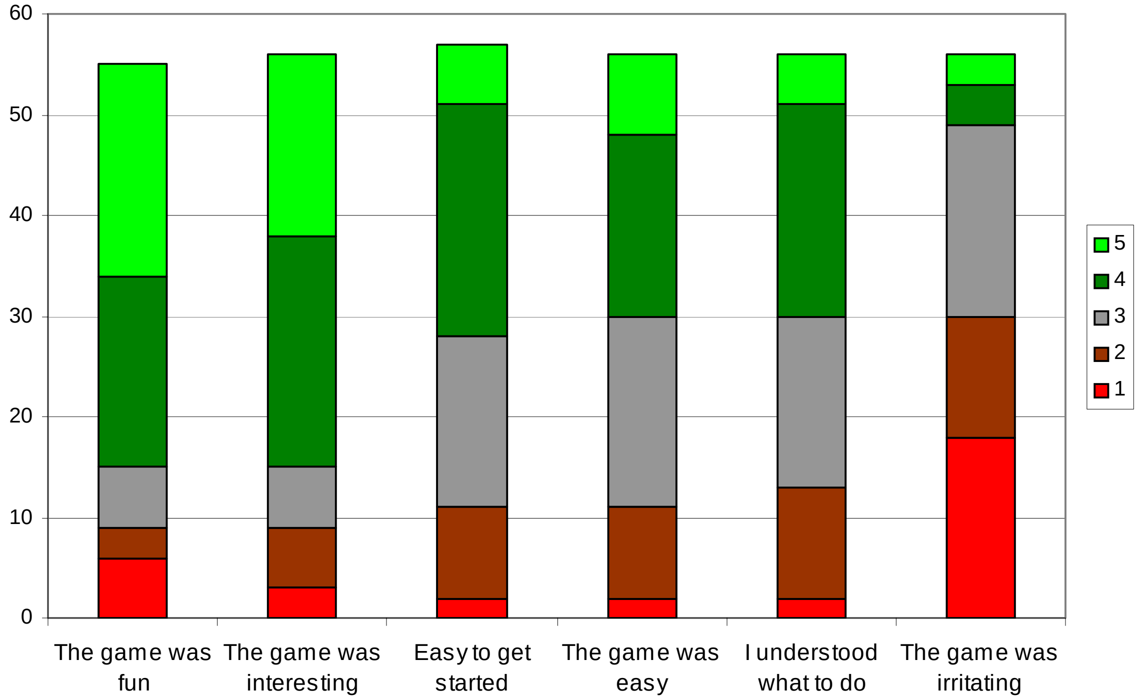

Figure 3.18. Distribution of all users' ratings of the system as a whole.

Question	Median scores
The game was fun	4.0
The game was interesting	4.0
It was easy to get started	4.0
The game was easy	3.0
I understood what to do	3.5
The game was irritating	2.0

Table 3.15. Median scores for questions about the game play across all four sub corpora

In the interviews, users unanimously reported that Cloddy Hans was a bit slow, but kind, while Karin being rather the opposite. Non-communicative as well as verbal and non-verbal behaviour of the two characters Cloddy Hans and Karin had been designed to convey differences in personality along several dimensions in the so-called OCEAN model (Gustafson, Bell, Boye, Lindström, & Wirén, 2004; McCrae & Costa, 1996). Analyses of data obtained from the post-experiment questionnaires showed that the two characters were indeed perceived as having different personalities in several respects.

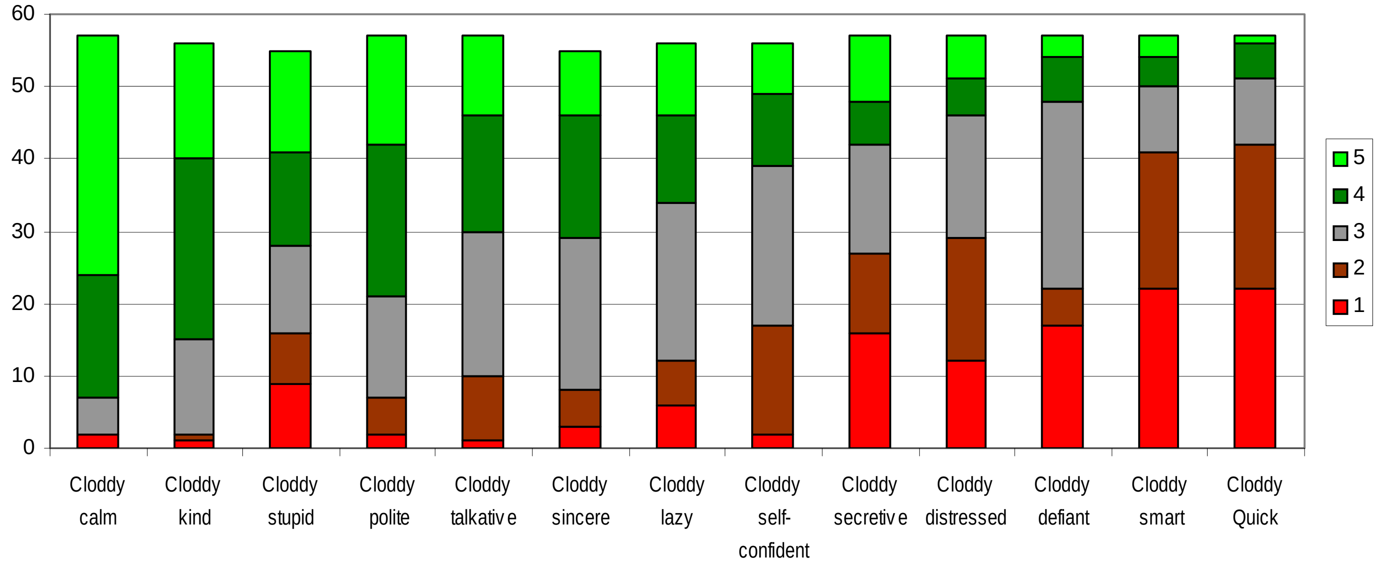

Figure 3.19. Distribution of all users' ratings of Clody Hans' personality traits.

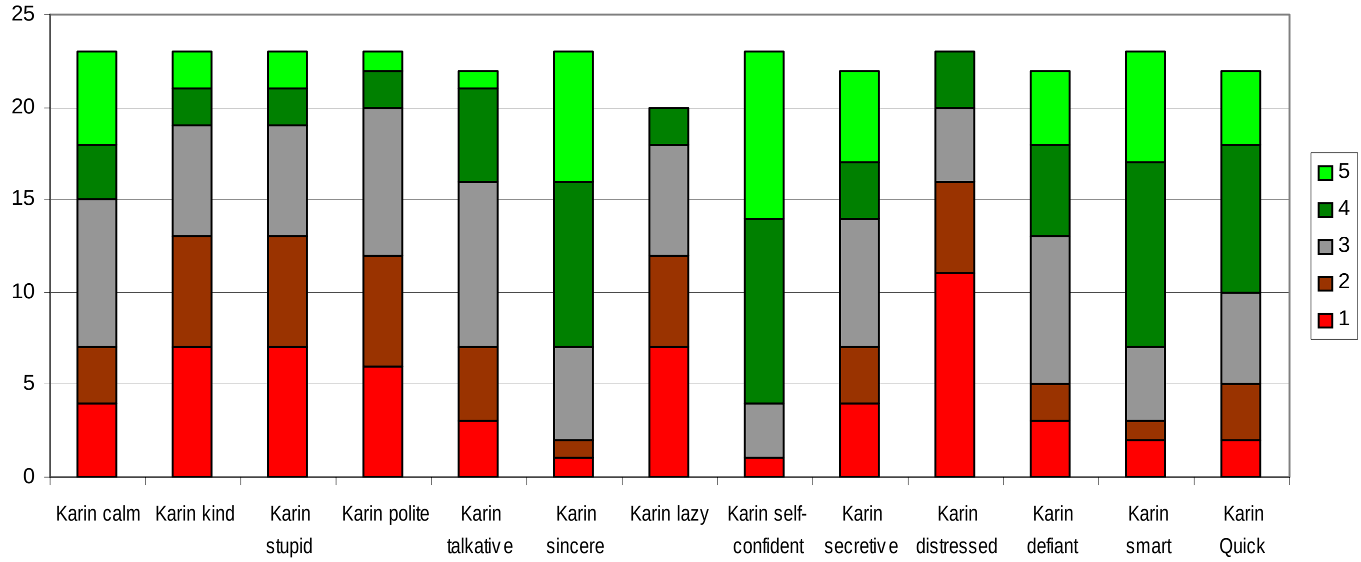

Figure 3.20. Distribution of all users' ratings of Karin's personality traits.

