The Swedish NICE Corpus – Spoken dialogues between children and embodied characters in a computer game scenario

Linda Bell, Johan Boye, Joakim Gustafson, Mattias Heldner, Anders Lindström, Mats Wirén

Voice Technologies, R&D division, TeliaSonera Sweden
{linda.bell|johan.boye|joakim.gustafson|mattias.heldner|anders.x.lindstrom|mats.wiren}@teliasonera.com

Abstract

This article describes the collection and analysis of a Swedish database of spontaneous and unconstrained children–machine dialogues. The Swedish NICE corpus consists of spoken dialogues between children aged 8 to 15 and embodied fairy-tale characters in a computer game scenario. Compared to previously collected corpora of children's computer-directed speech, the Swedish NICE corpus contains extended interactions, including three-party conversation, in which the young users used spoken dialogue as the primary means of progression in the game.

1. Introduction

During the past few years, some computer games where voice commands provide the primary means of control have been developed. In Lifeline, released in 2004, simple spoken dialogue commands can be used to navigate and direct the actions of the main character. Introducing more advanced spoken dialogue into computer games poses tremendous research challenges. Since the primary user group is children and adolescents, the state-of-the-art of understanding spontaneous conversational children's speech has to be advanced considerably. This involves research on the basic technologies including speech and gesture recognition, natural language understanding and dialogue management. There is also a need to develop know-how and technology that can equip the embodied conversational agents appearing in such games with appropriate behavior in every given dialogue situation. For instance, methods for the dynamic generation of verbal as well as non-verbal communicative behavior need to be developed, which puts high and partially novel demands on spoken language generation, the modularity and flexibility of the character animation system, and the synchronized, real-time control of the two. Not least importantly, knowledge is needed regarding how children adapt their speech when interacting with computers and animated characters, and, finally, there is a general need for a better understanding of how spoken dialogue can be incorporated into games in a useful and entertaining way.

The EU project NICE has attempted to address several of these issues. One of the results of the NICE-project is a corpus of spontaneous child–computer dialogue data in Swedish, which can be used to pursue the above-mentioned research goals. The aim of this paper is to describe this corpus, and to present some first observations.

The corpus was collected using a semi-automated version of the NICE fairy-tale game system [1], allowing users to interact with life-like conversational characters in a fairy-tale world inspired by the Danish author H. C. Andersen, using speech and 2D-gestures on the screen. The fairy-tale characters in the game move about in an interactive 3D environment, and possess rudimentary dialogue skills. The game is of a problem-solving nature, involving information-seeking utterances, commands, simple negotiation, but also social dialogue. The game features two main characters; Cloddy Hans and Karen. Cloddy Hans is a friendly 'helper' character who follows and guides the user throughout the course of the game. Karen is a sullen 'gatekeeper' who guards a drawbridge which the user must cross, and who has to be persuaded to let the user pass. The introduction of several interactive fairy-tale characters with distinct personalities was assumed to increase the feeling of interactivity and pace, and even allow for three-party dialogue, and thus increase the level of engagement and the game's entertainment value [2]. It was also a way to possibly engage the users in a conflict since Cloddy Hans and Karen do not like each other. If the users did take sides, it would also be interesting to see in which way this might influence the users' dialogue behavior.

2. Corpora of children's speech

What distinguishes the Swedish NICE fairy-tale corpus from previous corpora of recorded children's speech is that it contains computer-directed, spontaneous dialogue data. Several previously collected corpora consist of prompted speech and monologues where children recount stories, e.g. the American English corpora *KIDS* [3], and *CU Kids' Audio Speech Corpus* [4], and the British English, German, Italian and Swedish corpora collected within the EU project *PF_STAR* [5-9].

As concerns dialogue data, Batliner et al. [7] describe a data collection where children engaged in spoken interaction with a robot AIBO dog. The purpose of the experiment was to elicit spontaneous emotional speech by using one test condition in which the AIBO was 'disobedient' and disregarded the children's commands. However, since the AIBO did not answer back, the children's utterances mostly consisted of short commands and little dialogue interaction took place. Oviatt and Adams [10] describe a corpus where children between the ages of 6 and 10 interacted with either adults or an embodied Wizard-of-Oz interface with animated marine animals. The children's computer-directed speech was found to be less disfluent, more hyperarticulated, clearer and more repetitive. The authors report that about one-third of all content involved social interaction with the embodied agents. Narayanan and Potamianos [11] allowed children to play an interactive computer game using voice commands or keyboard and mouse in a Wizard-of-Oz scenario. The resulting corpus was used to create novel language models and understanding strategies for dialogue systems aimed towards young users. The authors found that user experience was improved by adding 'personality' to the interface, allowing for multimodal interaction and using animated sequences to convey information [11].

