NICE project (IST-2001-35293)

Natural Interactive Communication for Edutainment

NICE Deliverable D7.1

Evaluation criteria and evaluation plan

March 2003

Authors

Laila Dybkjær[1], Niels Ole Bernsen[1], Reinhard Blasig[2], Stéphanie Buisine[4], Morgan Fredriksson[5], Joakim Gustafson[3], Jean-Claude Martin[4], Mats Wirén[3]

1: NISLab, Odense, Denmark, 2: Scansoft Aachen GmbH, Germany, 3: Telia Research AB, Farsta, Sweden,

4: LIMSI-CNRS, Orsay, France, 5: Liquid Media, Stockholm, Sweden

Project ref. no.	IST-2001-35293
Project acronym	NICE
Deliverable status	Internal
Contractual date of delivery	28 February 2003
Actual date of delivery	22 March 2003
Deliverable number	D7.1
Deliverable title	Evaluation criteria and evaluation plan
Nature	Report
Status & version	Final
Number of pages	29
WP contributing to the deliverable	WP7
WP / Task responsible	NISLab
Editor	Laila Dybkjær
Author(s)	Laila Dybkjær, Niels Ole Bernsen, Reinhard Blasig, Johan Boye, Stéphanie Buisine, Morgan Fredriksson, Joakim Gustafson, Anders Lindström, Jean-Claude Martin, Mats Wiren
EC Project Officer	Mats Ljungqvist
Keywords	Evaluation, criteria, methods, natural interactivity, multimodal interaction.
Abstract (for dissemination)	This deliverable from the IST/HLT NICE (Natural Interactive Communication for Edutainment) project specifies a set of evaluation criteria and outlines an evaluation plan for the two prototypes to be developed in the NICE project.

Table of Contents

1 Introduction

Evaluation is an important part of the software life cycle and tightly interwoven with development. Its major function is to provide iterative feedback on the quality of each component as well as on the entire system throughout the development process. Despite its importance evaluation of multimodal and natural interactive systems is today as much of an art and a craft as it is an exact science with established standards and procedures of good engineering practice. This means that establishing an appropriate set of evaluation criteria and choosing the right evaluation methods for the NICE prototypes is not a straightforward task but presents a research challenge of its own, as will also appear from the discussions in the present report.

The following chapters first provide a brief overview of the background on which we view the NICE evaluation task. We then present a comprehensive set of evaluation criteria and plans to be used – possibly in a revised version - in the evaluation of the two NICE software prototypes which will be available by the end of month 20 and by the end of month 32, respectively. Chapter 2 briefly describes evaluation in NICE-related areas and discusses which strategy to follow in NICE. Chapter 3 presents criteria for evaluation of the entire system as well as component-related criteria. Chapter 4 discusses evaluation of the first and second prototype, respectively. Chapter 5 presents plans for how to involve representative users, how to carry out evaluation, and how to analyse and use evaluation results and feedback from users. Suggestions for improvements based on evaluation results and feedback will be provided to, and discussed with, the developers. Relevant suggestions will be taken into account in the next module/system iteration.

2 Background

Before deciding on an evaluation plan for NICE and on which evaluation criteria to include, it is useful to begin by recalling what distinguishes NICE from related systems in neighbouring areas:

First, what primarily distinguishes NICE from other computer games is spoken, multimodal dialogue. Moreover, dialogue is not supposed to be just an "add-on" but the primary means of progression in the game. The rationale for this is the great potential for more natural interaction we see in making available methods from multimodal dialogue systems as a means of controlling gameplay.

Secondly, what distinguishes NICE from typical spoken dialogue systems is the attempt to move away from task-oriented dialogue. Thus, the interaction with HCA is domain-oriented, which means that it concerns areas associated with his life and works, but without a clear goal-orientation and without other demands than it being entertaining and educational to the user. Furthermore, the interaction in the fairy-tale world (where the game proper takes place) can be seen as joint problem-solving between the user and the characters. Thus, although the user is initially presented with an overall problem, the purpose of this interaction is not to solve a "task", but to progress through a story by gradually overcoming various obstacles in which "socialising" with the characters of the game is an important ingredient too.

Thirdly, NICE shares many goals, such as believability of characters and natural expressions, with those of affective computing [Picard 1997, Höök 2002]. However, in NICE "affecting" the user is not a goal in itself but rather a means for achieving something else, namely, the timely unfolding of an engaging and challenging narrative which makes the user want to finish the game.

The first two points suggest that we should continue to look at evaluation methods for multimodal spoken language dialogue systems (SLDS), whereas the third point indicates that we also need to look at how embodied conversational agents are evaluated. Ideally, however, we need to go one step further and look at how well the game actually achieves its purpose of being engaging, fun, and creating a relationship with the user, and how much of this can be attributed to the conversational agents.

For SLDSs and their components, including speech recognition, natural language understanding and generation, dialogue management, speech synthesis, system integration, and human factors, there has been extensive work on evaluation documented in, e.g., the DISC Best Practice Guide (www.disc2.dk) which provides useful information on technical and usability evaluation, and in the US DARPA Communicator project which has used the PARADISE (Paradigm for Dialogue System Evaluation) framework [Walker et al. 1997] for usability evaluation.

For some areas of SLDSs there are standards or best practices, such as for speech recogniser evaluation, while for other areas, e.g. dialogue management, there is little in terms of standardisation. We will evaluate components according to existing standards and best practices when these exist and are relevant for a given component. However, this report will not list all such standard or best practice criteria. Rather it will focus on those criteria which we consider essential to measure the robustness, functionality, and value for the user of the NICE system. Due to the nature of NICE, several of these criteria may very well have to be new ones compared to what has so far been used in evaluation of multimodal SLDSs.