Table 3.16 shows which of the two characters displayed each trait in the most salient way, as judged by the users in Lab 2 and 3, who all interacted with both Karin and Cloddy Hans.

loddy Hans	Karin	Not significant
Kind	Smart	Defiant
Stupid	Quick	Secretive
Lazy	Self-confident	Sincere
Calm		Talkative
Polite		
Distressed		

Table 3.16. User judgments regarding which animated character displayed specific personality traits in the most salient way, based on questionnaire data from Lab 2 and 3. Differences between Cloddy Hans and Karin were tested for significance using Wilcoxon Signed Ranks Test ($p<0.05$).

The cases where no significant difference between Karin and Cloddy Hans could be found, can probably be explained by the fact that quite a few children had difficulties in understanding the words used to describe these traits, and therefore asked the experimenters about their meaning.

3.8.2 Perceived understanding capabilities and naturalness of the interaction

The system design was aimed at getting the users to perceive Cloddy Hans as friendly but a bit stupid, and Karin as being determined and smart. To evaluate if the children had perceived the intended understanding capabilities, they were asked to rate how much they agreed on the following statements on a Likert scale:

- I understood what Cloddy said
- Cloddy did what I said
- Cloddy understood what I said
- Talking with Cloddy was like talking with a human
- Cloddy helped me when things went wrong
- Cloddy did what he wanted
- I understood what Karin said
- Karin did what I said
- Karin understood what I said
- Talking with Karin was like talking with a human
- Karin helped me when things went wrong
- Karin did what she wanted

The users was also asked about how the perceived their interactions with Cloddy and Karin in the subsequent in-depth interviews.

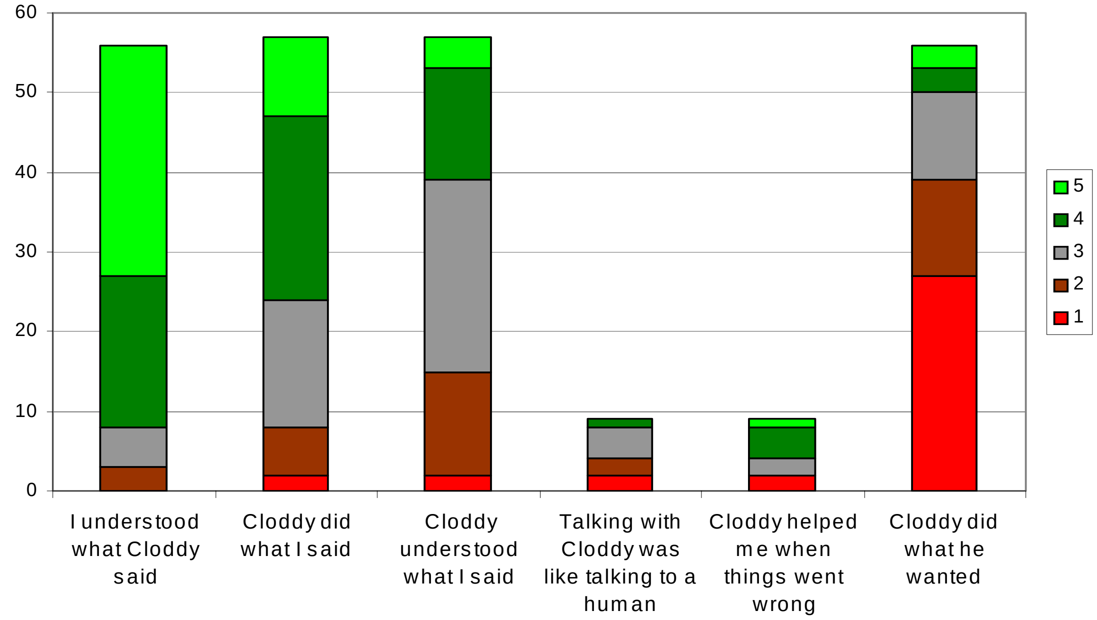

Figure 3.16. Distribution of all users' ratings of Cloddy Hans' conversational abilities.

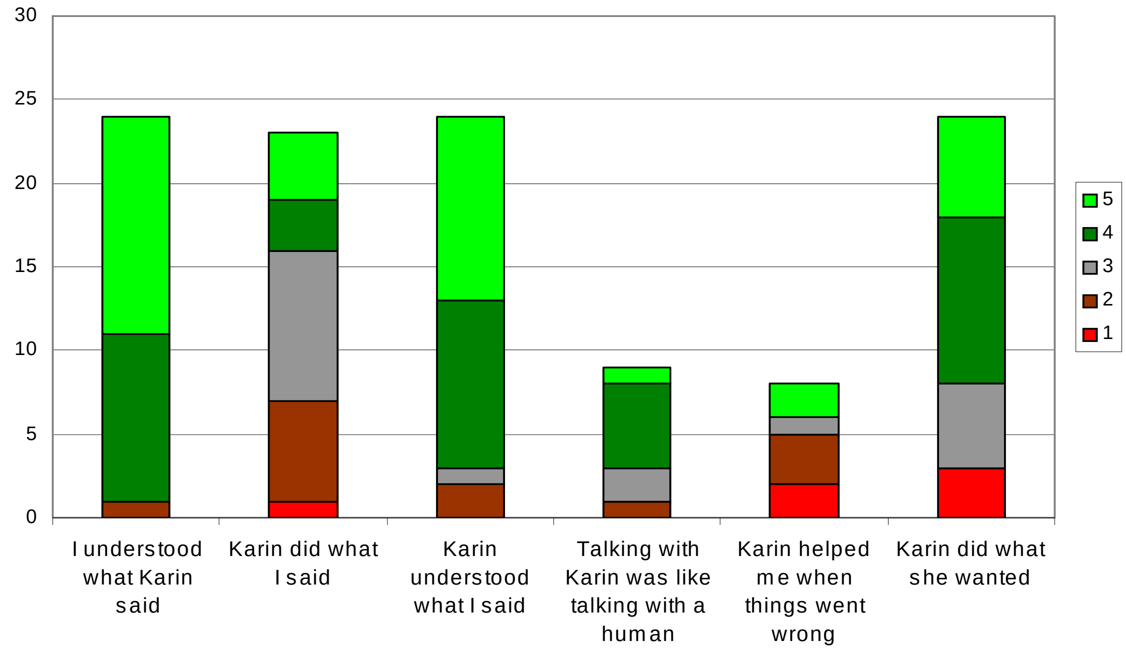

Figure 3.17. Distribution of all users' ratings of Karin's conversational abilities.

