H. C. Andersen Conversation Corpus

Niels Ole Bernsen, Laila Dybkjær and Svend Kiilerich

Natural Interactive Systems Laboratory
University of Southern Denmark
Campusvej 55, 5230 Odense M, Denmark
{nob, laila, kiil}@nis.sdu.dk

Abstract

This paper describes the design, collection and current status of the Hans Christian Andersen (HCA) conversation corpus. The corpus consists of five separate corpora and represents transcription and annotation of some 57 hours of English spoken and deictic gesture user-system interaction recorded mainly with children 2002-2005. The corpora were collected as part of the development and evaluation process of two consecutive research prototypes. The set-up used to collect each corpus is described as well as our use of each corpus in system development. We describe the annotation of each corpus and briefly present various uses we have made of the corpora so far. The HCA corpus was made publicly available at http://www.niceproject.com/data/ in March 2006.

1. Introduction

The complete Hans Christian Andersen (HCA) conversation corpus described in this paper was made available at http://www.niceproject.com/data/ by end of February 2006. The corpus includes five different corpora representing transcription of approximately 57 hours of English spoken user-system interaction recorded 2002-2005. Common to the corpora is that they present orthographically transcribed and transcription-tagged data from conversations between, primarily, children and youngsters and 3D life-like animated fairytale author HCA. Otherwise, the corpora are different in several respects, reflecting their different purposes in the development process.

In this paper, we briefly describe the HCA system and the role of the corpora in its development, followed by a detailed characterisation of the HCA corpus. We conclude by describing ongoing corpus work.

2. The HCA System

The HCA system was developed in the EU NICE project on Natural Interactive Communication for Edutainment, http://www.niceproject.com. A key project goal was to explore the feasibility for embodied conversational characters of carrying out domain-oriented conversation as opposed to standard task-oriented dialogue. Thus, with the HCA system, the 10-18 years old target users do not have any particular task(s) to solve during interaction. Rather, they are invited to have conversation with revived HCA about his fairytales, his life, his person, including his visual appearance on-screen, his study in which he receives his visitors (Figure 2.1), and the user. Jointly and separately, these domains of conversation are huge, and there was no way in which we could have provided virtual HCA with anything like the knowledge he once had. Thus, HCA tells his visitors on occasion that he is still far from knowing everything he once knew, and that users can ask him more about what he actually remembers.

The conversation with HCA is fully mixed-initiative and visitors may also use 2D deictic gesture when talking to HCA about objects in his study.

3. The HCA Corpus

The HCA corpus consists of two early corpora and three user test corpora, cf. Table 3.1. Major goals in corpus development were to collect:

Figure 2.1. HCA in his study (second prototype).

- substantial acoustic speech data for tuning project partner Scansoft's UK English speech recogniser with children's speech;
- data for speech recogniser vocabulary and language model development;
- data on how users express themselves for natural language understanding module development;
- deictic gesture and combined speech/gesture data;
- data on what the users want to talk to HCA about.

The corpora served various other purposes as well and were developed at different stages in the development life-cycle. In the following we provide more details on the five corpora, the setups used for collecting them, and their annotations (Section 4).

3.1. Early Wizard of Oz Corpora

The first two corpora were developed using Wizard of Oz (WoZ) simulation.

3.1.1. First Wizard of Oz Corpus

The purpose of the WOZ1 corpus was to explore children's conversation with HCA. WOZ1 demonstrated that Danish kids were quite happy talking to HCA in English and curious to explore his knowledge and opinions.

The corpus was collected at four schools in HCA's hometown, Odense, and (in one case) at NISLab. Therefore the setup had to be easily transportable by two persons (wizard and assistant) and also fairly quick to set up on the spot.