A major challenge in NICE is that we are moving away from the ordinary task-oriented (multimodal) SLDS and into domain-oriented dialogue. This is still to a large extent unexplored territory both as regards development and evaluation. We expect that the collected

data may bring us important information about new aspects of natural interactivity, not least if we manage to establish a set of useful evaluation criteria for this purpose. Several ongoing research projects have as part of their agenda to look into evaluation methods for various aspects of natural interactive and multimodal dialogue systems. For instance: SMADA, Speech Driven Multimodal Automatic Directory Assistance, 2000-2002, http://smada.research.kpn.com/MainPage/, looked at usability issues for small terminals for mobile internet access. INSPIRE, Infotainment management with speech interaction via remote-microphones and telephone interfaces, 2002-2004, http://www.inspire-project.org/, looks at usability and acceptability evaluation. MIAMM, Multidimensional Information Access using Multiple Modalities, 2001-2004, http://www.loria.fr/projets/MIAMM, looks at evaluation methods and protocols for multimodal interaction. And SmartKom, 1999-2003, http://smartkom.dfki.de/, addresses usability evaluation of multimodal systems from a general point of view. The evaluation work in SmartKom is inspired by PARADISE which perhaps is the most widely used framework for usability evaluation of SLDSs. This framework tries to model user satisfaction as a function of task success and various dialogue cost metrics. The PARADISE model is not unproblematic in itself [Dybkjær et al. 2003] and, as pointed out by [Beringer et al. 2002], the model will have to be extended for multimodal systems evaluation. Based on PARADISE, [Beringer et al. 2002] propose an extended evaluation framework called PROMISE (Procedure for Multimodal Interactive System Evaluation), for multimodal dialogue systems evaluation. The details of the model, however, seem not yet to have been fleshed out so this must be considered relatively early work. We will keep an eye on these projects to see if they deliver something which may be useful to NICE.

The evaluation of ECAs (embodied conversational agents) is another aspect to look at in NICE. ECA systems evaluation has been addressed by several empirical studies, see [Dehn and van Mulken 2000] for a review. These studies consist of experimental comparisons between different systems with respect to, e.g., presence versus absence of an agent, amount of non verbal behaviour displayed, amount of embodiment, style of rendering, etc. The main criteria used in experimental evaluation concern the influence of the quality of animated agents

- on the user's subjective experience of the system, i.e. perceived intelligence, perceived believability, likeability, engagingness/entertainment value, comfortability, smoothness of interaction, etc.
- on the user's behaviour while interacting with the system, including attention (eye contact, reaction time...), flow of communication (conversation index, e.g. number of topic shifts and successful answers per time units, number of repetitions and hesitations, communicational overlaps), etc.
- on the outcome of the interaction as indicated by performance data on., e.g., problem solving, learning, memory performance.

Several of these criteria may also prove useful for evaluation in NICE.

Given the emerging nature of the field, it is, as already hinted at, expected that some of the evaluation criteria developed and used in the NICE project will constitute improvements on the state-of-the-art in multimodal and natural interactive systems evaluation, for instance wrt. the notion of transaction success. The same applies to some of the results of evaluation, for instance regarding modality appropriateness, children's interaction with multimodal edutainment systems, edutainment value, or animated character usefulness. For this reason, the methods and criteria mentioned in the following chapters of this report should be seen as a first proposal for the evaluation of the NICE prototypes and not as a completely fixed strategy. It is likely that the presented set of criteria and plans will be subject to iterative

improvement based on the experience gained during the project with development and experimental evaluation, as well as due to progress in the field more generally.

3 Evaluation criteria

As indicated in the previous chapter, establishing a set of evaluation criteria for NICE is in part a research challenge in itself due to the innovative nature of the system. We shall use existing standards and best practices when possible and we shall keep an eye on results from projects exploring evaluation criteria for multimodal and natural interactive systems. However, to the extent that we cannot rely on existing criteria we have to set up our own and explore their usability for evaluating the NICE prototypes.

This section discusses which evaluation criteria we propose to use in the evaluation of the first and the second prototype, respectively. Clearly, the second prototype must demonstrate, among other things, increased robustness, functionality and linguistic capabilities as means to making it more engaging compared to the first prototype.

Each prototype must be evaluated both at component level and at the overall system level. Moreover, we must make sure that the H.C. Andersen (HCA) part as well as the fairy tale world part are evaluated, which may to some extent involve different evaluation criteria for each of the two parts.

We distinguish between technical evaluation and usability evaluation. Technical evaluation concerns the entire system as well as each of its components. It is usually done by developers through objective evaluation, i.e. through quantitative and/or qualitative evaluation. Quantitative evaluation consists in counting something and producing an independently meaningful number, percentage etc. Qualitative evaluation consists in estimating or judging some property by reference to expert standards and rules. Usability evaluation of a system is usually done by developers and users. Developers may to some extent draw on objective evaluation metrics but a substantial part of usability evaluation is done via subjective evaluation, i.e. by judging some property of a system or, less frequently, component by reference to users' opinions. All these kinds of criteria may be used for various purposes, such as for evaluation of the system's quality, its conformance with specifications, for comparing the system with other systems, or for measuring progress during system development.

3.1 Technical system evaluation

Technical evaluation focuses on robustness and functionality from a technical point of view. This section concentrates on technical evaluation at the overall system level. The primary purpose of the technical evaluation criteria is (i) to test if the system has the specified overall technical functionality, and (ii) to test if the system has the technical robustness required for users to interact with it sufficiently smoothly for usability evaluation to make sense. In all cases, objective measures can be applied. For many of the criteria listed below, a quantitative evaluation method can be used while in other cases qualitative evaluation will be needed. After each criterion a comment is added in parentheses which indicates whether evaluation is quantitative or qualitative and what we understand to be measured by the criterion.

- technical robustness
  quantitative; how often does the system crash (the system is hanging); how often does it produce a bug which prevents continued interaction (e.g. a loop)
- handling of out-of-domain input
  qualitative; to which extent does the system react reasonably to out-of-domain input
- time performance
  quantitative; how long does it usually take to get a reaction from the system during interaction

- barge-in
  quantitative; is barge-in implemented
- number of characters
  quantitative; how many characters are available in the fairy tale world
- number of emotions which can be expressed by characters
  quantitative/qualitative; how many different emotions can be conveyed
- number of input/output modalities
  quantitative, how many input modalities and how many output modalities does the system allow
- synchronisation of output
  qualitative; is output properly synchronised
- number of domains
  quantitative; how many domain can HCA talk about (his life, his fairy tales, etc.)
- number of different plots/scenes available
  quantitative; how many different plots/scenes can the user choose among

3.2 Usability evaluation of the system

To evaluate the usability of the system, we have established a set of evaluation criteria divided into two subsets. The first group includes the criteria which we consider basic to usability. Many of them are common for the HCA part and the fairy tale world part while some are specific to one of these parts only. Those which are specific are marked in parentheses either by HCA or by FT (for fairy tale world). If one of the criteria produces a very negative evaluation it may mean that the responsible module(s) must be improved before further evaluation is worthwhile. For example, if speech recognition adequacy is very bad this means that, basically, the user is not able to communicate with the system until recognition has been improved. From the technical component evaluation we will have measures of, e.g., speech recogniser performance, gesture recogniser performance, and parser performance. These are objective metrics which may be compared to perceived subjective recognition and understanding adequacy.