Wilcoxon signed rank tests were used to test for significant differences in the ordinal scale judgments of Cloddy Hans' and Karin's conversational abilities. These tests showed among other things that the users felt that Cloddy Hans was more obedient than Karin. Karin, on the other hand, was perceived as understanding what the users said better; was understood better by the users; was more like talking to a real human being; and did what she wanted herself more than

Cloddy Hans did. The characters were not perceived as different when it came to helping the users in case of errors.

3.9 Effects of the conversational behaviour of Cloddy Hans

Cloddy Hans' conversational behavior was correlated with a number of judgments in the questionnaire using Pearson's correlation coefficient, and the significance of these correlations was tested. Although this operation involves correlating numerical variables with ordinal ones we feel that this might be justifiable as we are only searching for possible dependencies between variables rather than assessing their absolute influence.

3.9.1 The conversational behaviour of Cloddy Hans and entertainment and ease-of-use aspects of the game

The way Cloddy Hans behaved in the conversation (described above) had no significant effect on how the users judged the entertainment and ease-of-use aspects of the game. Cloddy's behavior neither influenced whether the game was perceived as easy, funny, annoying or interesting, nor if the users thought it was easy to get started with and understood what to do.

3.9.2 Cloddy Hans conversational behaviour and the experience of his conversational abilities

However, Cloddy Hans' conversational behavior did influence how the users judged various aspects of his conversational abilities. Several significant positive or negative correlations ($p<.05$) were found.

One aspect of his conversational abilities was whether Cloddy understood what the users said, and a significant negative correlation with the proportion of safe utterances, and significant positive correlations with the proportion bad and domain utterances were found. There were also positive correlations between the proportion of meta and domain utterances and whether the users felt that Cloddy did what they told him, and a negative correlation with safe utterances. Another question was whether talking to Cloddy Hans was like talking to a real human being. Here, a negative correlation with the proportion of safe utterances, and a positive correlation with the proportion of meta utterances were found. Regarding the question whether Cloddy helped them when things got wrong, we found a negative correlation with the proportion of problem utterances. Finally, there was a negative correlation between the proportion safe utterances and the feeling whether Cloddy did what he wanted.

3.9.3 Cloddy's conversational behaviour and the experience of his personality traits

Cloddy Hans' behaviour also influenced how the users judged his personality traits. In particular, the proportion of problem utterances was positively correlated, and the proportion of bad utterances was negatively correlated with Cloddy being perceived as kind. Furthermore, the proportion nice utterances was positively correlated with Cloddy being polite. None of his other character traits were affected by his conversational behaviour, so that that the traits *smart* and *stupid* were not affected by Cloddy Hans' conversational behaviour.

3.10 Efficiency measures (number of user turns)

3.10.1 Effect on Cloddy Hans' perceived conversational abilities

We observed some dependencies between measures of dialogue efficiency and Cloddy Hans' conversational abilities. The number of user turns per phase in the dialogue in the fairy-tale machine scene was negatively correlated with if the users felt that Cloddy understood what they said; that Cloddy did what they said; and that talking to Cloddy was like talking to a real human being. That is the more user turns per dialogue phase, the less the users felt that Cloddy understood them, did what they told him and was like talking to a human being. Table 3.17 shows positive/negative Pearson correlations that are significant at the 0.05 level (2-tailed).

	Cloddy understood what I said	Cloddy did what I said	I understood what Cloddy said	Talking to Cloddy was like talking to a human	Cloddy helped me when things got wrong	Cloddy did what he wanted
Total number of phases	+					
Total number of user turns per phase	-					
Number of phases in machine scene			+			
Number of turns in machine scene		-			-	
Number of user turns per phase in machine scene	-	-		-		
Number of phases in world scene	+					
Number of turns in world scene						
Number of user turns per phase in world scene						

Table 3.17. Positive and negative Pearson correlations between efficiency measures and Cloddy Hans' conversational abilities that are significant at the 0.05 level (2-tailed).

3.10.2 Effect on Cloddy Hans perceived personal traits

There were also several correlations between measures of dialogue efficiency and Cloddy Hans perceived personality traits. For example, when there were many turns and dialogue phases in the machine scene Cloddy Hans was perceived as kinder, more polite, less stupid, but also more distressed. Positive and negative Pearson correlations between efficiency measures and Cloddy Hans' personal characteristics that are significant at the 0.05 level (2-tailed) are shown in Tables 3.18 and 3.19.

	Kind	Smart	Sincere	Quick	Lazy	Talkative	Stupid
Total number of phases							
Total number of user turns per phase		-					
Number of phases in machine scene	+						-
Number of turns in machine scene							-
Number of user turns per phase in machine scene							
Number of phases in world scene			+				
Number of turns in world scene							

Number of user turns per phase in world scene							

Table 3.18. Positive and negative Pearson correlations between efficiency measures and Cloddy Hans' personal characteristics that are significant at the 0.05 level.

	Secretive	Calm	Self-confident	Polite	Defiant	Distressed
Total number of phases		-				
Total number of user turns per phase						
Number of phases in machine scene						-
Number of turns in machine scene				+		-
Number of user turns per phase in machine scene						-
Number of phases in world scene						
Number of turns in world scene		-				
Number of user turns per phase in world scene						

Table 3.19. Positive and negative Pearson correlations between efficiency measures and Cloddy Hans' personal characteristics that are significant at the 0.05 level.

3.11 Dialogue phenomena

Several types of dialogue behaviour were observed on the part of individual users, indicating a high degree of social involvement with the characters. In addition to insulting the rather dunce Cloddy Hans, these behaviours included:

- either taking Karin's or Cloddy Hans's part when one of them offended the other,
- showing repent when being accused of deceipt,
- lying, making ironic, sarcastic and humorous remarks,
- reacting to the character's mood and adding politeness markers and explicit appeals in order to cheer the character up and thereby achieve the user's goals,
- repeated efforts of persuasion attempting to convince a reluctant Cloddy Hans to pick up a particular item or hand over items to Karin, and
- lecturing Cloddy Hans while making reference to common dialogue history.

Several dialogue excerpts exemplifying some of these dialogue behaviours are exemplified in the following sections.

3.11.1 Taking sides

The plot design included a traditional dramaturgic element of conflict, namely between Cloddy Hans and the girl with the red shoes, Karin, who acts as a kind of gate-keeper at the bridge in an effort to promote negotiative dialogue and possibly also three-party interaction. One further objective was to see if users would in any way engage in the conflict, e.g. by choosing sides. Very little is ever conveyed about the reason for the conflict by the two animated characters engaged in it, but as can be seen in the dialogue examples below, each character tries to project their own view of the other character onto the user, finally provoking the latter to take sides. Interestingly, in many cases this actually prompted the users to quite clearly take a standpoint as to which of the two characters was on the user's side. This was evident since the provocation in many cases elicited highly spontaneous emotively coloured verbal feed-back from the user. As can be seen in the examples below, some users sided by Cloddy Hans and some by Karin.

Karin	Why are you dragging along that cloddy hans by the way
User	Because he is fun
Cloddy	I don't like talking to her

Karin	What do you do in your spare time
User	Play soccer and practice athletics
Karin	I like to dance
User	Yes I also like dancing
Cloddy	She is quite often a bit grumpy
Karin	If you want me to lower the bridge I want something in return
User	You can have the book
Cloddy	She is known to take bribes
Karin	If you have nothing better to offer the deal is off
User	Okey
Karin	I thought I saw something falling down along with that dung-hill
User	The sword
Karin	Go find it and then let's discuss
Karin	Why are you dragging along that cloddy hans by the way
User	I don't know

Karin	Why are you dragging along that cloddy hans by the way
User	He refuses to go away
Cloddy	Uh-huh
User	Put the sword down

Karin	Why are you dragging along that cloddy hans by the way
Cloddy	What did you just say about me?
User	Well I don't know actually
Karin	Be careful so that you don't end up like him
User	I wouldn't think so!
Karin	Yes isn't he a pain in the ass
User	Uh-huh a real pain
Cloddy	Er what did you just say about me
User	You were ugly
Cloddy	Thanks for nothing
User	I have given you a lot
Cloddy	How mean
User	No-o

This excerpt starts in a situation where the user is trying to persuade Karin to let the user pass over the bridge.

