NICE project (IST-2001-35293)

Natural Interactive Communication for Edutainment

NICE Deliverable D9.3

Final report

4 May 2005

Authors

The NICE team

Project ref. no.	IST-2001-35293
Project acronym	NICE
Deliverable status	Public
Contractual date of delivery	4 May 2005
Actual date of delivery	4 May 2005
Deliverable number	9.3
Deliverable title	Final Report
Nature	Report
Status & version	Final
Number of pages	29
WP contributing to the deliverable	9
WP / Task responsible	SDU
Editor(s)	Niels Ole Bernsen and Svend Kiilerich
Author(s)	The NICE team
EC Project Officer	Mats Ljungqvist
Keywords	NICE project results and future perspectives
Abstract (for dissemination)	This final report from the HLT project Natural Interactive Communication for Edutainment (NICE), Deliverable 9.3, describes the results achieved in project and the use the partners are planning to make of the research progress made on multiple fronts in the course of the project.

Table of Contents

1 Project overview

1.1 Composition of the NICE consortium and roles of the partners

CNRS/LIMSI, France

- Gesture Recogniser (GR)
- Gesture Interpreter (GI)
- Multimodal input fusion (IF)
- Data collection

Liquid Media, Sweden

- Virtual world graphics
- System integration

NISLab, Denmark

- Hans Christian Andersen (HCA) conversation
- Natural Language Understanding (NLU)
- Conversation Management (CM)
- Response Generation (RG)
- Data collection
- Project coordination

ScanSoft, Germany

- Dedicated children's speech recognition

TeliaSonera, Sweden

- Computer gaming
- Natural Language Understanding (NLU)
- Conversation Management (CM)
- Response Generation (RG)
- Data collection

1.2 Main project achievements

1.2.1 Two demonstrators

During the past three years (March 2002 – February 2005), the NICE consortium has developed two interactive multimodal prototypes of each of two related systems, a system for edutaining English conversation with fairytale author Hans Christian Andersen (the HCA system) and a system for playful, computer-game style Swedish conversation with some of his fairytale characters (the Fairy Tale World, FTW, system). Both systems target 10-18 years old children and teenagers, enabling them to use combined speech and 2D gesture in conversation with 3D embodied conversational characters developed on a professional computer games platform. These relatively complex systems have been built close to their original specifications and both of their respective prototypes have been sufficiently mature to enable user testing with target group users, yielding valuable data for analysing novel aspects of human communication with 3D embodied conversational characters. Jointly, the two systems demonstrate entirely new opportunities for exploiting embodied conversational character technologies for purposes of edutainment and entertainment. Equally significantly for content-rich systems such as the ones described in this report, both systems are, to a significant extent, portable to new applications rather than being technical monuments which must be re-implemented from scratch for each new application.

1.2.2 Component-level brief

Both NICE systems are composed of a mixture of off-the-shelf components and components which have been iteratively developed in the project. *Off-the-shelf components* include the systems' common message broker and the speech synthesiser for English. All other components have been developed in the project, sometimes on the basis of existing technologies. The *components developed on the basis of existing technologies* include the gesture recogniser which is common to both systems and which has been developed on the basis of open-source software, and the two speech recognisers which were supplied by project partner ScanSoft. The *components that were developed entirely within the project* include:

- gesture interpretation, common to both systems
- natural language understanding, different in the two systems
- speech/gesture input fusion, common to both systems
- conversation/dialogue management, different in the two systems, and
- response generation, partially different in the two systems
- speech synthesis for Swedish

1.2.3 Main challenges

At the start of NICE, the consortium faced the challenge that key project aims had not been achieved before. This was true, in particular, of the aims of demonstrating in running systems:

- workable speech recognition children and teenagers' voices
- computer games with advanced spoken interaction
- use of spoken dialogue and conversation technologies for edutainment and entertainment, and
- domain-oriented, as opposed to task-oriented, spoken interaction in which the system is apple to handle spoken conversation in semi-open domains

A further challenge which should be mentioned here is that of integrating spoken interaction with 2D gesture input in the context of edutainment and entertainment applications.

1.2.4 Main NICE achievements

Throughout the NICE project, the focus has been on meeting those challenges by developing system architectures that combine state-of-the-art components for gesture and speech recognition, natural language and gesture understanding, dialogue and conversation management, and speech and graphical output. The resulting architectures are distributed and extensible and can well be reused for similar applications.

The consortium is pleased that, through three years of exiting and extensive cooperation involving all partners and both NICE systems, the challenges listed above have all been satisfactorily met in the project. The resulting system demonstrators extend the limits of technical feasibility in several respects. Examples of these achievements are:

- conversational speech recognition for (non-native) spontaneous children speakers in large and semi-open conversation domains and with users unfamiliar with the application
- natural language understanding components dealing robustly with the speech recognition error rates that can be expected in large domains

- a gesture input processing “chain” of gesture recognition, gesture interpretation, and semantic-level input fusion of linguistic and gesture contents, enabling interpretation of a wide variety of graphical input gestures
- advanced management of fully mixed initiative speech and gesture input/full natural interactive output conversation in large domains (HCA system)
- event-driven dialogue management of multi-party dialogues, processing user utterances and managing several game characters who interact through spoken dialogue with the user as well as with each other, as well as moving about in a three-dimensional fairytale world and carrying out various actions, such as picking up and putting down objects, pulling levers and winding down a drawbridge (FTW system)
- a professional computer game environment with an easy-to-use interface that allows for blending of animations, lip-synchronization, and sub-second control of the character's movements, enabling the creation of extremely reactive animated conversational characters

In addition, it should be mentioned that NICE has generated voluminous data resources, partly through the two user tests of both NICE systems, and partly through additional, extensive Wizard of Oz data collection exercises. The purposes of the latter were to collect speech data for training the speech recognisers with children’s voices, develop recogniser language models, develop gesture recognition and gesture interpretation approaches, develop natural language understanding, and iteratively explore and test emerging design specifications for the spoken interaction between user and system. NISLab’s data collections will be available to non-project parties and will certainly prove interesting for further research on speech recognition, children’s spoken conversational interaction with embodied animated characters, and children’s interaction through speech and gesture: children’s speech recognition, non-native speech recognition, hesitation modeling, speech endpointing and noise modeling, spoken social interaction, domain-oriented conversation, conversation through speech and gesture, etc. LIMSI’s data collections will be available as well, inviting further investigation of children’s use of gesture. The data resources available from the project will lead to even better methods for natural language understanding and conversation management for natural interactive systems. The simultaneous collection of speech *and* gesture data in this kind of domain is very difficult. Traditional simulation methods like Wizard-of-Oz are not really applicable, due to the problems for a manual operator to keep up the system's end of the dialogue (which might involve several characters), and control the movements and actions of the characters at the same time. TeliaSonera's methodology of collecting data with a supervised system, which is mainly automatic but allows for human intervention by an operator, combines the speed of an automatic system with the flexibility of a Wizard-of-Oz data collection tool.

2 Project objectives

2.1 The original project goal

In the project proposal and the following Technical Annex (TA) to the NICE project contract, the goal of the project was presented as being:

- to build a prototype system that will enable kids, youngsters, and anyone else to have entertaining multimodal conversation with H.C. Andersen and his fairy tale characters

The goal would be addressed

- through joint expertise in professional computer games, speech recognition, and multimodal spoken dialogue systems

The TA then goes on to present the main challenges addressed in NICE (see Section 1,2.3), all of which requiring, at the time, advances beyond the state of the art.