The second group includes the criteria which we consider essential to the evaluation of the NICE prototypes but many of which are new and may need subsequent re-definition because they represent research challenges. All criteria on this list are annotated with HCA, FT or both to show which part(s) of the system a particular criterion will be used to evaluate.

Most parameters below must be evaluated using a subjective method, such as questionnaire or interview. Again, the comment in parentheses after each criterion indicates whether evaluation is subjective or is objective in terms of being either quantitative or qualitative.

3.2.1 Basic usability criteria

- speech understanding adequacy
  subjective; how well does the system understand speech input
- gesture understanding adequacy
  subjective; how well does the system understand gesture input
- output voice quality
  subjective; how intelligible and natural is the system output voice
- output phrasing adequacy
  subjective; how adequate are the system's output formulations

- animation quality
  subjective; how natural is the animated output
- quality of graphics
  subjective; how good is the graphics
- ease of use of input devices
  subjective; how easy are the input devices to use, such as a joystick
- frequency of interaction problems
  HCA; quantitative; how often does an interaction problem occur (e.g. the user is not understood or is misunderstood)
- sufficiency of domain coverage
  HCA; subjective; how well does the system cover the domains it announces to the user
- number of characters the user interacted with in the fairy tale world
  FT; quantitative; serves to check if some character is difficult to find or for other reasons not used
- number of objects the subject interacted with
  HCA, FT; quantitative; serves to check to which extent the possibilities in principle offered by the system are also used by users
- navigation in the environment (number of places visited, etc. in the fairy tale world)
  FT; quantitative; serves to check to which extent the environmental possibilities in principle offered by the system are also found and used by users
- number of topics addressed in the conversation
  HCA; quantitative; serves to check how well the implemented domains cover the topics addressed by users

3.2.2 Core usability criteria

- transaction success
  HCA; quantitative; how often is a transaction exchange between the user and the system successful
- naturalness of user speech and gesture (including modality appropriateness)
  HCA, FT; subjective; how natural is it to communicate via the available modalities
- output behaviour naturalness
  HCA, FT; subjective; character believability, speech, graphics, coordination of speech and graphics, display of emotions, dialogue initiative, dialogue flow, synchronisation of verbal and non-verbal behaviour, non-communicative function, etc.
- sufficiency of the system's reasoning capabilities
  HCA; subjective; how good is the system at reasoning about user input
- ease of use of the game
  HCA, FT; subjective; how easy is it for the user to find out what he can do and how to interact in the HCA part and in the fairy tale world part, respectively
- error handling adequacy (such as detection of errors, how to handle them)
  HCA, FT; subjective; how good is the system at detecting errors and how well does it handle them
- scope of user modelling, i.e. the system’s request for, and use of, knowledge about its users
  HCA; subjective; to which extent does the system exploit what it learns about the user
- entertainment value
  HCA, FT; subjective; this measure includes issues such as quality of the game,

originality of the game, interest taken in the game, feeling like playing again, time spent playing the game, user initiative in game

- educational value
  HCA, FT; subjective; to which extent did the user learn something from interacting with the system
- user satisfaction
  HCA, FT; subjective; how satisfied is the user with the system

Usability evaluation will have high priority in NICE although the focus will be slightly different for the HCA part and the fairy tale world part. Section 5.4 on evaluation methods and approaches discusses the approaches and differences in more detail.

3.3 Technical component evaluation

While, in most cases, usability evaluation requires data from a full (simulated or implemented) system and therefore is mostly addressed at system level, technical evaluation at component level makes perfect sense and is crucial. Errors and weaknesses in individual components will negatively influence system integration and contribute to sub-optimal performance of the entire system. Exhaustive technical evaluation is very time consuming. There is no time in NICE for such a thorough technical evaluation. To make sure that proper evaluation is done in any case although it may not be exhaustive, we list below for each component those criteria that are considered most important and that we plan to evaluate.

Typically, individual component evaluation will be carried out with each significantly new component version, enabling progress evaluation for that component. Since we are collecting large amounts of data, test data provision is not likely to be a problem.

In the lists below those criteria which have standard definitions or appear self-explanatory are not commented further while those which may not be self-evident are explained in parentheses.

3.3.1 Speech recogniser

- Word error rate for English and for Swedish
- Vocabulary coverage for English and for Swedish
- Perplexity of English language model and of Swedish language model
- Time performance

The perplexity measure will quantify how well our collected data covers the usual conversations (as represented by some evaluation corpus).

3.3.2 Gesture recogniser

- Recognition accuracy regarding gesture type
- Number of recognition failures
- Number of interpretation errors

3.3.3 Natural language understanding

- Lexical coverage, English NLU and Swedish NLU for fairy tale world
- Lexical coverage, English NLU and Swedish NLU for HCA
- Parser error rate, English NLU and Swedish NLU for fairy tale world
- Parser error rate, English NLU and Swedish NLU for HCA
- Topic spotter error rate, English NLU and Swedish NLU for HCA
- Anaphora resolution error rate, English NLU and Swedish NLU for HCA

3.3.4 Gesture interpretation

- Selection of referenced objects error rate

3.3.5 Input fusion

- Robustness to temporal distortion between input modalities
- Fusion error rate
- Cases in which events have not been merged and should have
- Cases in which events have been merged and should not have
- Recognised modality combination error rate

3.3.6 Character modules

- Meta-communication facilities (which kind(s) of meta-communication are implemented)
- Handling of initiative (is the system able to give and take initiative)
- Performance of conversational history (HCA; does the history support smooth and reasonable interaction)
- Handling of changes in emotion (HCA; can emotions change based on the contents of user input)

3.3.7 Response generation

- Coverage of action set (communicative and other movements)

3.3.8 Graphical rendering (animation)

- Synchronisation with speech output
- Naturalness of animation, possibly including sub-division into modalities (gaze, gesture, etc.)