Karin	Why do you keep dragging along that Cloddy Hans figure, by the way
User	Because he is fair, a lot more so than you, in any case
Karin	If you want me to lower the bridge, I want something in return
Cloddy	She is so snappy
User	You are very very very kind

Eventually the user and Cloddy Hans are let over the bridge, at which point the dialogue continues:

Cloddy	Do you think we should give her our stuff or not?
User	Give it to her - we have no use for it
Cloddy	But I don't care!
Karin	You fooled me! I will remember this!

Cloddy	Yes! Now we have crossed the bridge
Karin	Hey – wasn't I supposed to get something in return for letting you across the bridge?
User	Yes
Cloddy	What do you want me to do?
User	Give her the emerald
Cloddy	OK. (Cloddy Hans hands over the emerald to Karin)
Karin	Thanks! It is a pleasure doing business with the two of you
User	I know! (sarcastically)
Karin	I wish you a pleasant stay in the fairy tale world!
User	OK
Cloddy	Do you want me to go the right?
User	Yes.
Cloddy	I have a bit of a problem in telling right and left apart, I never learned that as a child
User	But then go to the left!
Cloddy	I have a bit of a problem with right and left
User	But go straight ahead, then!
Cloddy	Do you want me to go over there? (starts walking towards the user)
User	No, you are supposed to turn around and go back!
Cloddy	My brain is disconnected
User	And this occurred to you only now, or what?

This example illustrates how the user sides with Cloddy Hans and gives him explicit positive verbal feed-back. On the other hand, when Cloddy Hans displays what the user thinks is dishonest behaviour, and is included in Karin's accusation, the user chooses to stand up for her own values, in a way which seems to reflect the socio-cultural rules that would apply in similar interaction between humans.

3.11.2 Freedom of expression

User	put the hammer by er that tube with the hammer and key
Cloddy	is it ok if we name that useful
User	er put the magic wand in the tube with the er hat
Cloddy	uh-huh
Cloddy	oh-kay

user	take the bag in your hand
cloddy	do you really think so?
user	yes
user	take the bag in your hand the sack ehh
cloddy	I'm already holding the sack
user	take the book from the shelf as well
cloddy	I can anly carry one thing at the time
user	but put the sack into the fairy-tale machine

This example shows how the user is free to use her own words to describe the objects. Most users picked up the terminology used by Cloddy Hans for the different slots of the fairy-tale machine, and several reported (in the interviews) that figuring out which words to use constituted one of the intricacies of the game itself.

3.11.3 Ad lib. excursions (tagged as out of phase)

Unlike interfaces based on direct manipulation and menus, spoken language interfaces allow for reference to abstract entities, as well as to entities which are not currently visible, tangible or audible. Several users also made use of the possibility to temporarily diverge (digress?) from the current topic, make minor excursions and later return to the task at hand, as in this example where a user enters a sub-dialogue about Thumbelina, who walks by:

user	try pulling the lever
user	pick up that little man
cloddy	now I don't understand
cloddy	oh you mean thumbelina
user	put down*
cloddy	she is so small that you can't hear her
user	ask her something
cloddy	I don't think she wants that
user	put something from the shelf in one of the tubes
cloddy	uh-huh
user	ask thumbelina to do something
cloddy	no I don't want to do that
user	are you afraid?
user	coward (LAUGHTER)
cloddy	what did you say about me?
user	what did you say?
cloddy	what did you say about me?
user	coward
user	coward
user	(LAUGHTER)
cloddy	thank you for nothing
cloddy	o-kay…
user	put that lamp*
cloddy	o-kay…
user	in the pipe to your farthest right
user	pull the lever
cloddy	I don't want to
user	coward
cloddy	you have yourself to blame
user	now what
cloddy	uh-huh
user	well what are you supposed to do now
cloddy	i am listening
user	can we go back to the previous image/scene

This dialogue also shows how Cloddy Hans is in fact insulted by the user, and clearly shows that, both verbally and by facial gestures. This dialogue snippet ends with an example of meta-reference to the entire previous scene, which the user obviously thinks is under the control of, and can be understood by, the animated character, although he is part of it.

3.11.4 Motivating the ECA

The following dialogue excerpt shows how a user, upon the explicit request of Cloddy Hans, motivates the latter by explaining the goal of the exercise. This particular user was indeed very satisfied with the game experience, despite being an experienced user of shoot-’em-up-games and strategic war games. This clearly shows that more or less political claims made within by

Shneiderman and others regarding the inappropriateness of social metaphors in human–computer interfaces (Shneiderman 2000), does not apply to this class of applications.

cloddy	I'm going to put the emerald in valuable
user	go back to the shelf
user	go back to the shelf
cloddy	I'm going over to the shelf
cloddy	what do you want me to do
user	can you fetch the wand
cloddy	uh-huh
user	please
cloddy	I'm going to fetch the wand
user	go back to the fairy-tale machine
cloddy	uh-huh
cloddy	why do you want me to do that
user	to put the wand in the magic slot
cloddy	I'm going to walk over to the machine
user	put the wand in the magic tube
cloddy	I'm going to put the wand in magic

3.11.5 Wrongful accusations

In this example, the user is convinced that Cloddy Hans is still holding the diamond despite being told twice that this is not the case. When the user finally discovers his mistake, he comments on that in a high-pitched, clearly apologetic voice, saying that it seems that Cloddy Hans was right, after all. This type of behaviour cannot be classified as anything else than socially naturalistic.

karin	look behind you in the grass
user	go to the er emerald
karin	you'd better concentrate wasn't there a shimmer in the grass over there
user	would you please walk over to the emerald
cloddy	uh-huh
cloddy	don't you want me to hold the emerald
user	yes I want you to hold the emerald
cloddy	no really
user	oh yes I want you to hold the emerald it would be most kind if you could walk over to it now
cloddy	oh well could you wait a minute and we will see…
cloddy	I'm going to the diamond
user	would you mind letting go of the diamond
cloddy	okay
cloddy	<ask-for_request>
user	let go of the diamond
cloddy	I've already done that
user	put the diamond down
user	put the diamond down Cloddy H*
cloddy	I've already done that
user	oh it seems you actually did! Could you pick up the emerald
cloddy	I'm going to pick up the emerald

3.11.6 Common ground, co-reference and multi-party dialogue

These two sections from one user's interaction illustrate several desirable features of the fairy-tale system. First, the user asks Cloddy Hans to walk over to the bridge, but as he approaches the girl, the user realizes that he has forgotten her name, but remembers that Cloddy Hans knows it. The user makes explicit reference to the dialogue history shared by himself and Cloddy Hans by use of the discourse particles "nu då" ("now again"). This is yet another feature, typical of human–human dialogue, and its presence here ought to be indicative of quite some degree of perceived naturalness.

cloddy	I'm going to the bridge
user	go to the bridge
cloddy	GOTO
user	what* .. what was her name now again
cloddy	her name is karin
user	oh yes
user	karin

(a number of turns later, after fetching the emerald)

user	would you be so kind as to give her the emerald
user	what was her name now again
user	would you mind*
cloddy	her name is karin
user	karin

When later finding himself in the same slightly embarrassing memory-lapse situation a second time, the user turns to Cloddy Hans, lowers his voice and makes it sound more intimate, probably to avoid being overheard by the girl whose name he has forgotten. These examples fit quite well with the "media equation", stating that people respond to the mediated world and the real world in the same fundamentally social and natural way (Reeves and Nass 1996).