2.2 Comparison with the actual project results

When comparing the main project achievements (Section 1.2.4) with the project goal quoted above, it may be noted that the NICE project has, in fact, developed *two partly different* research demonstrators (Section 1.2.2), the HCA system and the FTW system, in two different languages, i.e., English and Swedish, respectively. The project goal envisions a single system for HCA and the FTW but one which would be fully developed for both English and Swedish, enabling, in the process, linking of HCA's study in which he conducts domain-oriented conversation with the young users, with the fairytale world. This difference which is, in fact, the sole difference between the NICE goals and the project's actual achievements, has the following explanation.

Very early in the project, at the initial requirement specification workshops, it became clear that FTW spoken interaction, plot-structured, object-handling, and multi-character as it was, would require very different natural language understanding and dialogue/conversation management compared to the HCA spoken interaction which was aimed to be domain-oriented, non-plot-structured, and single-character-only. It was therefore agreed to develop natural language understanding and dialogue/conversation management independently for the two systems. The consequences of this consortium decision is described in Section 2.

Was it a good idea to develop two partly different systems in the NICE project? There are both pros and cons here.

On the "con" side, we have missed the opportunity to investigate linguistic portability from English to Swedish and vice versa, between the HCA and FTW systems. This is an important field of research which could have benefited from the project's results. We have also missed the opportunity to investigate challenges arising from the need to develop a single dialogue/conversation manager for two rather different purposes. Could we have succeeded in developing such a manager? We cannot demonstrate that we could have.

On the "pro" side, we have been able to proceed unhampered by partner compromises to develop appropriate natural language understanding and dialogue/conversation manager solutions to the rather different challenges faced by the HCA and FTW system, respectively, and we have actually succeeded in these endeavours, as shown in the public NICE report on the results of evaluating the user test with the second prototypes of the two NICE systems. Collaboration-wise, the approach adopted has implied less collaboration than anticipated between TeliaSonera and NISLab but as much collaboration as originally intended between the project partners in all other respects. Moreover, for the partners who have delivered components essential to both systems, such as graphical rendering, gesture processing, and speech recognition, the generalisations of approach needed for success have provided valuable insights.

Given the results achieved in the NICE project and the serious research challenges at stake, we feel that we made the right decision early in the project. For instance, it may be argued that NISLab's conversation manager is fully capable of managing FTW dialogue even though this was never a design objective. However, this example illustrates that NICE partners may have been able to penetrate further into unknown territory because of the decision made than would have been the case otherwise.

As for the linking between HCA's study and the fairytale world, this is a straightforward problem. Nothing significant, it appears, would have been gained from its solution which, in its basic form, would have been that of enabling a user to walk into the fairytale world from HCA's study if and when HCA allowed the user do this. In fact, NISLab has done the conversation design for this to happen but the design was never implemented for the reasons discussed above.

3 Project results and achievements

According to its (European Commission) specification, the present chapter should address the systems, services, tools, methods or models arrived at and how they relate to current or prospective markets or user needs. We would like to first remind the reader of the project's main achievements (Section 1.2.4) and then proceed to expand on their relationships with current or prospective markets or user needs. The issue of innovative methodologies will be discussed in Chapter 4:

- conversational speech recognition for (non-native) spontaneous children speakers in large and semi-open conversation domains and with users unfamiliar with the application
- natural language understanding components dealing robustly with the speech recognition error rates that can be expected in large domains
- a gesture input processing "chain" of gesture recognition, gesture interpretation, and semantic-level input fusion of linguistic and gesture contents, enabling interpretation of a wide variety of graphical input gestures
- advanced management of fully mixed initiative speech and gesture input/full natural interactive output conversation in large domains (HCA system)
- event-driven dialogue management of multi-party dialogues, processing user utterances and managing several game characters who interact through spoken dialogue with the user as well as with each other, as well as moving about in a three-dimensional fairytale world and carrying out various actions, such as picking up and putting down objects, pulling levers and winding down a drawbridge (FTW system)
- a professional computer game environment with an easy-to-use interface that allows for blending of animations, lip-synchronization, and sub-second control of the character's movements, enabling the creation of extremely reactive animated conversational characters

Recent advances in *speech and language technology* have mostly benefited adult users. The results generated by the NICE project contribute to improving the technologies also for young users. This opens up a new commercial market, namely, services for children and adolescents, but is also a democratic issue. Many speech-enabled services concern access to information, and as long as speech technology works badly for young users, those users will be shut off from the information such services provide. The NICE results will also be used to improve speech recognition for Swedish people in general, thus improving the prospects for future speech-enabled services in Sweden.

As regards *gesture input combined with spoken input,* NICE results suggest a dilemma for computer games designers. In brief, they have to consider choosing between mainly visual-motor interaction (e.g., navigation in the 3D world, moving 3D objects) which ignores spoken interaction, on the one hand, and considered, spoken and gesture interaction, on the other, the use of either tactile screen or a gyro-mouse requiring different communication infrastructure between the modules since eliciting different gesture and multimodal behavior. Game designers should also limit to a small set of semantic output in either of the modalities at a given step in the game in order to obtain robust multimodal integration. The permanent availability of both input modalities will nevertheless enable to cover different users' preference regarding the interaction style they are free to select at any given time. In practice, fast visual-motor interaction seems incompatible with more fully natural interactive "think-before-you-speak-and-point" interaction. Two rather different styles of computer gaming may result from this observation.

Like shallow-but-adequate natural language processing, *multimodal and natural interactive dialogue/conversation management* remains a field characterised by little or no de facto consensus and standardisation. This makes it difficult for newcomers in the field, including companies, to go full-speed towards a working solution for their applications. For these stakeholders, NICE offers a modality-independent, fully mixed-initiative conversation manager which can handle any combination of input modalities and which can be used for task-oriented dialogue as well as for domain-oriented conversation (HCA system). In addition, NICE offers a dialogue manager for multi-party interaction between embodied animated characters and users, pushing the state of the art forward towards the ultimate goal of natural interactive systems.

The NICE results also open up new possibilities in *computer gaming*. So far, games involving spoken user commands have occupied a small niche on the computer games market, and those games have not received very good reviews from the experts in the area. The knowledge gained within NICE can be put to use in designing better games involving non-command-based, spontaneous spoken dialogue and domain-oriented conversation, potentially opening up a huge market. In fact, the user tests of both of the NICE prototype systems showed that multimodal applications using speech I/O are highly enjoyable and indicated that there is a large market potential for games using the combination of AI character modeling, nonverbal behavior and speech I/O.

The NICE results will also be put to use in designing *more entertaining interfaces to commercial services*. 'Entertainment value' is increasingly becoming a competitive factor; if two services A and B are comparable except for the fact that A has a more entertaining interface, customers might be more likely to choose A. Also, the methods developed within NICE can be used for edutainment purposes in exhibitions, museums, etc. Domain-oriented, conversational and multimodal applications like the ones achieved in the NICE project can be expected to gain much interest in future markets. This type of system is appealing to a large number of users, because it is:

- **entertaining**: the combination of various system inputs and outputs (gesture, speech, graphics) enable lively and immersive interaction
- **easy to use**: the user does not have to read instructions to successfully interact with the system
- **versatile**: the system can be used for game playing, for education or for a combination of both. There are virtually no limits in terms of the game/educational content
- **cost-efficient**: the NICE prototypes are automatic systems that can be placed almost anywhere

As the NICE systems are still research prototypes, much work can be done to make the development of similar systems easier and faster. The basic architectures of the NICE prototypes seem to be flexible enough to host a large variety of similar edutainment applications.