3.3.9 Text-to-speech

- Speech quality, English and Swedish
- Intelligibility, English and Swedish
- Naturalness, English and Swedish (must be evaluated in the context of the chosen character)

Note, that TTS is not developed in the project so the evaluation criteria are primarily to choose the best TTS available for the purpose.

3.3.10 Non-speech sound (if any)

- Appropriateness in context of music/sound to set a mood

3.3.11 Integration

- Communication among modules (are messages sent and received as planned)
- Message dispatcher (does the message dispatcher communicate as needed with other modules)
- Processing time per module

A good management of speech turn-taking is clearly an issue in existing systems. For instance, the character should show non-verbal behaviour when she is going to talk to avoid that the user starts speaking; the character should display non verbal behaviour when she gives the floor to the user.

4 Evaluation of first and second prototypes

The plan is to develop two prototypes of which the first prototype will have fairly limited functionality but will demonstrate that modules work together and that it is possible to carry out a conversation with HCA and perform a game in the fairy tale world. Only a (relatively large) subset of the evaluation criteria from Chapter 3 will be applicable to the first prototype whereas all of them should be applied to the second prototype.

4.1 First prototype

The first prototype should be evaluated according to a relevant sub-set of the evaluation criteria in Chapter 3. The list below includes the most important points on which the first prototype differs from the second one apart from generally lower performance at component level as well as at system level. The relevant sub-set of evaluation criteria must take the points below into account and thus not measure, e.g., user modelling or English natural language understanding for the fairy tale world.

- One character will be available in the fairy tale world.
- One game will be available.
- HCA has limited domain coverage within each of the five domains (life, fairy tales, presence, gate keeper, and user with at least three topics in each).
- English HCA and Swedish fairy tale world. The two parts will not be connected in the first prototype but can be activated and used independently.
- No user model will be available.

4.2 Second prototype

The second prototype should have the full functionality outlined in the contract and specified in deliverables. There will be an English full version and a Swedish full version. The following list includes the major differences between the first and the second prototypes.

- At least five characters available in the fairy tale world.
- At least four plots/scenes available in the fairy tale world, cf. Appendix 1.
- Substantially extended coverage for each domain in the HCA part compared to first prototype.
- User model included in HCA part.
- HCA and the fairy tale world will be connected.
- Full system in two languages (English and Swedish).
- Close to real-time performance.
- General improvements across components as well as at system level.

5 Evaluation plan

This chapter presents a detailed plan for how to test the two NICE software prototypes. Chapter 3 discussed evaluation criteria, i.e. basically *what* to test. The plan in this chapter includes other issues, such as which users to involve, where to test (environments), when to test what, and how to test (methods).

There are two major evaluation checkpoints in the NICE project, i.e. prototype 1 and prototype 2. Before and after prototype 1, there may be several smaller iterations which are also being evaluated to continuously keep track of progress and performance.

Each of the two prototypes must be thoroughly evaluated against the (possibly later revised) criteria established in Chapter 3. To be considered successful, each prototype must meet the criteria following the specifications below.

Results from the major evaluations as well as from evaluation of any intermediate system or component version will feed into the continued development of the first/second prototype. The final evaluation report will include recommendations for improvements (needed as well as desirable) to be made during exploitation/product development.

5.1 Users

The primary target group for the NICE system are children and adolescents between 9 and 18 years of age. People above 18 years should also be able to use the system but emphasis will be on user satisfaction of the primary target group. An advantage of NICE compared to task-oriented multimodal dialogue systems, such as information systems, is that users never have a particular task to carry out no matter if they are test subjects or real users. For information systems, there is often a considerable difference in the behaviour of subjects and real users because, to the test subjects, the tasks they are asked to carry out are artificial ones in the test context even if the subjects are representative of the target user group. For instance, subjects don't have a need to know about, e.g., a particular train connection and often do not care much about the information they are offered by the system, which means that they are often more positive in their evaluation of the system than are the eventual real users. For a game system such as NICE, the key issue is that the subject or real user must be interested in interacting with a computer game. This interest may be the same no matter if the interaction is done in test environments with subjects or with a final system and real users.

When doing evaluation of the two prototypes as well as in-between system versions, attention should be paid to the following parameters as regards users:

- Numbers of test persons: each prototype should be evaluated by at least 12 test users.
- Age: at least 8 users should belong to the primary target group. At least 2 users should be older than the target group.
- Both genders should be represented approximately equally.
- Subjects' backgrounds: we should try to have users who attend different educations (for high school students), have different major interests, have different computer game literacy, etc.
- Language background: for the Swedish version, native Swedish users will act as test subjects. For the English version, non-native UK English speakers coming from Denmark, Germany, the US and Japan will be used as well as native UK English users.

For the second prototype, we will include new users (both Swedish and UK English speakers) in addition to some of those who used the system already, to see how they receive the second prototype system which should be considerably more advanced than the first one.

5.2 Test environments

Users from the primary target group are considered very important for the ongoing evaluation of the prototypes throughout the project. Some ideas for actively seeking user contact and user involvement for continuous evaluation with target users in target environments include:

- putting out the first prototype at a telecom museum site and at the H.C. Andersen museum (young museum visitors are seen as a potential user group);
- putting out pre-versions of the second prototype at museum/exhibition sites which are highly appropriate for showing H.C. Andersen/NICE-related technology;
- deploying and evaluating system prototypes to existing customer bases of partners; and
- deploying and evaluating system prototypes to computer clubs involving young people.

The primary target environment is museums and exhibitions. However, if time allows, we will also evaluate a NICE system version intended for home use. It should be borne in mind that museums and the home are two quite different environments and may impose different requirements on both hardware and software. For example, a larger screen may be appropriate for a museum than the one used at home. At the museum, there may be several other visitors wanting to watch what is going on while a user interacts. The noise conditions will also be different in the museum compared to a home environment, which implies different demands on the recogniser software and/or setup.

5.3 Evaluation performance

The two major checkpoints will be the two prototypes. Chapter 4 provides an overview of our plans for each of the two prototypes as regards functionality and performance, and thus also points to which evaluation criteria are relevant for the first prototype. All evaluation criteria listed are expected to be relevant for the second prototype.