In the following example, the Wizard took over control over the movements of Cloddy Hans as the latter was crossing the bridge, and surprised the user somewhat by letting Cloddy Hans seemingly climb the rail of the bridge and then levitate a couple of meters up in the air. The user reacts to this and bashes Cloddy Hans and in doing so refers to the spatial domain ("What are you doing *up there*?"). What is interesting is that the user here also immediately assumes that the same rules that apply to little children, kittens etc., namely that climbing trees inevitably means trouble, also should apply to a cartoon character.

user	would you mind crossing the bridge
user	would you mind*
cloddy	take it easy I am dashing
user	what are you doing up there come down immediately cloddy hans!
cloddy	take it easy I am dashing
user	would you mind*

3.12 Conclusions

Through a series of both quantitative and qualitative analyses based on data from 57 users interaction with the NICE fairy-tale prototype, we have shown how the original goals of the project have been met to a large degree, with users perceiving the game as unconstrained, fun and entertaining and with many examples of natural interaction, previously unseen in state-of-the-art multimodal man–machine interaction.

Specifically, in relation to the general requirements of NICE deliverable D1.1.b, the following can be noted:

1. Spoken multimodal dialogue indeed turned out to be the main vehicle of progression for the users, and being able to speak with the characters was also what many users thought made the game fun, according to the interviews.
2. The prototype seems to have been very well suited to the target group (children and adolescents). Evaluation scores and interviews show that the users found the game both fun and interesting, with some of the older children reporting that the plot was probably better suited for younger children.
3. The second prototype includes Cloddy Hans, Thumbelina and Karin, who are all (loosely) modeled on characters from H. C. Andersen's fairy-tales, and who have separate modules for spoken language understanding, dialogue modelling, as well as for multimodal generation.
4. Unlike in the first prototype, where explicit instructions were spoken by a narrator prior to the game, in the second prototype it was the helper character Cloddy Hans who explained the situation and the task at hand to the user as part of the initial scene. This method for enrollment turned out to be successful in the sense that all users completed all tasks in the introductory scene. Some users reported that they would have wanted more explicit instructions on what their task was, while others explicitly report that finding out how to interact multimodally and what to say was an intriguing and novel element constituting yet another "puzzle" to be solved as an integral part of the game.
5. The introductory scene, which was designed with the purpose of familiarizing the user with multimodal dialogue, was perceived as part of the overall plot, and fun and entertaining, as can be seen by comparing ratings by users who played scene one only with ratings by users who played both scenes.
6. The user interaction data shows many examples of how users have taken quite different paths through the game, changed their mind in the middle of sentences and been able to make excursions from the topic and still come back on track later. When interviewed about the degree of self-control, the only parameter that seems to have affected that is the personality of the two characters Cloddy Hans and Karin, with Karin being perceived as more dominant.
7. From the questionnaires as well as from the interviews, it can be concluded that the game was perceived by most users as fun, non-irritating and interesting, regardless of the amount of time spent playing or the number of scenes played.

3.13 Lessons learned and future work

In the original specification of the NICE project, a number of predictions were made regarding the potential of combining spoken and multimodal dialogue with technology, metaphors, know-how and certain dramaturgical devices from character-based computer games. Many of these have been proven correct in the course of the second Fairy-Tale World prototype evaluation, but should come as no big surprise. These include

- that the addition of spoken and multimodal dialogue should create a positive user experience, and a sense of freedom and self-control among the users
- that the tasks as well as the user interface should be readily understood and appreciated by the target user group since this type of interaction is natural
- that it should be possible to use spoken language as the main device for story progression

Some of the goals that the second prototype set out to reach (and indeed reached) were less self-evident, but not completely unexpected, including

- that Likert-type scales should be possible to use to rate different aspects of the characters and the game even with very young users (the youngest being 8 years)
- that the addition of spoken language and social relations between several animated characters and the user (sometimes in three-party dialogue) actually did evoke social behaviour and emotively coloured reactions
- that co-ordinated physical, verbal and non-verbal apparition and behaviour should make it possible to design characters which were perceived as having fundamentally different personalities and conversational abilities
- that three-party dialogue with several animated characters actually worked, probably due to the fact that users regarded the different characters as separate entities who do not always hear or understand each other. This made it possible for Cloddy to help the user in the dialogue with Karin by suggesting what to offer her, and conversely for Karin to tell the user to bring something shiny, in which case the user almost always asked Cloddy to go and fetch a jewel. This is means that it is possible to overcome the shortcomings of the speech recognizer, by letting the system tell the users (via Karin) what to say to Cloddy in the next turn. An indication that that this seemingly simple trick "worked", can be observed from the fact that users rated Cloddy as stupid and Karin as smart even though the trick was used in both directions.

However, what came as more of surprise to the experimenters was

- that no effect could be found of how well the dialogue went, how many subtasks were solved or how much time was spent interacting with the system on the users' perception of how fun and interesting the game was
- the perceived personality traits *smart* and *stupid* were NOT affected by the conversational behaviour of Cloddy Hans, e.g. the amount of his turns which explicitly indicated that he had not understood the user (tagged as problem) and the ones where he indicating unceratinty (labeled as safe). However these did affect the users' rating of his conversational capabilities. This apparent contradiction might be explained by the strong persona design of Cloddy Hans, which involded slow sluggish speaking style, and a large

amount of hestiational verbal behaviour (filled pauses and plogonguations), and because the children considered the task he had to carry out was quite simple.

- that several users explicitly perceived shortcomings of the natural language interface (especially the understanding components) as part of the game, constituting an obstacle to overcome

Finally, observations of the children's gestural and multimodal input leads us to question the adequacy of a gyro-mouse for gesture input in the context of this particular application (unless users are trained to use it to gesture at 3D object, or if the system becomes able to request on demand the status and location of the mouse). Repetitive multimodal behaviors could for example be avoided through faster feedback on their gestural input. Finally, temporal management in the IF can be improved with the management of start of speech and start of gesture.

3.14 References

Bell, L. (2003). *Linguistic adaptations in spoken human-computer dialogues: Empirical studies of user behaviour*. Stockholm: KTH Speech, Music and Hearing.

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human–computer interaction. In *Proceedings of ICPhS 03,* . Barcelona.

Boros, M., Eckert, W., Gallwitz, F., Görz, G., Hanrieder, G., & Niemann, H. (1996). Towards understanding spontaneous speech: Word accuracy vs concept accuracy. In *Proc. ISCLP'96* (pp. 1009-1012).

Gerosa, M., & Giuliani, D. (2004). Investigating automatic recognition of non-native children's speech. In *Proceedings of ICSLP* (pp. 1521-1524).

Giuliani, D., & Gerosa, M. (2003). Investigating recognition of children's speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China* (pp. 137-140).

Gustafson, J., & Bell, L. (2000). Speech technology on trial: Experiences from the August system. *Journal of Natural Language Engineering: Special issue on best practice in spoken dialogue systems, 6*(3-4), 273-286.

Gustafson, J., Bell, L., Boye, J., Lindström, A., & Wirén, M. (2004). The NICE Fairy-tale Game System. In *Proc. of 5th SIGdial Workshop on Discourse and Dialogue*. Cambridge, MA: NAACL.

Gustafson, J., & Sjölander, K. (2002). Voice transformations for improving children's speech recognition in a publicly available dialogue system. In *Proceedings of ICSLP 2002*.