4 Methodologies

This chapter includes comments on the advantages of methodologies used over other possible alternatives through a review of the world-wide state-of-the-art.

There is an important sense in which the NICE systems are simply interactive spoken computer games. This potentially revolutionary field of computer games was non-existent when the NICE project began. Computer games with *spoken output* did exist, to be sure. Today, we have seen the first products on the market which feature *spoken input* which can make the game characters do certain things. So far, these products do not seem to be terribly popular with the games reviewers, which seems to be due to the fact that most or all of them assume that the gameplayer is able to learn large numbers of spoken commands. It should be emphasised here that neither spoken output alone not spoken input alone makes an *interactive spoken* computer game. That requires the user to be able to use spoken input and to be appropriately responded to by the game character(s) in spoken output. To our knowledge, commercial interactive spoken computer games are today, at best, in their early infancy.

Thus, to find appropriate interactive spoken computer games with which to compare the NICE research prototype systems, one has to turn to the field of research on natural interactive communication with animated graphical characters which is sometimes called research on embodied conversational agents (ECAs, cf. Cassell et al. 2000). As it should, given the enormous challenges that have to be overcome in order to achieve full human-style natural interactive communication, research on ECAs is an extremely multi-dimensional endeavour, ranging from fine-tuning the details of lip synchronisation for some particular language through the addition of computer vision to ECAs to highly theory-based papers on social conversation skills and multiple emotions which ECAs might come to include in the future.

Among this wealth of research directions, the consortium finds that the two NICE prototype systems stand out in various ways, as will be explained in more detail below. For one thing, most of the ECA community put less emphasis on spoken interaction than we have had to do in NICE, given the initial challenges addressed by the project in the related fields of spontaneous speech recognition of children’s voices, high-complexity robust natural language understanding, spoken conversation management, and multi-character speech synthesis. Secondly, few members of the ECA community have the luck to collaborate with a professional computer games company as we have been able to in NICE. Thirdly, few ECA researchers have ventured into the highly complex territory of semantic gesture/speech input fusion. Fourthly, except for the NICE consortium, the ECA community has so far not addressed the challenges of domain-oriented conversation with a famous character from the past, such as HCA. There may be even more reasons but those just mentioned may serve to explain why we have identified precious few systems that come close to the NICE prototype systems in what in the end may be described as a determined effort to achieve complete demonstrators of what interactive spoken computer games for edutainment and entertainment could look like.

4.1 Speech recognition

4.1.1 Methodologies

As ScanSoft provided the speech recognition technology for the NICE project, and since NICE is mainly targeting children users, the focus of ScanSoft’s research was on improving recognition accuracy for children speech. Different approaches had to be taken for the

English and the Swedish NICE system, depending on the availability of in-domain speech training data.

For English, the data collections done within the NICE project (NISLab) were sufficient to train NICE-specific acoustic modeling resources. The main complication here was to optimize the training process by iteratively expanding the training set, starting from simple utterances, and ending with the complete data set. A set of criteria has been developed to define the range of utterances used at each training iteration. A standard training process would have generated significantly inferior acoustic models. Note that in the English version of the NICE prototype most users are non-native English-speaking children, which is the reason for an extraordinarily high amount of noise and hesitations within utterances.

The Swedish NICE prototype targets children speakers, too. In contrast to the English version, the speakers use their native language. Here, 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.

4.1.2 Recent related work

Looking at other recent research done on children speech recognition, 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 and 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 (Giuliani and Gerosa 2003, Gerosa and Giuliani 2004, Hagen et al. 2003), generally achieving error rate improvements in the usual range for VTLN, i.e. about 10% relative.

Li and Russell (2003) point out the importance of children-specific pronunciation modeling by using, e.g., 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.

4.2 Gesture input processing

Since the start of the NICE project, advances have been achieved in several directions in the area of multimodal interfaces. Prototypes have been developed in several task-based professional oriented applications such as: integration of 2D pen gestures for graphical design (Milota 2004) ; integrating speech, 3D gestures and 3D virtual reality graphics with a limited amount of possible commands in each modality (Kaiser et al. 2003); integration of a small set of spoken and gestural commands during mobile interaction in an outdoor rescue mission simulation (Kumar et al. 2004).

Evaluation of multimodal input systems and annotation of multimodal corpora have also progressed (Martin et al. 2004), including the study of multimodal integration patterns of seniors (Xiao et al. 2003), and the evaluation of Embodied Conversational Agents (Ruttkay & Pelachaud 2004).

Regarding specification and software architecture issues, the World Wide Web Consortium has recently started standardization initiatives in several directions of multimodal interaction in the "Multimodal Interaction Activity": the specification of gesture input representation in the InkML initiative (Chee et al. 2004), the specification of the communication between components of multimodal input systems in the EMMA initiative (Extensible MultiModal Annotation markup language) (Avaya et al. 2004), and more generally Multimodal Architecture and Interfaces (Barnett 2005).

Finally, a new area of bidirectionnal interaction with Embodied Conversational Agents has emerged lately as demonstrated by the Workshops held at the International conference on Autonomous Agent and Multiagent Systems (AAMAS) in 2003/2004, and by bidirectionnal systems and studies, such as the MAX system enabling analysis and synthesis of deictics and iconic gesture in an assembly task (Pfeiffer & Latoschik 2004).

Yet, to our knowledge, the NICE project remains the only attempt to study and achieve multimodal input in the case of children interacting with 3D objects and a conversational (non task based) edutainment character. One of the few contributions we know of is the study with a simulated system described in (Xiao et al. 2002) where the character and the graphics were in 2D. The various levels of simulation used in the different NICE prototypes enabled us to collect corpora which provide knowledge on how children use 2D gesture and their combination with speech while interacting with a character and 3D objects.

Regarding gesture and multimodal input, the main achievements of the project are thus at multiple levels:

- a better understanding of the requirements for the real-time processing of gesture and multimodal input raised by such a bidirectional conversation edutainment system

- the specifications of the individual modules processing individual modalities and their combination, and a XML specification language for representing the exchanges between these modules; indeed, these languages defined in the NICE project for describing the messages exchanged between the different input modalities share some common features with the more complex upcoming standards efforts at the W3C such as InkML and EMMA

- the required architectural principles, including combinations of feed-forward and feed-back messages between modules managing various input modalities at different levels of abstraction, temporality in order to ensure proper time management. and the production of output consistent with input

- the algorithms and the software implementation in Java of the Gesture Interpretation and the Input Fusion modules, which can combine temporal, singular/plural, semantic/ perceptual dimensions in a less restrictive and hence more robust fashion than the unification techniques used in task-based multimodal systems

- the definition of experimental protocols for evaluating such bidirectional edutainment systems used by children and teenagers and the proposal of typologies for the analysis of their multimodal behavior (Buisine & Martin 2005) ; software have been integrated or developed for annotating user's multimodal behaviour in such a complex set up. For instance, information logged by the modules during the execution have been incorporated into Anvil for comparative analysis of the videos and manual validation/annotation followed by statistics computation