Since the NICE project is an innovative research project with many unknowns that have to be explored, it is difficult at this point to set up precise targets for how well the evaluation criteria must be met, even as regards most of the quantitative aspects of technical system and components evaluation. Obviously, the evaluation results from the first prototype will serve as a background on which to compare progress in performance of the second prototype. However, it seems clear that there are few, if any, component and system evaluation baselines to be found in the field for the kind of system we are building. So, in general, we rather seem to be in the position of developing a new kind of system whose evaluation results might come to serve as baselines for comparison with subsequent systems and components in the field. We will of course also keep an eye on what is going on in the field of multimodal and natural interactive systems evaluation. Relevant results from the field may be used in a comparison to NICE results.

5.4 Evaluation methods and approaches

There exists a wealth of system and component evaluation methods. Some of these are only usable at certain stages of development while others can be used throughout the development life cycle. We shall not make exact and exhaustive plans for which methods to use in NICE. It

is likely that, among others, walkthroughs and Wizard of Oz experiments will be used at the earlier stages of development whereas blackbox tests of components, controlled experiments with implemented system versions, and field tests of system versions will be used later on when software is available and sufficiently stable for these purposes. Questionnaires and interviews will be used to collect feedback from users. Component and system performance will be compared across versions.

The focus of evaluation and the approach to evaluation will be slightly different in the HCA part and the fairy tale world part. In the following we briefly describe the planned approaches for the two parts.

5.4.1 HCA evaluation focus and approach

The HCA part of NICE will have its main emphasis on the dialogue between the user and HCA although input gesture will also be allowed. This makes the HCA part somewhat different from the fairy tale world part which will have less advanced dialogue but more gestures and a main emphasis on the game and play.

Thus the main focus of HCA usability evaluation will be on the dialogue and HCA's behaviour, i.e. on how the conversation with the user unfolds in a natural and edutaining way, how HCA handles input which is not understood for some reason or another, or is misunderstood, how he expresses emotional reactions to input, how he exploits what he learns about the user, etc.

Wizard-of-Oz studies are being used to collect information on what users may want to talk with HCA about, how they do it, and how they react to his reactions. By using Wizard-of-Oz we can relatively easily change HCA so that he, e.g., understands more or understands less, or reacts more or less expressively.

We will use questionnaires and interviews as a main source to iteratively collect information about users' opinions throughout the project and we will analyse video and audio recordings from users' interaction with HCA. The iterative results from users' input and from the analyses of recordings will serve as a basis for implementation and for later improvements of the implementation.

In the analyses we will pay special attention to issues such as modality appropriateness, naturalness of interaction, and transaction success.

To the extent possible we will follow standards when annotating data. As regards annotation tools we will consider what is the best solution when we have the need. One possibility might be to use the NITE workbench (nite.nis.sdu.dk) which is under construction, provided that it is ready by the time we need a tool. Annotation schemes and tools for natural and multimodal interactive behaviour, including best practice recommendations, are presented and discussed in a series of reports edited by NISLab and produced by the ISLE (International Standards for Language Engineering) NIMM (Natural Interactivity and Multimodality) Working Group. See in particular ISLE NIMM reports D9.1, D9.2, and D11.1 at isle.nis.sdu.dk.

5.4.2 Fairy tale world evaluation focus and approach

The primary objective in evaluating the usability of the fairy tale world part is to get a handle on how well the system succeeds in engaging and entertaining the user. To this end, we must try to give some substance to inherently vague notions like "narrative progression", "dramatic effect" and "entertainment value" as experienced by the user.

So how can we pin down and measure notions like these? To begin with, we will use a "check list" with properties such as the following that together make up what we are aiming at:

- What is the story according to the user's perception, and to what extent do its events unfold in a timely and engaging manner?

- Do the characters display meaningful roles and believable personalities that contribute to the story? Are they aware of the user? Do they signal (verbally and non-verbally) that they understand the user (to the extent that they want to)? Can they handle out-of-domain and inappropriate utterances from the user (à la Eliza/Perry)? In other words, do they succeed in knowing that they don't know and in bringing back the dialogue "on track"? ("Not losing their faces.").
- Is there a real choice for the user among actions and multimodal acts in any given situation? To what extent is the user able to affect the plot (the particular sequence in which the underlying story "reveals itself")? To what extent does the user feel that she can affect the plot (not the same question as the previous one)?
- To what extent are actions and dialogue turns "story-functional"? In other words, do they contribute to the user gradually progressing in the story or appreciating the personality of a character (which in turn forms part of the story)? (The rationale for story-functionality is that loose ends that don't contribute to the narrative progression or effect serve no purpose.)

We intend to use questions like these in the questionnaires and interviews that will be instrumental in performing subjective evaluation. It is important to realise that those dimensions do not necessarily map onto individual components of the system, but rather to dramatic means at our disposal for building an engaging story and affecting the user.

As for quantitative measures of entertainment value, we plan to take a step towards investigating if it is possible to arrive at a meaningful scheme. We think that such a process might, if nothing else, provide useful inspiration for developing and refining our set of evaluation criteria.

More specifically, we plan to try out the feasibility of labelling of game sessions. To this end, we have to ask ourselves what are the primary mechanisms for driving the game forward in NICE. From the user's point of view, the answer is clearly multimodal dialogue and, to a lesser extent, moving around. As explained in Appendix 1, we assume that the user is not capable of any physical action in the fairy-tale world apart from moving around. From the characters' point of view, the primary mechanisms are multimodal dialogue, non-verbal signalling and physical action, including moving around in the fairy-tale world. From the system's point of view more generally, it is also the display of objects that play a role in the story.

The labelling could then be done with respect to the dimensions indicated by the check list. It should be centred around the primary mechanisms for driving the game forward, that is, the "turns" of multimodal acts as well as major changes of scenery and objects. In other words, the labelling should reflect how well these mechanisms serve to advance the game (tell a story) in a timely and engaging fashion, according to our chosen dimensions.

So what kind of tool do we need to be able to do labelling? Given that game sessions can be saved as video (such that we also overhear the user's utterances and keep track of his/her graphical gestures), one option is to use WaveSurfer with its new video plug-in:

WaveSurfer is a transcription tool from KTH developed by Jonas Beskow and Kåre Sjölander. It is freely downloadable from http://www.speech.kth.se/wavesurfer/. As a recent add-on, they have released a video plug-in that makes it possible to open video files in WaveSurfer, and to do labelling, annotations and simple cut-and-paste editing: http://www.speech.kth.se/wavesurfer/video/video.html

By defining tiers corresponding to relevant dimensions from the check list, we could label critical moments of game sessions while watching them as "movies", being able to freely play, rewind and replay snippets of them. Doing (and developing) a labelling like this is likely

to be very time-consuming, so we plan to try it out in a limited scale, since it might help us in developing and refining our set of evaluation criteria.