Hagen, A., Pellom, B., & Cole, R. (2003). Children's speech recognition with application to interactive books and tutors. In *Proceedings of IEEE ASRU Workshop*.

Li, & Russell, M. (2002). An analysis of the causes of increased error rates in children speech recognition. In *Proceedings of ICSLP 2002*.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology, 140*, 1-55.

McCrae, R., & Costa, P. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives*. New York: Guilford.

Potamianos, A. (2003). Robust recognition for children's speech. *IEEE Transactions on Speech and Audio processing, 11*(6).

Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.

Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM, 43*(9), 63-65.

Sjölander, K. (1997). *The Snack Sound Toolkit*, from http://www.speech.kth.se/snack/

Sjölander, K., & Gustafson, J. (2004). Voice creation for conversational fairy-tale characters. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*. Pittsburgh.

Appendix 1a

A complete dialogue example constructed of phases from different users.

speaker	verbal	utterance type	sub-task inititalizer	phase (user ID)
user	hej	social	user	cloddy social (Eke_fp02)
cloddy	hej själv	nice		
user	vad heter du	social	user	
cloddy	jag heter dummerjöns	nice		
user	jag heter NN	social		
cloddy	hur gammal är du	nice	cloddy	
user	elva	social		
user	vad är det där för något	domain	user	machine plot description (Eke_fp26)
cloddy	<introduce_machine_plot_goal>	domain		
user	okej...är det bara att köra på	meta		
cloddy	du kan prata med mig och be mig göra saker	domain		
user	ta den där saken där i hyllan	domain	user	first object to machine (Eke_fp19)
user	hallå	meta		
cloddy	hallå själv	meta		
user	gå till sagohyllan	domain	user	
cloddy	vill du att eeh , jag ska gå bort till , hyllan	safe		
user	ja…gör det	yes		
cloddy	då gör jag väl det *[cloddy walks over to the shelf]*	meta		
user	ta något ifrån hyllan	domain		
user	be tummelisa göra nåt	domain	user	out of phase (Han2_fp11)
cloddy	nä det vill jag inte	bad		
user	är du feg…fegis ha ha ha	test		
cloddy	vad sa du om mig	bad		
user	fegis	test		
user	ta något från hyllan	repetition		first object to machine (Eke_fp19)
cloddy	vilken vill du att , jag ska ta	safe		
user	den där ..säcken visst är det	domain	user	
cloddy	vill du att , jag ska ta , säcken	safe		
user	ja	yes		
cloddy	då gör jag väl det *[cloddy picks up sack]*	meta		
cloddy	vad ska vi göra	safe		
user	ta boken ifrån hyllan också	domain		
cloddy	jag kan bara bära en sak i taget	problem		
user	men lägg i säcken där i sagomaskin *[user clicks at useful slot]*	domain	user	
cloddy	är det , säcken , du vill att , jag ska lägga eeh , där	safe		

user	ja	yes		
user	lägg påsen i sagomaskinen	repetition		
cloddy	vill du att , jag ska lägga eeh , säcken i , #nyttigt	safe		
user	ja	yes		
cloddy	då gör jag väl det *[cloddy puts sack in useful slot]*	meta		
user	kan du dra i spaken	domain	user	pull lever (Han2_fp12)
cloddy	jag vill inte	bad		
user	varför inte	meta		
cloddy	om man drar i spaken skapas en saga	domain		
user	varför vill du inte det	meta		
cloddy	mm jag får egentligen inte röra sagomaskinen	domain		
user	men det vill jag	meta		
cloddy	mm jag är ju ensam i sagoverkstan idag	domain		
user	okej	meta		
cloddy	jaa vi kanske kan ta oss en titt	meta		
user	mmm det tycker jag	yes		
cloddy	vill du att jag drar i spaken	safe		
user	mmm	yes		
cloddy	jag ska eeh , dra i spaken *[cloddy pulls lever falls down into the world}*	safe		
user	var är vi nu någonstans då	repetition		world intro (Han1_fp16)
cloddy	nu är vi i sagovärlden	domain		
cloddy	vad ska vi göra	safe		
user	jag vet inte	meta		
cloddy	inte jag heler	meta		
user	var snäll och gå upp gå och ta upp svärdet	domain	user	explore world (Han2_fp06)
cloddy	vill du att , jag ska ta , svärdet	safe		
user	ja	yes		
cloddy	bra ide *[cloddy walks over to the sword]*	meta		
cloddy	jag ska ta, svärdet *[cloddy picks up the sword]*	safe		
user	tackar hugg i trädet med svärdet	test		
cloddy	jag vill inte	bad		
user	gå bort till tjejen därborta	domain	user	
cloddy	vill du att eeh , jag ska gå bort till , bron	safe		
user	aa	yes		
cloddy	ja varför inte *[cloddy walks over to the bridge]*	meta		
user	tack …BUURP…ha ha ha	social		
user	kan du vara snäll att prata med flickan där framme	repetition	user	karin intro (Han2_fp11)
cloddy	hon heter karin	domain		
user	kan du vara snäll och prata med karin	repetition		
user	fråga karin om hon behöver hjälp	repetition		
cloddy	kan inte du prata med henne istället, jag förstår mig inte på kvinnor	safe		

user	kan du i alla fall ba gå bort dit och hälsa snällt på henne	repetition		
cloddy	kan inte du prata med henne istället	safe		
user	okej	meta		
cloddy	jag vågar inte	bad		
user	kan jag få chansen till det också	meta		
karin	vem är du	nice	karin	karin_social (Han2_fp06)
user	johanna	social		
user	du då	social		
karin	jag heter karin	nice		
cloddy	hon brukar vara lite sur	bad		
karin	hur mår du	nice	karin	
user	bra hur mår du	social		
karin	det är väl sådär...jag är föräldralös och har knappt några vänner	nice		
user	vill du bli min vän	social		
karin	ja	nice		
user	kan du vara snäll och dra ner bron åt oss	domain	user	Bridge negotiation (Han2_fp06)
karin	om jag ska fälla ner bron vill jag ha nåt i gengäld	domain		
cloddy	hon är svag för mutor	meta		
karin	jag tyckte jag såg något glittrande falla från himlen samtidigt som den där dynghögen kom rasande	domain		
user	du kan få en rubin .. jag kommer tillbaka med något fint åt dig	domain		
cloddy	tjejer gillar väl sånt som glittrar	domain		
karin	varför släpar du med dig den där dummerjöns förresten	bad	karin	
user	han vägrar gå härifrån	social		
cloddy	<mmm>	meta		
user	lägg ner svärdet	domain	user	
cloddy	*[cloddy drops the sword]*			
user	kan du vara snäll och ta upp smaragden	repetition		
cloddy	då gör jag väl det *[walks over to the emerald and picks it up]*	safe		Bridge negotiation (Han2_fp10)
user	gå tillbaka till karin	domain	user	
cloddy	jag ska gå bort till , bron *[cloddy walks to karin]*	safe		
cloddy	kan inte du prata med henne istället ...jag är lite rädd för det tecknade könet	nice		
karin	vad har du att erbjuda	domain	karin	
user	en diamant	domain		
karin	nu börjar det likna nåt	nice		
cloddy	tjejer gillar väl sånt som glittrar	meta		
karin	det kanske kan duga	nice		
user	kan du öppna bron	domain	user	
karin	okej *[karin lowers the drawbridge]*	meta		
cloddy	hurra	nice		
user	gå över bron	domain	user	

cloddy	[cloddy walks over the bridge]			
user	sen ger du diamanten för att hon har varit snäll och fällt ner bron	domain	user	give_toll (Han2_fp02)
cloddy	ska vi ge henne grejerna eller låta bli	domain		
user	alltså jag vet inte	meta		
cloddy	ska vi ge henne grejerna eller låta bli	domain		
user	ge den till henne även fast hon är dum	repetition		
cloddy	okej	accept		
karin	ni lurade mig det här ska jag minsann komma ihåg	bad		
user	nej men ge henne diamanten fort	repetition		
cloddy	okej	accept		
cloddy	hon är ju bara en liten fjutt	bad		
user	jaa, men ändå hon släppte ju ner bron	meta		
karin	om ni bara hade bett snällt hade jag släppt över er ändå	bad		
user	men du sa ju att du behövde diamanter för att släppa ner bron	meta		
cloddy	varför säger du så	meta		
user	ja men hon sa ju att hon skulle fälla ner bron ifall vi bedde snällt, hon sa ju det	meta		
karin	skiter väl jag i	bad		
cloddy	vad är det du inte förstår	meta		
user	därför du sa att om*	meta		
karin	jag kan väl inte ha svar på allt	bad		
user	näe	meta		
cloddy	hon är så snäsig	bad		
user	vill du ha diamanten så får du be om ursäkt	domain		