- a corpus of 2D gestures during interaction with 3D objects which can be useful for training gesture recogniser in similar context

4.2.1 Natural language understanding

Approaches to robust parsing can be divided into data-driven and symbolic methods, the former of which have been the focus of a steadily growing interest during the last decade. One strand of work in this area deals with syntactic parsing in the sense of deriving a constituent structure or a dependency structure (for example, Collins 1999, Charniak 2000, Nivre and Scholz 2004), but without the specific requirement of producing output that serves the needs of a dialogue manager. Another strand of work, namely, "How may I help you" type systems, explicitly aims at integrating robust understanding with a dialogue system, but with a semantic representation that is limited to atomic categories. Thus, parsing here corresponds rather to classification of utterances into a small set of categories — for example, 15 in the classic AT&T "How may I help you" system (Gorin et al. 1997), and generally not more than a few hundred in more recent systems. We are not thus aware of any approaches that make use of automatic, data-driven methods to derive the kind of complex semantic structures that are needed by a dialogue manager in a domain like the fairy-tale game (FTW system).

Turning to symbolic, rule-based approaches to robust parsing, one option, pioneered by Ward (1989), is to rely on pattern-matching and to use a relatively coarse-grained semantic representation, such as a variable-free slot–filler list. Other instances of work in this shallow-parsing direction are Jackson et al. (1991) and Aust et al. (1995).

However, conversational applications such as the NICE systems tend to require more fine-grained semantic formalisms in order to sufficiently capture the meaning of user utterances. For example, variable-free slot–filler lists are not suitable for negotiative dialogue, in which several alternative solutions are simultaneously discussed and compared (Boye and Wirén 2003b; Larsson 2002). On the other hand, the computational price for adopting a general-purpose logic-based formalism and general semantic reasoning is likely to be too high in an application where savvy users will not accept having to wait for the system to come back with an answer.

Several attempts at finding a suitable trade-off by synthesizing the shallow and logic-based approaches have been made. One possibility is to "robustify'" some general-purpose linguistic method, either by homing in on the largest grammatical fragment (Boye et al. 1999), or on the smallest set of grammatical fragments that span the whole utterance (see for example van Noord et al. 1999 and Kasper et al. 1999). Another possibility is to extend the pattern-matching approach with the capability of handling general linguistic rules. For example, the parser of Milward and Knight (2001) makes use of linguistically motivated rules, representing the analysis as a chart structure. Semantic interpretation is carried out by mapping rules that operate directly on the chart. These rules incorporate task-specific as well as structural (linguistic) and contextual information. By giving preference to mapping rules that are more specific (in the sense of satisfying more constraints), grammatical information can be used whenever available. However, the semantic representations produced are still limited to that of variable-free slot–filler lists. In contrast, Boye and Wirén (2003a, 2003b) put forward a more fine-grained formalism in which a type system is used instead of general semantic reasoning; hence, the system is still much more restricted than general-purpose logic-based formalisms. The parser and semantic formalism used in the fairy-tale game constitute a further development and application to new domain of that framework.

4.2.2 Dialogue management

The fairy-tale game (FTW) system addresses the problem of managing conversational speech with animated characters that reside in a 3D-world. Few such systems have been built; the one which most closely resembles the fairy-tale game is the Mission Rehearsal Exercise (MRE) system from the USC Institute of Creative Technologies (Swartout et al. 2004). The MRE system has more complex interaction between animated characters than the fairy-tale game system, and uses a sophisticated model for emotion (Traum et al. 2004). On the other hand, the MRE domain is a military one with a more codified way of speaking, whereas we aim to handle more spontaneous speech. To the best of our knowledge, the fairy-tale game system and the MRE system are to date the only two dialogue systems in existence that allow for multi-party dialogue, with all its additional complexity (Traum 2004).

Another problem addressed by the fairy-tale game system is the handling of asynchronous, multimodal input. Here "asynchronous" means that a strict turn-taking scheme, where speakers proceed in alternation, need not be upheld. In particular, this means that the user can make several dialogue contributions in sequence, without needing to wait for the system's reply. It also means that not only user utterances can trigger reactions from the system, but a fairy-tale character can also be triggered to speak as a reaction to events in the environment (e.g., that some other character says or does something). Existing asynchronous dialogue systems mostly work in robot domains (see, e.g., Rayner et al. 2000, Lemon et al. 2001, Sandewall et al. 2005), where stimuli from the sensors of the robot trigger utterances from the spoken dialogue interface.

We are not aware of any system comparable to the HCA system in its handling of fully mixed-initiative conversation in multiple semi-open domains. As for the FTW system, one of the closest related systems is the MRE system mentioned above. Methodologically, to start developing the HCA system, we first developed a theory of social conversation based on story-telling which was tested in the user test of the first HCA system prototype [Bernsen and Dybkjær 2004]. This process included development of coding schemes and metrics for measuring new properties of spoken conversations, such as relative target group success [Bernsen 2004] or symmetry of knowledge and initiative among the interlocutors (the user and HCA), and began our ongoing work on discourse coherence in conversation with machines and on conversation success metrics. The user test of the first HCA system prototype confirmed that we were on the right track with the theory of social conversation, which meant that we could start work on the second HCA prototype by focusing on dramatically improving the flexibility of the system's conversational abilities, arriving at the second HCA system prototype's capability of fully mixed-initiative conversation.

An interesting point of methodology wrt. the HCA system is the following. In the summer of 2003, the detailed first HCA system prototype specification was Wizard of Oz-simulated to users during two weeks at the HCA museum in Odense, Denmark. Somewhat unexpectedly to us, the data gathered on this occasion proved to be of continuing interest during the remainder of the project, for the following reason. Even though the wizards mostly followed the detailed system specification when responding to the users' input, they did so in a thoroughly human way, i.e., by normally understanding the user's full communicative intention in context in each turn. The resulting, highly flexible and context-sensitive wizard responses continue to provide a "gold standard" for how to achieve a conversation style for the machine which is as close as possible to that of human-human communication. Thus, the museum data turned out to provide us with several basic conversational structures which have been implemented in the second HCA system prototype, such as "slow" users who spent up to three conversation turns to achieve what could easily have been achieved in a single turn, and clear evidence that, in general, the simulated HCA virtually always "followed the user" when the user changed topic or domain of conversation. In other words, some of the most important conversation management results produced in the Wizard of Oz simulation in the HCA museum did not come from the wizards' strict adherence to the conversation specification but from aspects of their specifically human style of conversation which had not been anticipated in the specification.

4.2.3 Voice Creation for Conversational Embodied Characters

The fairytale characters have to be lifelike and believable in their roles in the FTW game which means that they need to be provided with natural-sounding voices conveying distinct personalities. The characters furthermore have to be able to engage in clarification dialogues that makes the number of possible system utterances prohibitively large. For these reasons, a limited domain unit selection synthesis system was developed, based on the following design criteria: initially covering conversational spoken Swedish, feasible to add new languages, high quality sounding with personal speaking style, easy to build new voices, easy to integrate as a system component in a dialogue system, faster than real-time speech generation, prosodic control (emphasis and speaking rate), voice design using acoustic signal processing (giant/gnome), equipped with fillers and filler words with different prosodic traits that make them useful for attitudinal feedback as well as for turn management, and with extralinguistic sounds (inhalations, coughs and laughter) that are used to indicate emotional and attitudal state, methods to generate lip-synch track as well as coordinating body gestures and facial gestures with verbal output.