5.4.3 Test scenarios

To ensure common agreement of what the two parts of the NICE system should be able to do, and how, we propose to specify a set of test scenarios for the first and second prototype, respectively. Example scenarios are described below. It should be noted that the wording used in the examples will not necessarily be exactly the wording used in the actual test scenarios. We don't know exactly how HCA and the other characters will phrase themselves yet. Thus, the test scenario descriptions are rather meant to exemplify the kind of dialogue and interaction we expect a user will be able to have with the system. Sections 5.4.3.1 and 5.4.3.2 describe examples of dialogue and interaction in the first and second prototype, respectively.

5.4.3.1 Test scenarios for first prototype

The dialogue and interaction description examples are split into a sub-section relating to HCA and a sub-section relating to the fairy tale world characters. These two worlds will not be connected in the first prototype.

HCA

Test scenario dialogues for HCA will adhere to the following restrictions:

- be able to conduct five different conversations, one about each of his domains, including successful start and end greetings with users;
- the conversations will be about the topics we plan to implement for the first prototype;
- at least one topic will include pointing gesture by the user;
- whilst conducting the conversations, HCA will be in one of several different physical states in his study, such as walking around thinking, writing at his desk, etc.; and
- HCA will have to face sorting out a number of input error cases, such as out-of-domain (OOD) topics, OOD words, topic ambiguity, etc.

An example scenario could be the following:

In his study, HCA is standing at his desk reading a book. A user approaches and addresses him. HCA interrupts what he is doing and greets the user. The dialogue may evolve as follows.

U: Hello, you must be HCA
HCA: Yes, I'm HCA, hello
HCA: Who are you
U: I'm Mary
HCA: How old are you
U: I'm 15
U: I like your fairy tales
HCA: I'm delighted to hear that
HCA: Which one is your favourite
U: Perhaps Cloddy Hans or the Ugly duckling
U: How many did you write
HCA: I wrote about 180 fairy tales
HCA: My favourite is the ugly duckling but I also like the mermaid a lot
HCA: Do you know that one

U: Yes but it has a sad ending
U: The Disney version is much more positive so I like that one better
HCA: That must be something written after my time [HCA is a bit annoyed]

Figure 5.4.1. HCA's study.

Fairy tale world

Only one scene, one character (Cloddy Hans) and one plot, namely Plot 1 as described in Appendix 1, will be available.

The first prototype will include Swedish speech recognition and a NLU trained on Wizard-of-Oz data to be collected during spring and summer 2003. Response generation is coupled with turn-taking gestures, but without lip synchronisation. Simple 2D gestures at least comprising referral to/selection of physical objects present in the scene will be enabled.

5.4.3.2 Test scenarios for second prototype

HCA

The second prototype will enable a more sophisticated dialogue and improved animation. The underlying implementation will be improved on the basis of collected data and it will be extended with a user model, among other things. The HCA part will be linked to the fairy tale world part so that HCA will be able to guide the user to go though the door to the fairy tale world.

Fairy tale world

Prototype 2 will be as described in Appendix 1, with improvements relative to prototype 1 based on evaluation of this prototype as well as on successive Wizard-of-Oz data and iterative

development. There will be more characters and more plots compared to Prototype 1. The characters will use lip synchronised speech and they will have a richer repertoire of gestures and actions.

5.5 Data collection methods

Whenever users interact with the NICE system – no matter if simulated or real – system and user speech will be recorded and logged, and the user and the game hardware will be recorded on video using at least two cameras, to the extent possible. The video must record users' facial expressions, gestures, and body movements, and it must record what is on the NICE screen at the same time. It must always be ensured that the audio is captured in a quality sufficient to be used as training data for the recogniser.

The NICE software must enable logging of input and output as well as communication between each of the main modules in the system. The resulting session log files will support analysis and diagnosis of problems that may occur. Also, test suites should be generated which will allow systematic testing of components and of the entire system. Logfiles will be generated when the test suites are being run.

As mentioned above, subjective evaluation is needed for usability evaluation. The key methods for collecting data which can be used for usability evaluation remain questionnaires and interviews. These methods will be used extensively to collect information about users' opinion on the system both in controlled and in uncontrolled tests. Many tests will be uncontrolled in the sense that users will not be asked to perform particular scenarios but rather to just go ahead and use the system as they want to. A draft questionnaire is presented in Section 5.7.

During surveyed experiments, the experimenter often makes notes on any observation which s/he may find of importance or otherwise noteworthy. Such notes are also data which will feed into the analysis process and contribute to the evaluation of the system.

5.6 Data analysis

The following raw data resources will be included in the evaluation process:

- Audio and video recordings.
- Log files as soon as we have software to involve in experiments.
- Questionnaires and interviews.
- Notes made during test trials.

Raw data will be annotated as appropriate. For example, audio files will be transcribed. Exactly which additional annotation to add will depend on the information one wants to extract from the data. Of the raw data listed above, primarily audio and video files will be annotated. For example, we may want to mark up communication problems, the domain and topics the user is talking about, interaction initiative, and modality use. Video annotation may help us obtain a behavioural index of the quality of interaction. The sky is the limit here. Having annotated, e.g., the domains and topics found in the data, the domain coverage of the system compared to what users talk about may be evaluated. In the first prototype evaluation, we have to distinguish between user input which is not yet covered but is within the planned domain coverage for the second prototype, and input which is outside the domains altogether and thus is not planned to be covered by the second prototype either. If much user input concerns a domain which has not been included it should be considered for inclusion in the second prototype.

From questionnaire answers ticked off on a Likert scale, it is pretty easy to extract statistical information on users' opinions on the system and to run comparisons across versions. Free-style comments in questionnaires and interviews can be difficult to evaluate but will often contain valuable information on problems with the system which may otherwise be difficult to become aware of.

Log files are very useful for locating bugs and other problems observed during component testing, system interaction, or in the audio/video files. It may be helpful to annotate the log files in a way which makes it easy to extract core information and convert it into an easy-to-read format.