Appendix 1b

Acomplete dialogue example constructed of phases from different users.

speaker	verbal	utterance type	sub-task inititalizer	phase (user ID)
user	hey	social	user	cloddy social (Eke_fp02)
cloddy	hey you	nice		
user	whats your name	social	user	
cloddy	my name is Cloddy Hans	nice		
user	my name is NN	social		
cloddy	how old are you	nice	cloddy	
user	eleven	social		
user	what is that	domain	user	machine plot description (Eke_fp26)
cloddy	<introduce_machine_plot_goal>	domain		
user	okey...is it alright to go on then	meta		
cloddy	you can talk with me and ask me to do thing	domain		
user	take that thing there on the shelf	domain	user	first object to machine (Eke_fp19)
user	hello	meta		
cloddy	hello yourself	meta		
user	go over to the fairy-tale shelf	domain	user	
cloddy	you want me too eeh , walk over to the shelf	safe		
user	yes…do that	yes		
cloddy	alright I'll do that then *[cloddy walks over to the shelf]*	meta		
user	take something from the shelf	domain		
user	ask thumbelina to do something	domain	user	out of phase (Han2_fp11)
cloddy	no I don't want to	bad		
user	are you a cowars...coward ha ha ha	test		
cloddy	what did you say about me	bad		
user	coward	test		
user	take something from the shelf	repetition		first object to machine (Eke_fp19)
cloddy	which one do you want me to take	safe		
user	that one [points at the sac] ..a sack isn't it?	domain	user	
cloddy	do you want me to take the sack	safe		
user	yes	yes		
cloddy	then I'll do that *[cloddy picks up sack]*	meta		
cloddy	what do you want me to do	safe		
user	take the book from the shelf as well	domain		
cloddy	I can only carry one thing at the time	problem		
user	but place the sack there in the fairy-tale machine *[user clicks at useful slot]*	domain	user	
cloddy	is it the sack you want me to place ehh there	safe		

user	yes	yes		
user	put the bag in the fairy-tale machine	repetition		
cloddy	do you want me to place the sack in the useful slot	safe		
user	yes	yes		
cloddy	then I'll do that *[cloddy puts sack in useful slot]*	meta		
user	could you pull the lever	domain	user	pull lever (Han2_fp12)
cloddy	I don't want to	bad		
user	why not	meta		
cloddy	if you pull the lever a fairy-tale is created	domain		
user	why don't you want that to happen	meta		
cloddy	mmm I actually not supposed to touch the fairy-tale machine	domain		
user	but I want you to	meta		
cloddy	mmm well I'm alone in the fairy-tale lab today...	domain		
user	okey.	meta		
cloddy	yees mayb we can have a look then	meta		
user	mmm I think you should	yes		
cloddy	do you want me to pull the lever	safe		
user	mmm	yes		
cloddy	I'm going to eeh , pull the lever *[cloddy pulls lever falls down into the world}*	safe		
user	where are we?	repetition		world intro (Han1_fp16)
cloddy	now we are in the fairy-tale world	domain		
cloddy	what should we do?	safe		
user	I don't know	meta		
cloddy	me neither	meta		
user	please walk over and pick up the sword	domain	user	explore world (Han2_fp06)
cloddy	do you want me to pick up the sword?	safe		
user	yes	yes		
cloddy	good idea *[cloddy walks over to the sword]*	meta		
cloddy	I'm going to pick up the sword *[cloddy picks up the sword]*	safe		
user	thanx use the sword on the tree	test		
cloddy	I don't want to	bad		
user	walk over to the girl over there	domain	user	
cloddy	Do you want me to eeh , walk over to the bridge	safe		
user	yeah	yes		
cloddy	yes why not *[cloddy walks over to the bridge]*	meta		
user	thanx …BUURP…ha ha ha	social		
user	could you please talk with the girl over there	repetition	user	karin intro (Han2_fp11)
cloddy	her name is Karin	domain		
user	could you please talk with Karin	repetition		
user	ask Karin if she needs help	repetition		
cloddy	can't you talk with her instead, I don't understand women	safe		

user	couldn't you just walk over and say hi to her nicely	repetition		
cloddy	why don't you talk with her instead	safe		
user	okey	meta		
cloddy	I don't dear to	bad		
user	could you give me the chance to do it then	meta		
karin	who are you	nice	karin	karin_social (Han2_fp06)
user	johanna	social		
user	and you	social		
karin	my name is Karin	nice		
cloddy	she usually is a bit grumpy	bad		
karin	how are you	nice	karin	
user	fine how are you	social		
karin	well so so ... I'm an orphean and I hardly have any friends	nice		
user	would you like to become my friend	social		
karin	yes!	nice		
user	could you please lower the drawbridge for us	domain	user	Bridge negotiation (Han2_fp06)
karin	if I am to lower the drawbridge i want something in return	domain		
cloddy	she is easily bribed	meta		
karin	I thought I saw something shineng falling down at the same time as the smuck	domain		
user	you can have a ruby .. I'll return with something nice for you	domain		
cloddy	girl like shiny things!	domain		
karin	Why do you keep dragging along that Cloddy Hans figure, by the way	bad	karin	
user	he refuse to leave	social		
cloddy	mmm (with negative prosody)	meta		
user	drop the sword	domain	user	
cloddy	*[cloddy drops the sword]*			
user	could yo please pick up the emerald	repetition		
cloddy	ok I'll do that *[walks over to the emerald and picks it up]*	meta		Bridge negotiation (Han2_fp10)
user	go back to Karin	domain	user	
cloddy	I'll walk over to the bridge *[cloddy walks to karin]*	safe		
cloddy	couldn't you talk with her...I'm a bit afrain of the animated sex (word joke)	nice		
karin	what do you have for me?	domain	karin	
user	a diamond	domain		
karin	now you're talking	nice		
cloddy	girl like shiny things!	meta		
karin	it might be sufficient	nice		
user	could you open the bridge	domain	user	
karin	okey *[karin lowers the drawbridge]*	meta		
cloddy	hurray	nice		
user	walk over the bridge	domain	user	

cloddy	[cloddy walks over the bridge]			
user	then give the diamond to her since she was nice and lowered the bridge	domain	user	give_toll (Han2_fp02)
cloddy	do you think we should give her our stuff or not?	domain		
user	well I don’t know	meta		
cloddy	should we give her our stuff or not?	domain		
user	give it to her...even though she is naughty	repetition		
cloddy	okey	accept		
karin	You fooled me! I will remember this!	bad		
user	no but...give her the diamon fast	repetition		
cloddy	okey	accept		
cloddy	she is only a little fjutt	bad		
user	yeah, but anyway she did lower the bridge	meta		
karin	if you had only asked me nicely I would have let you over anyway	bad		
user	but you said that you needed diamonds to lower the bridge	meta		
cloddy	why do you say that?	meta		
user	yes but she said that she would lower the bridge if we aske nicely, sha said that	meta		
karin	I don’t care	bad		
cloddy	what don’t you understand?	meta		
user	because you said the if*	meta		
karin	jI can’t have answers to anything	bad		
user	no	meta		
cloddy	she is so grumpy	bad		
user	if you want the diamond you’ll have to appologize to me	domain		