Several other systems exist that make it possible to build natural sounding synthetic voices in limited domains, e.g., Festival/FestVox (Black 1998, Black & Lenzo 2000b). Limited-domain synthesizers (Black & Lenzo 2000a) have been developed for a number of applications, including task-oriented dialogue systems in the travel domain (Rudnicky et al 2000) as well as for animated characters in a military training domain (Johnson et al 2002). There are also a number of multimodal speech synthesizers that allow for rendering of lip-synchronized facial animation and speech synthesis (Cohen and Massaro 1993, Beskow, 2004, Pelachaud et al. 1994, Cassell et al. 2001). In the development of these synthesizers, however, focus has mainly been on the multimodal, facial and lip-synchrony-related aspects, and only to a limited extent on voice quality, personality and naturalness of the voices.

The faster-than-real-time automatic generation of conversational spoken Swedish coordinated with facial as well as bodily animations for several ECAs with distinct personalities brings together state-of-the-art results from several disciplines. It represents a breakthrough not only because, to the best of our knowledge, this is the first time all these features have been combined in one and the same system. It does so also because the results were integrated in a partially supervised system, which made it possible to immediately test different aspects of naturalness, conversational abilities and percieved personality traits. Such tests have been performed within the project, resulting in a reasonably large corpus of linguistic and other interactional data obtained from children and adolescents using the system. The children also filled in questionnaires and answered questions in deep interviews about their perception of the different characters' personality and conversational abilities, verifying that the young users did in fact perceive the characters' difference in personality and also rated the understanding capabilities of the characters to be different. Several types of behaviour, previously unseen in human–machine multimodal conversational data, were also observed in this unique material.

5 European added value

The wide range of complementary competences necessary to construct systems, such as the NICE FTW and HCA systems, is difficult to find in a single European country, especially in small countries like Sweden and Denmark. In addition, as we all know, finding, or having, some particular competence does not always imply the ability to fit smoothly into an academic-industrial consortium, such as our NICE consortium. Thus, ScanSoft finds that the NICE consortium had a good mix of partners from industry and universities. Doing joint work with partners from different European countries helps getting a better view on market conditions and possible future cooperations at European level. Since ScanSoft sells Europe-wide (world-wide), joining into other EU cooperations will be attractive for ScanSoft in the future. For ScanSoft as a company working on speech recognition, international projects like NICE have the additional appeal of dealing with different languages within a single project. So, if research innovations show improvements within one language, results can be verified in other languages. Liquid Media finds that shared ideas and a shared workload have given the project a good momentum and although we at Liquid are used to a more straightforward approach to development, the cooperation took our work into new areas previously unconsidered by Liquid. For LIMSI, the different challenges to gesture input processing arising from the HCA system and the FTW system, respectively, forced them to arrive at a suitable level of generality of approach. Finally, for NISLab as coordinating site, it has been a pleasure to coordinate an EU project in which the contractual deliverables actually did come in, and generally in good quality, if not always on the exact date and time originally planned, then again almost never with delays which put unexpected burdens on other partners.

6 Outlook

This chapter provides a short description of how the results and achievements of the project have benefited each partner and how the partners intend to use and exploit these further.

6.1 LIMSI

In the NICE project, LIMSI has developed new competence about the management of multimodal input in conversational applications with ECAs (embodied conversational agents). LIMSI intend to continue to use the specifications and software modules (GR / GI / IF), and the annotation methodology for other research projects on multimodal interfaces. The 2D Lea cartoon-like agents developed by LIMSI for early testing at the beginning of the NICE project have been used by other members of LIMSI. LIMSI is also setting up cooperation at a national level via a national working group on Embodied Conversational Agents which has been funded in France and is co-organised by LIMSI. Protocols for experimental evaluation of ECAs are being adapted for the evaluation of emotional expressivity in ECAs.

6.2 Liquid Media

Liquid have developed an advanced skeletal animation system with multiple concurrent and combinable animation tracks, hierarchical sorting, high level controls for nonverbal gestures and real time lip sync. All of these new functionalities have been implemented in a commercial game engine.

Also, the inner workings of speech have opened up a new market for us and possibly for the gaming industry as well. Future projects with several partners are under discussion.

6.3 NISLab

As originators of the NICE project, i.e., the ones who kicked off the small rocks that became the avalanche of the NICE project, we at NISLab have gained valuable experience in several key areas of natural interactive systems. First and foremost, towards the end of the project and far from fully realised in the second prototype of the HCA system, we believe to have made a major advance in general-purpose management of systems with more than a single input modality, several different output states, and full in-principle natural interactive 3D animated character output. This conversation manager is capable of managing both domain-oriented, semi-open domain, multiple-domain conversation and more traditional task-oriented natural interaction. Secondly, and again not fully represented in the second HCA prototype system, we have gained valuable experience with the challenges involved in building ontology-based natural interactive systems in which natural language understanding, conversation management, and response generation share a common ontology. In addition, we have: gained valuable experience in the difficult discipline of speech recogniser optimisation; investigated and built a system with emotions; come to realise the enormous challenges involved in full-scale conversational speech and gesture semantic input fusion; and gotten a solid dose of experience in 3D character animation.

We are keen to move on to exploring the perspectives just described in future projects and collaborations, and have been invited to join several consortia in which we could have the opportunity to do this in contexts, such as mobile phone technologies, cultural heritage technologies, and technologies enabling the character to see and sense emotions. One of these consortia includes Liquid Media as well. In addition, we are interested in exploring commercialisation opportunities for some of the results achieved in the NICE project.

6.4 ScanSoft

The NICE project had very ambitious goals, setting requirements for several of the system components that exceeded the state-of-the-art at the start of the project. Nevertheless, the project team managed to create working prototypes for both the English and the Swedish version. In the course of doing system integration and testing, the functionality of the complete system design and implementation could be verified. For future systems of similar complexity, this made the NICE prototypes a very helpful test bed in terms of overall system design and component requirements.

From the speech recognition perspective, the main challenges originate from the fact that (mostly) children speakers talk to the system in a conversational style, and that the amount of speech data available to train the ASR engine is very limited. For the English version, an additional complication was introduced by having a large proportion of non-native speakers.

Children are currently not in the focus of commercial ASR applications, so there are few publicly available speech databases for training the NICE ASR engine. Consequently, ScanSoft mainly had to rely on the speech data collected within the NICE project. On the other hand, ScanSoft was very interested to participate in the project in order to get access to such data collections and reuse them in commercial applications.

On the technological field, the project further helped ScanSoft to develop and validate new technology aimed at

- improving speech recognition for children speech in general
- dealing with ASR applications under tight training data constraints
- investigating ways to improve speech recognition for conversational speech with a high degree of noise and speaker hesitations

The new technology is already partly included in the latest versions of the ScanSoft OSR speech recognition product. Other ASR research results generated in the course of the NICE project will be integrated into future OSR product versions.

Continuing the investigations of the NICE project, TeliaSonera and ScanSoft intend to set up a bilateral project, doing further research on ASR systems for children, and for conversational and multimodal applications.