Observation notes are rarely annotated systematically but are rather scrutinised and used as keys to analyse in detail, e.g., certain interaction patterns, apparent bugs, or particular interaction problems.

5.7 Draft questionnaire

The following is a draft proposal for a questionnaire for use with users trying a simulated or implemented version of (parts of) the system. The questionnaire is likely to be modified according to the focus point of the test set up. For example, for a test of the fairy tale world, part the questionnaire will probably add a number of questions related to edutainment value and perhaps leave out some of the questions related to basic usability criteria.

Computer games experience	None at all						More than 500 hours
Knowledge about HCA	None						Knows about everything
Knowledge about HCA fairy tales	None						Knows about everything
Age (years)							
Gender (male/female)							
What was your overall satisfaction level with the system	Very low						Very high
Did you have any difficulties in getting started	A lot						None at all
Did you know what to do	Never						Always
How did you find the output voice quality	Very incomprehensible						Very intelligible
	Very synthetic						Very natural
How did you find the output animation quality	Very bad						Very good
How did you find the quality of graphics	Very bad						Very good
How was the output behaviour (the combination of speech and graphics, the handling of	Very artificial						Very natural

emotions)							
How was the spoken output phrasing	Very inadequate						Very adequate
How was the system's understanding of spoken input	Very bad						Very good
How was the system's understanding of gesture	Very bad						Very good
How was the use of the gesture input device (to be specified as soon as this device is determined).	Very difficult						Very easy
How was your interaction with the system (modalities, dialogue flow)	Very artificial						Very natural
Could you talk to the system about the topics you wanted to talk about	Never						Always
Could you control the dialogue if you wanted to	Not at all						Any time
How was the system's output in return to your input	Always inappropriate						Always appropriate
Was the system able to treat misunderstandings or errors in an adequate way	Never						Always
How did you find the type of game	There are many other games like this						Very original
How did you find the entertainment value	Very low						Very high
Did you learn something from using the game	Nothing at all						A lot
How interesting was the game	Very boring						Very exciting
Would you like to play the game again	Definitely not						I'd love to
What did you like about the NICE system							
What didn't you like about the NICE system							
What should be improved							
Other comments							

6 References

Beringer, N., Hans, S., Louka, K. and Tang, J.: How to Relate User Satisfaction and System Performance in Multimodal Dialogue Systems? - A Graphical Approach. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Copenhagen, June, 2002, 8-14.

Dehn, D.M. and van Mulken S.: The impact of animated interface agents: a review of empirical research. International Journal of Human-Computer Studies, 52, 2000, 1-22.

Dybkjær, L., Bernsen, N. O. and Minker, W.: Overview of Evaluation and Usability. In Minker, W., Buhler, D. and Dybkjær, L. (Eds.): Spoken Multimodal Human-Computer Dialogue in Mobile Environments. Kluwer Academic Publishers, 2003 (to appear).

Höök, K.: Evaluation of Affective Interfaces. AAMAS 2002 (First International Joint Conference on Autonomous Agents and Multi-Agent Systems), Bologna, Italy, July 15-19, 2002.

Picard, R.W.: Affective computing MIT Press, Cambridge, MA, USA, 1997

Walker, M., Litman, D., Kamm, C. and Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics , ACL 97, 1997.

Web sites:

DISC: www.disc2.dk

ISLE: isle.nis.sdu.dk

NITE: nite.nis.sdu.dk

WaveSurfer: www.speech.kth.se/wavesurfer

7 Appendix 1: Outline of game and story in NICE fairy-tale world

This appendix summarizes the discussions so far between Telia and Liquid Media on the characteristics of the game and story in the fairy-tale world of NICE. In particular, it focuses on:

- What characters inhabit the world ("casting") and what objects they manipulate.
- What the general story is about ("storyline").
- The user's role in the game.

7.1 Characters

The fairy-tale world is inhabited by autonomous animated characters inspired by figures from H C Andersen's stories. The characters are equipped with personalities as well as short- and long-term goals that together make up their personal traits. The original Andersen characters that we have discussed so far are:

Cloddy Hans, adapted as follows: He is a bit stupid, or so it seems. He cannot read, and only understands oral/multimodal instructions at a rather detailed level. Lacks initiative, but honest and anxious to try to help the user. Physically strong and fearless. In spite of his limited intellectual capabilities, he may sometimes provide important clues through sudden flashes of insight. Most importantly, he is the user's faithful assistant who follows him throughout the game.

Thumbelina, adapted as follows: She has been enticed to evil by a wicked user, and H.C. Andersen has lost all control over her. In connection with her transformation, she has also increased her size to that of a small girl and has got supernatural physical powers. She terrorizes the fairy-tale world by scaring and deceiving its creatures and by physical destruction. The only way to save her (and the world) is to bring her back to Andersen for re-programming.

In addition, we plan to use a set of more prototypical characters that may still be inspired by Andersen's stories, for example:

- A witch.
- A prince.
- A princess.
- A soldier.
- A peasant girl.

Finally, we plan to re-use H.C. Andersen himself as a kind of meta character.

7.2 General story

7.2.1 Structure of a story

We assume that the (generic) story consists of a series of *plots*, which can be enacted in differing orders according to a partial ordering. A plot corresponds to a problem or hindrance of some kind that needs to be overcome, and which is functional from the point of the user progressing through the general story. To further increase the variation allowed by the partial ordering, it is possible to turn from one plot to another before the preceding one is brought to an end, thereby leaving "dangling" plots (which may be resumed later on).

A plot consists of one or several *scenes*, each of which is enacted at a particular physical location. A plot may thus require the user to move between several locations in order to solve the corresponding problem. Scenes are in turn divided into *beats*, which typically correspond to dialogue about a subtopic in the scene.

In the rest of this section, we outline the beginning and end of the story, and give some ideas on intermediate plots.

7.2.2 Introduction

H.C. Andersen has succeeded in building a thriving fairy tale world full of harmony and beauty, namely, the game world that the user is about to enter. The world is good because it complies with Andersen's philosophy that everybody and everything should belong in their proper place based on their abilities. As someone has expressed it: "Down with those who are unfit, ahead with those who are fit, regardless of whether they are rich or poor."

There is a problem, however. One of Andersen's most gentle and beloved characters, Thumbelina, has been enticed to evil by a wicked user. She now spreads physical destruction and fear among all creatures in the fairy tale world. There is a risk that she will destroy the entire world if nothing is done to prevent her.