	WoZ1	WoZ2	WoZ3 (PT1)	PT2	PT2 (English)
Purpose	recogniser training system specification	recogniser training design specification test	recogniser training PT1 evaluation	full system test PT2 evaluation	full system test with native speakers PT2 evaluation
When collected	1-31 October 2002	21-30 July 2003	20-22 January 2004	15-17 February 2005	30 March / 13 April 2005
Where collected	local schools	HCA Museum, Odense	NISLab	NISLab	NISLab
WoZ type	controlled in-field	in-field	controlled laboratory	controlled laboratory	controlled laboratory
Users	7-19 year olds	all ages	10-18 year olds	11-16 year olds	10-14 year olds
Number of users	60-70	approx. 500	18	13	4
User input	speech	speech	speech and gesture	speech and gesture	speech and gesture
System output	speech, small animated HCA face, text reminders	speech, modestly animated HCA in his study	speech, gesture, lip-sync	speech, gesture, lip-sync, facial expression	speech, gesture, lip-sync, facial expression
Implemented system parts involved	none	none	all except speech recogniser	all	all
Role of wizard(s)	Act as HCA	Act as HCA	Type spoken input	none	none
Assistant(s)	One acting as experimenter and technician	Student helper in museum, inviting young visitors to talk to HCA	Experimenter, two typing wizards, technician, observers, two interviewers	Experimenter, two system operators, calibrator, technician, observer, two interviewers	Experimenter, two instructors / observers / interviewers, system operator / calibrator / technician
Raw data collected	audio: user and wizard no logfiles	audio: user and wizard no logfiles	audio: user and wizard video incl. audio: interaction complete logfiles	audio: user video incl. audio: interaction complete logfiles	audio: user video incl. audio: interaction complete logfiles
Speech data volume	approx. 6 hours	approx. 30 hours	approx. 11 hours	approx. 8 hours	approx. 2 hours
Video data volume	N/A	N/A	approx. 11 hours	approx. 8 hours	approx. 2 hours
No. of conversations	35-40	502	36	26	6
No. of user utterances	2.047	6.870	1.206	1.101	276
Raw transcription	yes	yes	yes	yes	yes
Transcription coding	yes	yes	yes	yes	yes
Transcription tool	Transcriber	Transcriber	Transcriber	Transcriber	Transcriber
Semantics annotation	no	yes approx. 50% of data	no	no	no
Topic annotation	yes	yes approx. 70% of data	no	no	no
User language skill evaluation	yes	yes	no	no	no

Table 3.1. Overview of the five corpora constituting the HCA corpus.

Users spoke through a microphone placed on the table and listened to HCA responses via headphones, cf. Figure 3.2. Headphones were used instead of loudspeakers to avoid the wizard's output being recorded together with the user's input since we wanted to have the users' voice material separately. The wizard spoke through a microphone and listened to the user via speakers. To render the voice of the wizard as neutral and computer-

like as possible, the interface Sound Blaster Live! was used for distortion along with the Sound Blaster card. The wizard was responsible for controlling the mouth, eye and head movements of a cartoon-like HCA face, which could be seen by the user via a mini-hub connection.

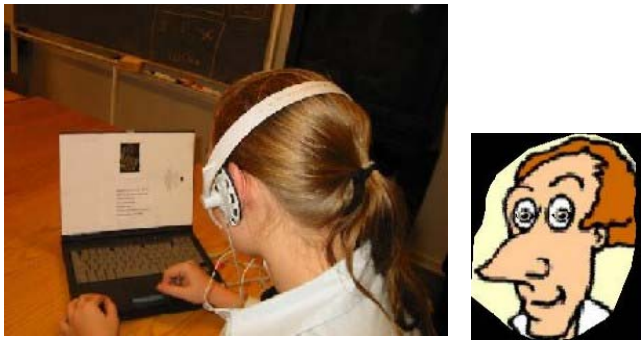

Figure 3.2. WOZ1 user (left) and (right) the HCA the user saw on the screen.

The user's screen was covered with paper and through a hole HCA was visible (Figure 3.2). Below HCA a list of topics was written (e.g., childhood, youth, and travels). There were no restrictions on what the user could talk about. However, the list was meant to inspire the user in case s/he couldn't think of anything to talk about. Figure 3.3 shows a conversation snippet.

user	why do you begin to read histories
wizard	well at first i wanted to be an actor i tried to get into the theatre but they didn't want me so instead i started to write
user	where do you live the where do you live that you was little where you was little
wizard	my childhood i spent in Odense i lived in Odense 'till i was fourteen years old then moved to Copenhagen
user	where do you live in Copenhagen

Figure 3.3. A dialogue snippet from WOZ1.