6.5 TeliaSonera

TeliaSonera has gained knowledge in particular in system architecture and distributed system engineering, natural language understanding, dialogue management, computer games technology, and the design and implementation of entertaining interfaces and applications. All this knowledge will be put to use in future internal TeliaSonera projects.

The collected data will eventually lead to better speech recognition for Swedish in general, and for Swedish children in particular. This improved technology might find a direct use in future commercial services provided by TeliaSonera. Continuing the investigations of the NICE project, TeliaSonera and ScanSoft intend to set up a bilateral project, doing further research on ASR systems for children, and for conversational and multimodal applications.

7 Conclusions

NICE has successfully demonstrated their software prototypes for multimodal conversational edutainment systems. The cooperation with all project partners was not only constructive, but also enjoyable. ScanSoft regards NICE as an interesting and successful project and wants to thank all consortium partners for their cooperation. Special thanks to NISLab for their good project management.

So, summarising this final NICE report, the consortium partners agree that each and every one of them have benefited from the work done in the project to an extent that we could only dream of at the start of the NICE project.

8 References

Aust, H., Oerder, M., Seide F. and Steinbiss, V. (1995) The Philips automatic train timetable system. *Speech communication* 17, pp. 249-262.

Avaya, W. C., Dahl, D., Johnston, M., Pieraccini, R. and Ragget, D. (2004). EMMA: Extensible MultiModal Annotation markup language. *W3C Working Draft* 14 December 2004., W3C. http://www.w3.org/TR/emma/

Barnett, J. (2005) Multimodal Architecture and Interfaces. W3C Working Draft. 22 April 2005.

Bernsen, N. O.: Measuring relative target user group success in spoken conversation for edutainment. In J.-C. Martin et al. (Eds.): *Proceedings of the LREC 2004 Satellite Workshop on Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces.* Lisbon, Portugal, 25.5.04. Paris: European Language Resources Association (ELRA), 17-20.

Bernsen, N. O. and Dybkjær, L.: Evaluation of Spoken Multimodal Conversation. *Proceedings of The Sixth International Conference on Multimodal Interaction, ICMI 2004,* Pennsylvania, USA, October 2004a. New York: Association for Computing Machinery (ACM) 2004, 38-45.

Beskow, J.(2003). "Talking Heads - Models and Applications for Multimodal Speech Synthesis", PhD thesis

Black A., Taylor P., and Caley R., "The Festival speech synthesis system," http://festvox.org/festival. 1998.

Black, A. and Lenzo, K. "Limited Domain Synthesis", ICSLP2000, Beijing, China, 2000a.

Black, A. and Lenzo, K. "Building Voices in the Festival Speech Synthesis System," http://festvox.org/bsv, 2000b.

Boye, J., Wirén, M., Rayner, M., Lewin, I., Carter, D. and Becket, R. (1999) Language processing strategies and mixed-initiative dialogues. *Proc IJCAI workshop on knowledge and reasoning in practical dialogue systems*, Stockholm, Sweden.

Boye, J. and Wirén, M. (2003a) Robust parsing of utterances in negotiative dialogue. *Proc. Eurospeech*, Geneva, Switzerland.

Boye, J. and Wirén, M. (2003b) Negotiative spoken-dialogue interfaces to databases. *Proc. Diabruck* (*7th workshop on the semantics and pragmatics of dialogue*), Wallerfangen, Germany.

Buisine, S. and Martin, J.-C. (2005). Children's and Adults' Multimodal Interaction with 2D Conversational Agents. CHI'2005 Portland, Oregon, 2-7 April.

Buisine, S., Martin, J.-C. and Bernsen, N. O. (2005). Children's Gesture and Speech in Conversation with 3D Characters. HCI International 2005 Las Vegas, USA, 22-27 July 2005.

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (Eds.): Embodied conversational agents. Cambridge, MS: MIT Press 2000.

Cassell, J., Vilhjálmsson, H. & Bickmore, T. (2001) "BEAT: The Behavior Expression Animation Toolkit". Proceedings of SIGGRAPH '01, Los Angeles, CA, 2001, pp. 477–486.

Charniak, E. (2000). A Maximum-Entropy-Inspired Parser. Proc. NAACL (North American Chapter of the Association for Computational Linguistics).

Chee, Y. M., Magaña, J.-A., Franke, K., Froumentin, M., Russell, G., Madhvanath, S., Seni, G., Tremblay, C. and Yaeger, L. (2004). Ink Markup Language. *W3C Working Draft*. 2004. http://www.w3.org/TR/InkML/

Cohen, M. & Massaro, D. (1993) "Modeling coarticulation in synthetic visual speech". In: Thalmann, N. M. & Thalmann, D., (eds.), Models and Techniques in Computer Animation, Springer–Verlag, Tokyo, pp. 139–156.

Collins, M. (1999). Head-Driven Statistical Models for Natural Language Parsing. Ph.D. Dissertation, University of Pennsylvania.

Gerosa and Giuliani, "Investigation automatic recognition of non-native children's speech", ICSLP 2004

Giuliani and Gerosa, "Investigating recognition of children's speech", ICASSP 2003

Gorin, A. L., Riccardi G. and Wright, J. H. (1997). How May I Help You? *Speech Communication* Vol. 23, pp. 113–127.

Gustafson and Sjölander, "Voice transformations for improving children's speech recognition in a publicly available dialogue system", ICSLP 2002

Gustafson, J and Sjölander, K (2004) "Voice creation for conversational fairy-tale characters", Proceedings of the 5th ISCA Speech Synthesis Workshop, Carnegie Mellon University 14-16 juni 2004.

Hagen et al., “Children’s speech recognition with application to interactive books and tutors“, ASRU 2003.

Jackson, E., Appelt, D., Bear, J., Moore, R. and Podlozny, A. (1991) A template matcher for robust NL interpretation. *Proc. DARPA speech and natural language workshop*, Morgan Kaufmann.

Johnson, W., Narayanan, S., Whitney, R. Das, R., Bulut, M. and LaBore, C. “Limited Domain Synthesis of Expressive Military Speech for Animated Characters,” In Proceedings of the IEEE TTS Workshop, 2002.

Kaiser, E., Olwal, A., McGee, D. and Benko, H., Corradini, A., Li, X., Cohen, P., Feiner, S. (2003). Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. Fifth International Conference on Multimodal Interfaces (ICMI'03) Vancouver, British Columbia, Canada, ACM Press. http://www1.cs.columbia.edu/~aolwal/projects/maven/maven.pdf

Kasper, W., Kiefer, B., Krieger, H., Rupp, C. and Worm, K. (1999) Charting the depth of robust speech processing, *Proc. ACL.*

Kumar, S., Cohen, P. R. and Coulston, R. (2004). Multimodal Interaction under Exerted Conditions in a Natural Field Setting. Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI 2004), Pennsylvania, USA.

Lemon, O., Bracy, A., Gruenstein, A. and Peters, S. (2001) Information states in a multi-modal dialogue system for human-robot conversation. *Proc. Bi-Dialog, 5th Workshop on Formal Semantics and Pragmatics of Dialogue*, pages 57 – 67.