In a study using the same database, it was shown that younger children use less overt politeness markers and verbalize their frustration more than older children do [12].

3. The NICE fairy-tale game scenario

The initial scene of the game was designed as a sort of grounding game with the purpose of allowing the user to get acquainted with Cloddy Hans and learn how to interact with him and the physical environment displayed on the screen [1]. The user meets Cloddy Hans in H. C. Andersen's study, where the fairy-tale machine normally used by Andersen to construct new stories is situated. There is also a shelf in the study filled with various fairy-tale objects (gems, a sword, poison flasks etc.) that have to be put in one of several icon-labeled slots in the fairy-tale machine in order to construct a new story and thereby get transferred into the fairy-tale world, where the second scene takes place. The user can talk to Cloddy Hans and use a mouse for pointing and making gestures, but cannot directly manipulate the objects. Instead, she needs to agree with Cloddy Hans on what the different objects can be used for and how to refer to them, so that she may ask Cloddy Hans to put the objects in the appropriate slots. In the second scene, Cloddy Hans and the user find themselves on a rather small island, along with all the objects they previously chose to put in the fairy-tale machine. The island is separated from the mainland by a drawbridge, guarded by Karen, who has deliberately been designed to differ from Cloddy Hans in terms of personality, as conveyed by both her verbal and non-verbal behavior. Karen will only lower the drawbridge when offered something she finds acceptable in return, which she never does until the user's third attempt, thereby encouraging negotiative behavior. Furthermore, both Cloddy Hans and Karen openly show some amount of grudge against each other, with both characters occasionally prompting the user to choose sides.

4. Data collection using the NICE system

During 2004–2005, data was collected on several occasions using the NICE system at different stages during its development. The system could be run either in fully automatic mode or in supervised mode, in which a human operator had the possibility to intervene and replace or modify the output of system components. This made it possible to develop the system in a data-driven, iterative fashion, by initially gathering data in partially supervised mode and by running several cycles of data collection, data analysis and corresponding system development.

Four sub-corpora were collected over a period of 5 months. The recording conditions are described in Table 1 and the sub-corpora will be labeled "School", "Lab 1", "Lab 2" and "Lab 3" in the rest of this paper. During this period a fair amount of changes to the system took place, including adding the second scene in which Karen appears, as well as considerably improving the system's spoken language understanding capabilities. Thus, the four sub-corpora consist of data collected from heterogeneous user groups under differing conditions during several stages of the development of the NICE system (cf. Table 1). Speech data was collected when users were interacting with the system, as well as during a post-session interview. All subjects were recorded using a close-talking head-mounted wireless microphone, and subjects in sub-corpora Lab 1–3 were also recorded on video. Data from all major sub-components of the NICE system was also logged. Prior to the interaction, each user was given a short instruction and was also asked to fill out a questionnaire, recording demographic data and self-estimates of computer and video game use. The instructions were deliberately sparse–the users were told that they would be testing a research prototype of a new kind of computer game, where they would be able to talk to fairy-tale characters adopted from H. C. Andersen's stories. Following the interaction with the system the subjects were interviewed about their experiences with the game and the characters involved in it. After this, the subjects were given a second questionnaire assessing various aspects of the game as well as properties of the characters involved in it. This questionnaire used 5-point Likert scales [13], with which even the youngest subjects were familiar through the use of such instruments in school.

Some data was discarded for reasons such as drop-outs or failure in logging one or more of the involved modalities (cf. Table 1). All remaining speech was automatically segmented using the speech detection algorithm of a commercially available speech recognizer for Swedish, yielding close to six hours of spoken language data of which approximately two thirds were computer-directed speech. This material was orthographically transcribed, with special symbols employed to denote disfluencies, non-speech sounds etc. and analyzed in search of interesting interaction phenomena.

Table 1: Recording conditions for the four different sub-corpora

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

5. Findings

5.1. Corpus statistics

The total number of user sound files in the human–computer dialogue corpus was 5,580. This material was tagged in terms of utterance types, the distribution and individual variation in use of these utterance types is shown in Table 2.

Table 2: Distribution of utterance types and individual variation in use of utterance types

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

Utterance fragments were identified and joined into turns, following which the number of turns for each interlocutor was calculated. The database obtained in this way contains 5,583 Cloddy Hans turns, 255 Karen turns and 5,144 user turns. The average number of turns per user was 90, with individual variation ranging from 26 to 210 turns.

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

5.2. Interview results

The interviews were centered around the following questions:

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

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

5.3. Gameplay and personalities

Judging from the interviews, the game seems generally to have been perceived as fun, interesting and non-irritating even by users who found it difficult. This is supported by the results of the questionnaire (cf. Table 3).