3.1.2. Second Wizard of Oz Corpus

Partly based on WOZ1, we developed a more detailed design specification for the first system prototype, including HCA's output in his domains of knowledge and discourse. The specification was simulated in a 10 days, in-field WOZ2 exercise at the HCA Museum in Odense (Figure 3.4). It was judged that a single wizard at-a-time could do the simulation without the need for an assistant. However, since the simulations would go on all day from 9 AM to 7 PM for 10 days in succession, two wizards took turns talking to the museum visitors.

The wizard was wirelessly connected to the laptop which showed a primitive HCA mock-up (Figure 3.5) and navigated HCA's designed output in hypertext while speaking to the visitors using voice distortion. In addition to following the specification, the wizard made several planned improvisations, for instance to explore users' interest in topics outside the specification, such as modern technical inventions which was a great interest of HCA's.

In the museum, posters were put up inviting visitor to talk to "a nice person". Moreover, an assistant invited visitors to try the system, answered questions and helped the users put on the microphone headset, ensuring as much as possible that the microphone had the right distance to the mouth of the user. Then the assistant instructed the users to say hello to the system in order to start the conversation. Figure 3.6 shows a snippet from a conversation.

Figure 3.4. User talking to HCA in the museum.

Figure 3.5. HCA on-screen in WOZ2. He could only move his head and lips.

user	i haven't heard much about the Little Mermaid could you please tell me 'bout it
wizard	yes the mermaid saves a prince from drowning and falls in love with him. to become human and get legs she pays with her beautiful voice. because she is now dumb she cannot tell the prince who she is and he marries another causing the mermaid to die
user	okay a
wizard	do do you like the story
user	yea i like i've watched a c cartoon about it a ca a cartoon movie
wizard	what is a cartoon movie
user	well they use drawings and they use a lot o' them and then they make this whole movie about it

Figure 3.6. Conversation snippet from WOZ2.

3.2. User Test Corpora

Three more corpora all derive from user tests with different versions of the implemented system.

3.2.1. Third Wizard of Oz (PT1) Corpus

WOZ2 provided early confirmation that the target users appreciated HCA's story-telling approach. The first prototype, PT1, was developed following a modified version of the WOZ2 output design specification. Our third corpus, WOZ3, derives from the user test of PT1. WOZ3 is a "bionic" WOZ corpus in the sense that, when testing PT1, we still had wizards simulating the speech recogniser whereas the rest of the system was running.

Users came in parallel. Thus, two offices had been prepared for user testing. In one room, the user had both a mouse and an 19" touch screen for producing 2D gesture input. In addition, the user could watch HCA on a 42" flat-panel screen (Figure 3.7). In the second room, the user had a mouse and an ordinary 19" screen. All users wore a headset for oral communication. The two wizards (one per user) were placed in a third room. Their job was to type in what the users said, thus replacing the speech recogniser that was not yet integrated with the rest of the system.

Figure 3.7. Controlled laboratory test of PT1.

user	can you tell me about one of your adventures fairy tales
HCA	I was only thirty-one years old when I wrote The Little Mermaid.
user	tell me about one of your fairy tales
HCA	I like several of my fairy tales very much. One of my favourites is the ugly duckling. # The ugly duckling is not accepted by the other ducks and animals because it looks different. It has a hard time of suffering and troubles. # Eventually it wishes only to die till it discovers that it is in fact a beautiful swan. # The morale of the fairy tale is that if you are different then you will get bullied.
user	tell me about the little mermaid tell me about the little mermaid
HCA	Did I already tell you that only the bible has been translated into more languages than my fairy tales.

Figure 3.8. A conversation snippet from WOZ3 (PT1).

Each user test session had a duration of 60-75 minutes. A session included conversation with HCA in two different conditions followed by a post-test interview. In the first condition, the users only received basic instructions on how to operate the system, i.e. speak using the headset, control HCA's movements, control the four virtual camera angles, and gesture using mouse or touch screen. After 15 minutes, the first session was interrupted and the user received a set of 13 typed scenario problems to be solved through speech or gesture input in the second session, such as "Find out if HCA has a preferred fairy tale and what it is" and "Tell HCA about games you like or know". The problems might be addressed in any order and the user was not necessarily expected to carry out all of them. The purpose was to ensure a significant amount of user initiative to explore how the system would respond under the resulting user initiative pressure. Figure 3.8 shows a snippet from the WOZ3 (PT1) corpus.