Li and Russell, “An analysis of the causes of increased error rates in children speech recognition”, ICSLP 2002

Martin, J.-C., den Os, E., Kuhnlein, P., Boves, L., Paggio, P. and Catizone, R. (2004). Workshop "Multimodal Corpora: Models Of Human Behaviour For The Specification And Evaluation Of Multimodal Input And Output Interfaces". In Association with the *4th International Conference On Language Resources And Evaluation LREC2004* http://www.lrec-conf.org/lrec2004/index.php Centro Cultural de Belem, LISBON, Portugal, 25th may.

Milota, A. D. (2004). Modality Fusion For Graphic Design Applications. Proceedings of the Sixth International Conference on Multimodal Interfaces (ICMI 2004), Pennsylvania, USA.

Milward, D. and Knight, S. (2001) Improving on phrase spotting for spoken dialogue systems. *Proc WISP.*

Narajanan and Potamianos, “Creating conversational interfaces for children”, IEEE Transactions on Speech and Audio processing, vol. 10 No. 2, February 2002

Nivre, J. and Scholz, M. (2004). Deterministic dependency parsing of English text. *Proc. COLING 2004*, Geneva, Switzerland.

van Noord, G., Bouma, G., Koeling, R. and Nederhof, M-J. (1999) Robust grammatical analysis for spoken dialogue systems. *Journal of natural language engineering,* 5(1), pp. 45-93.

Pelachaud, C. & Prevost, S. (1994) “Sight and Sound: Generating Facial Expressions and Spoken Intonation from Context”. Proceedings of the second ESCA Workshop on Speech Synthesis, New Paltz, NY, USA, September 1994, pp. 216–219.

Pfeiffer, T. and Latoschik, M. E. (2004). "Resolving Object References in Multimodal Dialogues for Immersive Virtual Environments." Proceedings of the IEEE VR2004.

Potamianos et al., “Automatic speech recognition for children”, Eurospeech 1997

Potamianos, „Robust recognition for children’s speech“, IEEE Transactions on Speech and Audio processing, vol. 11 No. 6, November 2003

Rayner M., Hockey B.A. and James, F. (2000) A compact architecture for dialogue management based on scripts and meta-outputs. *Proc. Applied Natural Language Processing (ANLP)*

Rudnicky, A., Bennett, C., Black, A., Chotomongcol, A., Lenzo, K., Oh, A., Singh, R. “Task and domain specific modelling in the Carnegie Mellon Communicator system,” Proceedings of ICSLP, 2000..

Ruttkay, Z. and Pelachaud, C. (2004). From Brows to Trust - Evaluating Embodied Conversational Agents, Kluwer.

Sandewall, E., Linblom, H. and Husberg, B. (2005) Integration of live video in a system for natural language dialog with a robot. *Proc Dialor* (*9th workshop on the semantics and pragmatics of dialogue*), Nancy, France.

Swartout, W., Gratch, J., Hill, R., Hovy, E., Marsella, S., Rickel, J. and Traum, D. (2004) Toward virtual humans. Working notes of the AAAI Fall symposium on Achieving Human-Level Intelligence through Integrated Systems and Research.

Traum, D. (2004) Issues in multi-party dialogues, in Dignum (ed.) *Advances in Agent Communication*, pp. 201-211, Lecture notes in artificial intelligence 2922, Springer-Verlag.

Traum, D., Marsella, S. and Gratch, J. (2004) Emotion and dialogue in the MRE virtual humans. Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems (Kloster, Irsee, June 2004).

Ward, W (1989) Understanding spontaneous speech. *Proc. DARPA speech and natural language workshop,* pp. 137-141, Philadelphia, USA.

Xiao, B., Girand, C., Oviatt, S.L. "Multimodal Integration Patterns in Children," in Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP'2002), ed. by J. Hansen & B. Pellom, Casual Prod. Ltd.: Denver, CO, Sept. 2002, 629-632.

Xiao, B., Lunsford, R., Coulston, R., Wesson, M. and Oviatt, S. L. (2003). Modeling multimodal integration patterns and performance in seniors: Toward adaptive processing of individual differences. *Proceedings of the International Conference on Multimodal Interfaces* (ICMI’2003), Vancouver, ACM Press.

9 Annex

9.1 Deliverables

Deliverable No	Deliverable title	Nature
D1.1	Requirements and design specification for domain information, personality information and dialogue behaviour for the first and second prototype	Report
D1.2	Formal representation and coding of domain information, personality information and dialogue behaviour for the first and second prototype	Report + Software
D2.1	Speech data collection system aimed at collecting multimodal input data (speech and gesture)	Software
D2.2	Collection and analysis of multimodal speech and gesture data in an edutainment application	Report
D3.1	Trained acoustic models for Swedish recogniser	Report + Software
D3.2	English recogniser and improved Swedish recogniser	Report + Software
D3.3	Analysis and specification of cooperation between input modalities and cooperation between output modalities	Report
D3.4	Gesture interpretation module, first and second prototype	Report + Software
D3.5-1	Natural language understanding module: Swedish	Report + Software
D3.5-2	Two natural language understanding modules: Swedish and English (month 32)	Report + Software
D3.6	Multimodal input understanding module for the first and second prototype	Report + Software
D3.7	Multimodal output generation module for the first and second prototypes	Report + Software
D4.1	The beta version of the virtual world and a number of embodied agents	Report + Software
D4.2	First and second version of the system prototype, inhabited by characters from H.C. Andersen's world	Report + Software
D5.1	First prototype version of dialogue management and response planning	Report + Software
D5.2	Second prototype version of dialogue management and response planning	Report + Software
D6.1	Version 1 and 2 of “wrappers” for system modules	Report + Software
D6.2	Integrated prototype 1	Software
D6.3	Integrated prototype 2	Software
D7.1	Evaluation criteria and evaluation plan	Report
D7.1 Addendum	Addendum to NICE deliverable D71: Evaluation criteria and evaluation plan	Report
D7.2	Evaluation of the first and second NICE prototype versions	Report + CD-ROM
D8.1	Establishment and maintenance of the NICE website and establishment of project communication infrastructure.	

D8.2	As a minimum, six workshops as evenly spaced as possible. The first workshop will be due month 1. Responsible: a different NICE partner for each workshop (iterative).	
D8.3	Dissemination and use plan	Report
D8.4	Technology implementation plan	Report
D9.1	First annual progress report	Report
D9.2	Second annual progress report	Report
D9.3	Final report	Report
NICE data resources policy	NICE data resources policy	Report
Extra	Speech recognition and synthesis, natural language understanding robustness	Report