A complete dialogue example constructed of phases from different users.

speaker	verbal	utterance type	sub-task inititalizer	phase (user ID)
user	hey	social	user	cloddy social (Eke_fp02)
cloddy	hey you	nice		
user	whats your name	social	user	
cloddy	my name is Cloddy Hans	nice		
user	my name is NN	social		
cloddy	how old are you	nice	cloddy	
user	eleven	social		
user	what is that	domain	user	machine plot description (Eke_fp26)
cloddy	<introduce_machine_plot_goal>	domain		
user	okey...is it alright to go on then	meta		

cloddy	you can talk with me and ask me to do thing	domain		
user	take that thing there on the shelf	domain	user	first object to machine (Eke_fp19)
user	hello	meta		
cloddy	hello yourself	meta		
user	go over to the fairy-tale shelf	domain	user	
cloddy	you want me too eeh , walk over to the shelf	safe		
user	yes…do that	yes		
cloddy	alright I'll do that then *[cloddy walks over to the shelf]*	meta		
user	take something from the shelf	domain		
user	ask thumbelina to do something	domain	user	out of phase (Han2_fp11)
cloddy	no I don't want to	bad		
user	are you a cowars...coward ha ha ha	test		
cloddy	what did you say about me	bad		
user	coward	test		
user	take something from the shelf	repetition		first object to machine (Eke_fp19)
cloddy	which one do you want me to take	safe		
user	that one [points at the sac] ..a sack isn't it?	domain	user	
cloddy	do you want me to take the sack	safe		
user	yes	yes		
cloddy	then I'll do that *[cloddy picks up sack]*	meta		
cloddy	what do you want me to do	safe		
user	take the book from the shelf as well	domain		
cloddy	I can only carry one thing at the time	problem		
user	but place the sack there in the fairy-tale machine *[user clicks at useful slot]*	domain	user	
cloddy	is it the sack you want me to place ehh there	safe		
user	yes	yes		
user	put the bag in the fairy-tale machine	repetition		
cloddy	du you want me to place the sack in the useful slot	safe		
user	yes	yes		
cloddy	then I'll do that *[cloddy puts sack in useful slot]*	meta		
user	could you pull the lever	domain	user	pull lever (Han2_fp12)
cloddy	I don't want to	bad		
user	why not	meta		
cloddy	if you pull the lever a fairy-tale is created	domain		
user	why don't you want that to happen	meta		
cloddy	mmm I actually not supposed to touche the fairy-tale machine	domain		
user	but I want you to	meta		
cloddy	mmm well I'm alone in the fairy-tale lab today...	domain		
user	okey.	meta		
cloddy	yees mayb we can have a look then	meta		
user	mmm I think you should	yes		

cloddy	do you want me to pull the lever	safe		
user	mmm	yes		
cloddy	I'm going to eeh , pull the lever *[cloddy pulls lever falls down into the world}*	safe		
user	where are we?	repetition		world intro (Han1_fp16)
cloddy	now we are in the fairy-tale world	domain		
cloddy	what should we do?	safe		
user	I don't know	meta		
cloddy	me neither	meta		
user	please walk over and pick up the sword	domain	user	explore world (Han2_fp06)
cloddy	do you want me to pick up the sword?	safe		
user	yes	yes		
cloddy	good idea *[cloddy walks over to the sword]*	meta		
cloddy	I'm going to pick up the sword *[cloddy picks up the sword]*	safe		
user	thanx use the sword on thae tree	test		
cloddy	I don't want to	bad		
user	walk over to the girl over ther	domain	user	
cloddy	vill du att eeh , jag ska gå bort till , bron	safe		
user	yeah	yes		
cloddy	yes why not *[cloddy walks over to the bridge]*	meta		
user	thanx …BUURP…ha ha ha	social		
user	could you please talk with the girl over there	repetition	user	karin intro (Han2_fp11)
cloddy	her name is Karin	domain		
user	could you please talk with Karin	repetition		
user	ask Karin if she needs help	repetition		
cloddy	can't you talk with her instead, I don't understand women	safe		
user	couldn't you just walk over and say hi to her nicely	repetition		
cloddy	why don't you talk with her instead	safe		
user	okey	meta		
cloddy	I don't dear to	bad		
user	could you give me the chance to do it then	meta		
karin	who are you	nice	karin	karin_social (Han2_fp06)
user	johanna	social		
user	and you	social		
karin	my name is Karin	nice		
cloddy	she usually is a bit grumpy	bad		
karin	how are you	nice	karin	
user	fine how are you	social		
karin	well so so ... I'm an orphean and I hardly have any friends	nice		
user	would you like to become my friend	social		
karin	yes!	nice		

user	could you please lower the drawbridge for us	domain	user	Bridge negotiation (Han2_fp06)
karin	if I am to lower the drawbridge i want something in return	domain		
cloddy	she is easily bribed	meta		
karin	I thought I saw something shineng falling down at the same time as the smuck	domain		
user	you can have a ruby .. I'll return with something nice for you	domain		
cloddy	girl like shiny things!	domain		
karin	Why do you keep dragging along that Cloddy Hans figure, by the way	bad	karin	
user	he refuse to leave	social		
cloddy	mmm (with negative prosody)	meta		
user	drop the sword	domain	user	
cloddy	*[cloddy drops the sword]*			
user	could yo please pick up the emerald	repetition		
cloddy	ok I'll do that *[walks over to the emerald and picks it up]*	meta		Bridge negotiation (Han2_fp10)
user	go back to Karin	domain	user	
cloddy	I'll walk over to the bridge *[cloddy walks to karin]*	safe		
cloddy	couldn't you talk with her...I'm a bit afrain of the animated sex (word joke)	nice		
karin	what do you have for me?	domain	karin	
user	a diamond	domain		
karin	now you're talking	nice		
cloddy	girl like shiny things!	meta		
karin	it might be sufficient	nice		
user	could you open the bridge	domain	user	
karin	okey *[karin lowers the drawbridge]*	meta		
cloddy	hurray	nice		
user	walk over the bridge	domain	user	
cloddy	[cloddy walks over the bridge]			
user	then give the diamond to her since she was nice and lowered the bridge	domain	user	give_toll (Han2_fp02)
cloddy	do you think we should give her our stuff or not?	domain		
user	well I don't know	meta		
cloddy	should we give her our stuff or not?	domain		
user	give it to her...even though she is naughty	repetition		
cloddy	okey	accept		
karin	You fooled me! I will remember this!	bad		
user	no but...give her the diamon fast	repetition		
cloddy	okey	accept		
cloddy	she is only a little fjutt	bad		
user	yeah, but anyway she did lower the bridge	meta		
karin	if you had only asked me nicely I would have let you over anyway	bad		
user	but you said that you needed diamonds to lower the bridge	meta		
cloddy	why do you say that?	meta		
user	yes but she said that she would lower the bridge if we aske nicely, sha said that	meta		

karin	I don’t care	bad		
cloddy	what don’t you understand?	meta		
user	because you said the if*	meta		
karin	jI can’t have answers to anything	bad		
user	no	meta		
cloddy	she is so grumpy	bad		
user	if you want the diamond you’ll have to appologize to me	domain		