3.2.2. PT2 Corpus

Following analysis of WOZ3 (see, e.g., [Bernsen and Dybkjær 2004]) and substantial system re-design, not least with the aim of obtaining far more flexible conversation, the second prototype, PT2, was built and tested with users in February 2005. Although the user test conditions were very similar to those in the PT1 test, the resulting corpus is different from WOZ3 because it shows the effects of (i) incorporating a speech recogniser and (ii) strongly improving the system's conversational abilities.

Each user used a microphone/loudspeaker headset, a touch screen for gesture input, and a keyboard for controlling multiple virtual camera angles and for controlling HCA's locomotion. Two cameras captured the user's behaviour during interaction. A developer sat behind the user, monitoring the workings of the system modules, making observations, being ready to re-start a hanging or crashed system component, help users in case of difficulties with the equipment, etc. In about half of the user tests, an observer was present in the test room as well. Figure 3.9 shows a user interacting with PT2 HCA while Figure 3.10 shows an excerpt from the PT2 corpus.

Figure 3.9. Controlled laboratory test of PT2.

3.2.3. PT2 Corpus with native English speakers

The final, fifth corpus in the HCA corpus was also made with PT2. Basically the same set-up was used as in the first PT2 test. The main difference from the first PT2 corpus is that the second test involved English native speakers instead of children with Danish as their mother tongue. Figure 3.11 shows a conversation snippet from the second PT2 corpus.

HCA	Do you want to hear other stories about my life?
user	no thank you you are very old
HCA	sorry
user	you are very old
HCA	I am not old! As you can see I am a man in my best age. I know many things about the world and I can do many things or at least I used to. I still have much hair on my head and I dont wear a wig. I even had my hair curled lately and my beard shaven off completely. In spite of that many people think that I look strange I know that big feet and all. I will not speak of my teeth so don't you dare to ask
user	you are very old
HCA	No no I am not old at all. I am only fifty five years old that is nothing compared to two hundred years. Well I was born two hundred years ago. So I am wondering how I can be fifty five especially since I was born again just a few months ago. I wish somebody would tell me! Maybe I will never die again?
user	you are ugly

Figure 3.10. A dialogue snippet from PT2.

user	how was your mother
HCA	The name of my mother was Anne Marie Andersdatter. Oh my poor mother! She was a washerwoman. She had her washing place beside the Odense river. It was hard work not even well paid and cold too especially in the winter! She had to warm herself quite often with a drop of aquavit that became her destiny! My dear mother died when I was twenty eight.
user	who is this person
HCA	In the picture you see Riborg Voigt. I fell in love with her but she did not fall in love with me. She had the most beautiful brown eyes. Actually I wrote a poem about them: Two brown eyes I have espied. Therein my home and my world reside. There dwelleth talent and childlike peace. The memory in me will never cease.
user	wow
HCA	thanks at first you go through a lot of troubles and then you get famous

Figure 3.11. A snippet from the PT2 corpus with native English speakers.

4. Corpus annotation

All five corpora were transcribed after their collection and in some cases also further annotated as shown in Table 3.1. Table 4.1 provides an overview of the transcription tags used in each corpus.

For WOZ1, two students were used as transcribers and for topic tagging. The transcribers were instructed orally and also received two documents, one containing instructions and guidelines for transcription and one with the topics to use. A supervisor was responsible for checking the work of the students and for answering all questions arising. Spell-checking was made after the dialogues had been transcribed and topic tagged.

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

For WOZ2, students transcribed the segmented audio files. The transcription files were spell-checked using Microsoft Word. The speech recognition partner again wanted markup of the speaker's level of English inserted into the annotation files. The rules used in WOZ1 were reused in an elaborated version. Four levels of expertise, i.e. bad, medium, good, and native, were distinguished. The annotation was done by a phonetician. However, the idea behind the rules in the elaborated version is that it should be possible to arrive at the same evaluation results also if the rules are applied by non-phoneticians.

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

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

The PT1 corpus and the two PT2 corpora have all been transcribed using the tags shown in Table 4.1. Students transcribed the PT1 corpus with sporadic checks by NISLab staff. PT2 was transcribed by NISLab staff and not checked by others. Both PT1 and PT2 were spell-checked. Annotation rules were extended and got increasingly explicit based on lessons learned from transcription flaws made during transcription of the previous corpora.