Table 1. The NICE deliverables

9.2 References

Author(s)	Title	Published
Abrilian, S., Martin, J.-C. & Buisine, S.	Algorithms for controlling cooperation between output modalities in 2D Embodied Conversational Agents	Proceedings of the Fifth International Conference on Multimodal Interfaces (ICMI'2003), Vancouver, British Columbia, Canada, ACM Press, 5-7 November 2003, 293-296.
Bell, L.	Linguistic adaptations in spoken human-computer dialogues - Empirical studies of user behavior	PhD thesis, dept Speech, Music and Hearing, KTH, Stockholm, 2003.
Bell, L. and Gustafson, J.	Child and Adult Speaker Adaptation during Error Resolution in a Publicly Available Spoken Dialogue System	Proceedings of Eurospeech 03, Geneva, Switzerland, 2003.
Bell, L., Gustafson, J. and Heldner, M.	Prosodic adaptation in human-computer interaction	Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 03), Barcelona, Spain, 2003.
Bernsen, N. O.	When H. C. Andersen is not talking back.	In Rist, T., Aylet, R., Ballin, D. and Rickel, J. (Eds.): Proceedings of the Fourth International Working Conference on Intelligent Virtual Agents (IVA'2003), Kloster Irsee, Germany, 15-17 September 2003. Berlin: Springer Verlag 2003, 27-30.
Bernsen, N. O.	Measuring relative target user group success in spoken conversation for edutainment	Proceedings of the LREC 2004 Workshop on Multimodal Corpora. Lisbon, Portugal, 25 May 2004, 17-20.
Bernsen, N. O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D., and Mehta, M.	First prototype of conversational H.C. Andersen	Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI 2004), Gallipoli (Lecce), Italy, 25-28 May 2004, 458-461.
Bernsen, N. O., Charfuelàn, M., Corradini, A., Dybkjær, L., Hansen, T., Kiilerich, S., Kolodnytsky, M., Kupkin, D.,	Conversational H. C. Andersen	First prototype description. Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems (ADS04), Kloster Irsee, Germany, 14-16 June 2004, 305-308.

and Mehta, M.		
Bernsen, N. O. and Dybkjær, L.	Domain-oriented conversation with H.C. Andersen	Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems (ADS04), Kloster Irsee, Germany, 14-16 June 2004, 142-153.
Bernsen, N. O. and Dybkjær, L.	Managing domain-oriented spoken conversation	Proceedings of the Third International Conference on Autonomous Agents and Multi-Agent Systems (AMAAS) Conference Workshop on Embodied Conversational Agents: Balanced Perception and Action, New York, 20 July 2004.
Bernsen, N. O. and Dybkjær, L.	Structured interview-based evaluation of spoken multimodal conversation with H.C. Andersen	Proceedings of The International Conference for Spoken Language Processing, ICSLP 2004, South Korea, October 2004. International Speech Communication Association (ISCA) 2004, Vol. 1, 277-280.
Bernsen, N. O. and Dybkjær, L.	Evaluation of Spoken Multimodal Conversation	Proceedings of The Sixth International Conference on Multimodal Interaction, ICMI 2004, Pennsylvania, USA, October 2004a. New York: Association for Computing Machinery (ACM) 2004, 38-45.
Bernsen, N. O., Dybkjær, L., and Kiilerich, S.	Evaluating conversation with Hans Christian Andersen	Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 25-28 May 2004, 1011-1014.
Boye, J. and Wirén, M.	Robust parsing of utterances in negotiative dialogue	Proceedings Eurospeech, Geneva, Switzerland, 2003.
Boye, J. and Wirén, M.	Negotiative spoken-dialogue interfaces to databases	Proceedings Diabruck (7th workshop on the semantics and pragmatics of dialogue), Wallerfangen, Germany, 2003.
Boye, J., Wiren, M., and Gustafson, J.	Contextual Reasoning in Multimodal Dialogue Systems: Two Case Studies	Proceedings of The 8th Workshop on the Semantics and Pragmatics of Dialogue Catalogue'04, Barcelona, 19-21 July 2004.
Buisine, S., Abrilian, S., Martin, J.C.	Evaluation of Multimodal Behaviour of Embodied Agents	In "From Brows to Trust. Evaluating Embodied Conversational Agents". Edited by Zsófia Ruttkay & Catherine Pelachaud. Book Series: Human-Computer Interaction Series: Volume 7, Kluwer. ISBN 1-4020-2729-X, August 2004.
Buisine, S., Abrilian, S., Martin, J.C.	Evaluation of Individual Multimodal Behavior of 2D Embodied Agents in Presentation Tasks	Proceedings of the Workshop Embodied Conversational Characters as Individuals, Melbourne, Australia, in conjunction with the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems, 15 July 2003.
Buisine, S., Abrilian, S., Rendu, C., Martin, J.-C.	Towards Experimental Specification and Evaluation of Lifelike Multimodal Behavior	In: Marriot, A., Pelachaud, C., Rist, T., Ruttkay, S., Vilhjalmsson, H. (Eds.): Proceedings of the Workshop on Embodied conversational agents - let's specify and evaluate them!, held in conjunction with The First International Joint Conference on Autonomous Agents & Multi-Agent Systems, Bologna, Italy, July 2002.
Buisine, S., Martin, J.C.	Experimental evaluation of bi-directional multimodal interaction with conversational	Proceedings of INTERACT'2003, Zurich, Switzerland. 2003, 168-175.

	agents	
Buisine, S., & Martin, J.-C.	(In press) Children's and Adults' Multimodal Interaction with 2D Conversational Agents	To appear in Proceedings of CHI'2005, 2-7 April 2005.
Buisine, S., Martin, J.-C., & Bernsen, N.O.	(In press). Children's Gesture and Speech in Conversation with 3D Characters	To appear in Proceedings of HCI International 2005, 22-27 July 2005.
Corradini, A., Fredriksson, M., Mehta, M., Königsmann, J., Bernsen, N. O. and Johanneson, L.	Towards believable behavior generation for embodied conversational agents	Proceedings of the Workshop on Interactive Visualisation and Interaction Technologies, IV&IT 2004, Krakow, Poland, June 7-9, 2004, in conjunction with the International Conference on Computational Science 2004 (ICCS 2004).
Corradini, A., Mehta, M., Bernsen, N. O., and Martin, J.-C.	Multimodal input fusion in human-computer interaction	To appear in Proceedings of the NATO-ASI Conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management. Yerevan, Armenia, 18-29 August 2003.
Gustafson, J., Bell, L., Boye, J., Lindström, A. and Wiren, M.	The NICE Fairy-tale Game System	Proceedings of SIGdial 04, Boston, 30 April-1 May 2004.
Gustafson, J, Sjölander, K.	Voice creation for conversational fairy-tale characters	Proceedings of the 5th ISCA Speech Synthesis Workshop, Carnegie Mellon University 14-16 June 2004.
Gustafson, J, Sjölander, K.	Voice Transformations For Improving Children's Speech Recognition In A Publicly Available Dialogue System	In Proceedings of ICSLP02, Colorado, USA, 2002.
Lindström, A.	English and Other Foreign Linguistic Elements in Spoken Swedish	PhD thesis no 887, Linköping University, Sweden, 2004.
Martin, J.-C., Buisine, S., Abrilian, S.	2D Gestural and Multimodal Behavior of Users Interacting with Embodied Agents	Proceedings of the workshop Embodied Conversational Agents: Balanced Perception and Action, Edited by Catherine Pelachaud, Zsófia Ruttkay & Kris Thorisson. Held during the Third International Joint Conference on Autonomous Agents & Multi Agent Systems, New York, USA, 19-23 July 2004, 34-41.
Reeves, L.M., Lai, J., Larson, J.A., Oviatt, S., Balaji, T.S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J.C., McTear, M., Raman, T.V., Stanney, K.M., Su, H., Wang, Q.Y.	Guidelines for multimodal user interface design	Communications of the ACM – Special Issue on Multimodal Interfaces, 2004, vol. 47(1), 57-59.

Table 2. The NICE references.

9.3 Print and electronic material illustrating the results

D6.3: Videos demonstrating the integrated prototype 2 of the HCA system (NISLab) and the fairy tale world system (TeliaSonera) have been sent to the Commission on 29 April 2005. The videos will be made available on the NICE website in early May, 2005.