Andersen urgently requests the user to help him to restore his world by finding Thumbelina and bringing her back to him for re-programming. To this end, it is important to understand the following keys to her behaviour:

- Since it was a user who transformed Thumbelina, she despises fairy tale characters and only respects and listens to users. Hence, the user's participation and help is vital from Andersen's point of view.
- Like a classical serial killer, she leaves behind on the scene of every criminal act a clue or riddle that allows the user to eventually track her down. Deep down, she wants to be caught and restored to her original self, but only by a sufficiently cunning "detective".

The task of the user is thus to make an odyssey in the fairy tale world in order to find Thumbelina and somehow bring her back to her creator. In the course of this, he will both have to deal with various obstacles that are instigated by Thumbelina or that provide clues to her doings.

This introduction (whether it is just given as a briefing outside of the game, displayed in a cut scene or communicated through game dialogue) ends with the user entering the fairy tale world through a stargate door.

7.2.3 Plot 1

Scene 1

Location: Behind the ravine.

Characters: Cloddy Hans and the user.

Beat 1: Cloddy Hans introduces himself.

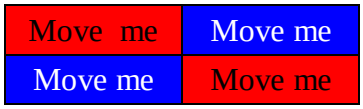
Beat 2: Cloddy Hans describes his problem and asks the user for help.

Beat 3: The user and Cloddy Hans together solve the problem (that is, the user gives piecemeal instructions to Cloddy Hans on what to do through multimodal dialogue).

The user finds himself trapped in an isolated corner of the fairy tale world, beside Cloddy Hans. A ravine separates the location from the rest of the fairy tale world, which is seen in the background. The sole connection across the ravine is a bridge cut off by a closed door.

Cloddy Hans informs the user that Thumbelina has locked him in here and that they must cross the bridge to find her in the fairy tale world and save her.

The entrance to the bridge is blocked by four similar boxes — two red and two blue ones — which are piled immediately in front of it. Each box carries the text ”Move me”. From straight ahead it looks like this, in principle:



Thus, to get across the bridge, the user must somehow get these boxes out of the way.

A bit aside, a similar red and a similar blue box are sitting on the ground. These boxes do not carry any text. Their sole purpose is to increase the ambiguity, thereby forcing the user to be more precise when referring to boxes.

The purpose of this overall set-up is to train the user in a basic multimodal utterance, namely, ”move X” (where the reference to X should preferably be expressed by a graphic gesture). If the user only makes use of a single modality, Cloddy Hans will ask a clarification question in such a way that the other modality is likely to be elicited. For example, if the utterance did not include graphics, Cloddy Hans might ask the user to literally point at the relevant box for clarification. It is also not certain that Cloddy Hans understands utterances like ”left” and ”right”. Another constraint is that he can only lift one box at a time.

7.2.4 Intermediary plots

The plots have to be carefully designed to ensure that multimodal dialogue between the user and characters of the game has a real purpose. Basically, the solutions of the problems should be more or less obvious even to nine-year olds, as we want to allow our users to concentrate on what is new in this game — multimodal reference and dialogue. (Hopefully, plot 1 is a good example of this.) Another constraint on plots is that they should be functional to the overall story, and not just added as arbitrary obstacles that are independent of the story line.

There are three classes of sources for the plots that the user will encounter:

1. Conflicting goals between the fairy tale characters lead to problematic situations
2. Fairy tale characters have been manipulated or put in hard situations
3. Objects and locations have been manipulated or removed

Some ideas on intermediate plots:

- A gorge guarded by a witch who requires the user to solve a new problem each time he wants to pass. This might be instigated by Thumbelina or it could be that the witch provides some crucial information on the doings of her.
- Someone has been deprived of a magic object which makes people reveal their thoughts to the owner of the object. The user can find it by solving a graphical problem and may then keep the object for later use or give it back.

- Someone has been locked into a room by Thumbelina. Outside the room, there is a written query that the user has to answer (perhaps by asking the person locked in) in order to give the right instruction to Cloddy Hans on how to unlock the door.
- The user has to obtain a piece of information from character A needed by character B and go tell B about it.

7.2.5 Plot N (final)

Scene 1 (part of introduction?)

Location: Any location where the user is.
Character: HCA.
Beat 1: HCA tells the user that he must bring Thumbelina to him.

Scene 2

Plats: Somewhere else.
Character: Thumbelina
Beat 1: The user must persuade Thumbelina to go to HCA.

Scene 3 (possibly cut scene)

Plats: HCA's study.
Character: Thumbelina and HCA
Beat 1: Thumbelina is somehow reset/re-programmed by HCA.

7.3 Role of the user

The task of the user is to find Thumbelina and somehow bring her back to her creator by making an odyssey in the fairy-tale world together with Cloddy Hans. In the course of this, he will have to deal with various obstacles somehow related to Thumbelina. To the extent that the user remains passive, various events are going to take place anyway (in other words, there should be a sense of real time and actions taken by other characters, independently of the user).

The user will perceive the fairy-tale world through a first-person perspective. Thus, there will be no user avatar. (We still assume that the user will be perceived as appearing in the world by other characters in the game.) Furthermore, the user's only means of physical action in the world are:

- moving around;
- changing his camera (looking around);
- pointing at arbitrary characters, objects and locations.

There are two reasons for these limitations: To begin with, what distinguishes NICE from other games is multimodal dialogue. Hence, we have to make sure that multimodal dialogue is appreciated by the user not just as an "add-on" but as *the primary means of progressing in the game*. Our key to achieving this is to deliberately limit the capabilities of the key actors — the user and Cloddy Hans — in such a way that they can succeed only by cooperation and by engaging in dialogue. Thus, the user is intelligent but cannot himself affect objects in the world; Cloddy Hans is fairly stupid but capable of physical action according to what he gets told (and may occasionally also provide tips to the user).

Secondly, spending a lot of resources on animation and control of the user's hands and the detailed actions carried out with them would risk side-tracking the project.

The user interface is based on a first-person perspective (as mentioned above), in which the user has access to a mouse-compatible input device for the purpose of moving around in the 3D fairy-tale world. When the user meets fairy-tale characters they will be shown in full-body camera angle. Typically a number of 3D objects will also be seen on the screen. The user can ask Cloddy Hans to manipulate objects in the screen by referring to them verbally and/or by using the mouse.