5. Corpus Exploitation

So far, we have investigated the HCA corpus to research user behaviour, develop various metrics, and test-drive theory development.

Based on the large WOZ2 corpus, Bernsen [2004] presents metrics for measuring the extent to which a system actually succeeds in targeting a particular user group, such as the 10-18 year olds targeted by the HCA system, rather than other user groups, in this case the under-10 year olds and the over-18 year olds.

Bernsen et al. [2004] analyse the top-ten conversations in terms of number of turns in the WOZ2 corpus. The average number of turns is 109. The metrics developed and applied all concern symmetry in conversation as manifested by (1) symmetry in presenting expertise in domains of common interest, (2) symmetry in taking initiative to change the domain/topic of conversation, and (3) symmetry in being an active contributor in driving the conversation forward. The assumptions are that (1) through (3) are key properties of successful prototypical

human-human conversation and that *enabling* the HCA system to support these properties is an important and difficult goal for "real" conversational systems as opposed to task-oriented spoken dialogue systems.

Martin et al. [to appear] present results from analysing deictic gesture and combined speech and deictic gesture input in PT2.

Ongoing work addresses (1) quantitative metrics conversation robustness based on PT2, (2) quantitative metrics for conversation success based on PT2, and (3) coding scheme and theory for spoken, deictic gesture, and combined speech-gesture reference to virtual scenes. This latter work is based on PT2 English.

Tags	**WoZ1**	**WoZ2**	**WoZ3 (PT1)**	**PT2**	**PT2 English**
Unfilled pause (one second or more)	•	•	•	•	•
Background noise (produced by the speaker or by something else, e.g., coughs, telephone, music, or people speaking)	•	•	•	•	•
Spelled input	•	•			
Restart (false starts, restarts, self-repairs and repetitions)	•	•	•	•	•
Directives (part of the experiment, occurring in-between actual wizard-user dialogues, where testing is performed or instructions are given)	•	•	•	•	•
Filled pause (a filled pause consists of discourse-oriented noise produced by the speaker, e.g. hmm or ahh)	•	•	•	•	•
External event (intrusive event which influences the speaker, but which does not necessarily include a noise)	•	•	•	•	•
Unknown word	•	•	•	•	•
Mispronounced word (this includes word fragments as well as full words which are mispronounced)	•	•	•	•	•
Grammatical error (e.g. how many books did you wrote?)	•	•			
Background speech (indicating a conversation between the user and another person, also when this other person is speaking to the user without the user answering)		•	•	•	•
Fade-out signal		•	•	•	•
Long phoneme (when a phoneme is longer than normally, e.g. moooove)		•			
Begin laughing / end laughing (used to delimit a segment of words pronounced while the user is laughing. The begin tag is put before the beginning of the segment and the end one is put after its end)		•	•	•	•

Table 4.1. Overview of transcription tags used in the five corpora constituting the HCA corpus.

6. Conclusions

We have described the collection of the HCA conversation corpus consisting of five different corpora and representing transcription and annotation of some 57 hours of English spoken and deictic gesture user-system interaction. Most users are children and youngsters. The set-up used to collect each corpus has been described as well as our use of each corpus in system development. We have also described the annotation (transcription and otherwise) of each corpus and briefly presented various uses we have made of the corpora so far. If you are interested in obtaining the HCA corpus, please fill the registration form at http://www.niceproject.com/data/.

Acknowledgement

We gratefully acknowledge the NICE project support by EU's Human Language Technologies programme, contract IST-2001-35293. We would also like to thank Michel Généreux who contributed to the development of the WOZ1 corpus, and Thomas K. Hansen who analysed users' English proficiency in WOZ1 and WOZ2.

References

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. 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, Penn State University, USA, 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'2004), Vol. III, Lisbon, Portugal, May 2004, 1011-1014.

Martin, J. C., Buisine, S., Pitel, G., and Bernsen, N. O.: Fusion of Children's Speech and 2D Gestures when Conversing with 3D Characters. Signal Processing, Special Issue on Multimodal Interfaces (to appear).