Table 3: Median scores for questions about the game play in the questionnaire across all four sub corpora

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

In the interviews, users unanimously reported that Cloddy Hans was a bit slow, but kind, while Karen being rather the opposite. Non-communicative as well as verbal and non-verbal behavior of the two characters Cloddy Hans and Karen had been designed to convey differences in personality along several dimensions in the so-called OCEAN model [2, 14]. Analyses of data obtained from the post-experiment questionnaires showed that the two characters were indeed perceived as having different personalities in several respects. Table 4 shows which of the two characters displayed each trait in the most salient way, as judged by the users in Lab 2 and 3, who all interacted with both Karen and Cloddy Hans.

Table 4: User judgments regarding which animated character displayed specific personality traits in the most salient way, based on questionnaire data from Lab 2 and 3. Differences between Cloddy Hans and Karen were tested for significance using Wilcoxon Signed Ranks Test (p<0.05).

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

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

5.4. Dialogue phenomena

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

- either taking Karen's or Cloddy Hans's part when one of them offended the other,

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

Dialogue excerpts exemplifying some of these dialogue behaviors are shown below. The excerpt starts in a situation where the user is trying to persuade Karen to let the user pass over the bridge.

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

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

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

6. Discussion

In this paper we have described a new corpus of multimodal spontaneous child–computer dialogue in Swedish which was collected while users were interacting with several embodied conversational agents, sometimes engaging in three-way dialogue, in a computer game where spoken and multimodal dialogue constituted the primary means of progression. Users found the game to be fun and spoken dialogue to be a natural part of the game. Deliberate differences in the persona design of the animated characters and the introduction of plot elements requiring negotiation seems to have resulted in high degrees of naturalness, spontaneity and engagement on the users' part (as shown by examples). The corpus as well as the system used for data collection will be useful tools for research on technologies required for accommodating children and adolescent users in future multimodal dialogue systems.

7. Acknowledgements

This work was carried out within the EU-funded project NICE (IST-2001-3529, http://www.niceproject.com).

8. References

[1] Gustafson, J., Bell, L., Boye, J., Lindström, A., and Wirén, M., "The NICE Fairy-tale Game System," in *Proc. 5th SIGdial Workshop on Discourse and Dialogue*. Cambridge, MA: NAACL, 2004.

[2] Gustafson, J., Boye, J., Fredriksson, M., Johannesson, L., and Königsmann, J., "Providing computer game characters with conversational abilities," in *Proc.of Intelligent Virtual Agent (IVA05)*. Greece, forthcoming.

[3] Eskenazi, M., "KIDS: A database of children's speech," *Journal of the Acoustical Society of America*, vol. 100, 1996.

[4] Hagen, A., Pellom, B., and Cole, R., "Children's speech recognition with application to interactive books and tutors," in *Proc. IEEE ASRU Workshop*, 2003.

[5] D'Arcy, S. M., Wong, L. P., and Russell, M. J., "Recognition of read and spontaneous children's speech using two new corpora," in *Proc. ICSLP*, 2004.

[6] Giuliani, D. and Gerosa, M., "Investigating recognition of children's speech," in *Proc. ICASSP*, 2003, pp. 137-140.

[7] Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S. M., Russell, M. J., and Wong, M., "'You stupid tin box' - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," in *Proc. LREC*. Lisbon, 2004.

[8] Blomberg, M. and Elenius, D., "Collection and recognition of children's speech in the PF-Star project," in *Proc. Fonetik 2003*. Umeå, 2003, pp. 81-84.

[9] Gerosa, M. and Giuliani, D., "Investigating automatic recognition of non-native children's speech," in *Proc. ICSLP*, 2004, pp. 1521-1524.

[10] Oviatt, S. and Adams, B., "Designing and evaluating conversational interfaces with animated characters," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds. Cambridge, MA: MIT Press, 2000, pp. 319-343.

[11] Narayanan, S. and Potamianos, A., "Creating conversational interfaces for children," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 65-78, 2002.

[12] Arunachalam, S., Gould, D., Andersen, E., Byrd, D., and Narayanan, S. S., "Politeness and frustration language in child-machine interactions," in *Proc. Europeech*, 2001, pp. 2675-2678.

[13] Likert, R., "A Technique for the Measurement of Attitudes," *Archives of Psychology*, vol. 140, pp. 1-55, 1932.

[14] McCrae, R. and Costa, P., "Toward a new generation of personality theories: Theoretical contexts for the five-factor model," in *The five-factor model of personality: Theoretical perspectives*, J. S. Wiggins, Ed. New York: Guilford, 1996.