International Journal of Humanoid Robotics

Children's Turn-Taking Behavior Adaptation in Multi-Session Interactions with a Humanoid Robot

Ivana Kruijff-Korbayová, Heriberto Cuayáhuitl, Bernd Kiefer

German Research Center for Artificial Intelligence (DFKI), Stuhsatzenhausweg 3, 66440 Saarbrücken, Germany
ivana.kruijff@dfki.de, heriberto.cuayahuitl@dfki.de, bernd.kiefer@dfki.de

Ilaria Baroni, Alberto Sanna

Fondazione Centro San Raffaele via Olgettina 60, 20132, Milan, Italy
baroni.ilaria@hsr.it, sanna.alberto@hsr.it

Marco Nalin

Telbios S.p.A. via Olgettina 58, 20132, Milan, Italy
marco.nalin@telbios.com

Received March 25, 2013
Revised Day Month Year
Accepted Day Month Year

If we want to advance human-robot interaction towards long-term sustainability, we need to understand how users' perceptions and responses to robot behavior develop across multiple sessions. To this end, we set up a study in which children interacted face-to-face in three sessions on different days with a humanoid robot that engaged in a quiz, dance or imitation activity, in one of two conditions: the robot either gave explicit verbal and non-verbal signals of being familiar with the user from previous interactions, or it did not. The robot system relied on a human wizard to interpret the user's speech and gesture input, but the rest of the system behavior was produced automatically. A preliminary analysis of a small subset of the interactions published elsewhere indicated that the children adapted various aspects of their verbal and non-verbal conversational behavior to the robot, just as humans generally adapt to their conversational interlocutors in a way that fosters the predictability, intelligibility, and efficiency of communication. We therefore carried out a follow-up systematic analysis of all quiz interactions focusing on the children's verbal turn-taking behavior. We found that communication problems such as speech overlaps and child speech ignored by the robot are decreasing across the three sessions. Moreover, these problems are fewer and decrease faster when the robot explicitly signals familiarity with the user. In this paper we present the experiment method, describe the dialogue management and verbal output production implemented in the system, and report the results of the children's turn-taking adaptation analysis.

Keywords: child-robot interaction; long-term interaction; verbal behavior adaptation; turn-taking; familiarity display; dialogue management; natural language generation

1. Introduction

As social robots are getting more commonplace, it is likely that they will interact face-to-face with humans over longer, discontinuous stretches of time. Various experiments in long-term face-to-face human-robot interaction have already been carried out.[25,23,26,29] These have for example attempted to identify factors that contribute to long-term engagement. In order to enable robots to engage in and sustain effective long-term face-to-face communication, it is of course important to understand what social competencies the robots need to have. We believe, however, that besides understanding suitable robot behavior, it is equally very important to understand how humans behave in interaction with robots and how their perception of and response to robot behavior develops over time in multiple sessions. One such aspect is adaptation of verbal behavior.

Interpersonal conversation is a dynamic adaptive exchange where an interlocutor's verbal and non-verbal signals are adjusted to the conversational partner (and the situation) in a way that fosters the predictability, intelligibility, and efficiency of communication, and also manages social impressions (cf. for example the Communication Accommodation Theory).[22,11] Since it is by now also well established that humans tend to treat computers as social actors and respond to them as they would to another person,[35] it can also be expected that humans adapt their conversational behavior to computers. And indeed, there is growing evidence that humans adapt various aspects of their verbal and non-verbal behavior to those of the computer interfaces they interact with. Concerning linguistic adaptation, for example, experiments with text-based human-computer interaction show lexical and syntactic adaptation of users to the system.[21,10,9]. Systematic work on speech signal feature adaptation of users in spoken human-computer interaction is also starting to emerge.[33] However, verbal behavior adaptation in face-to-face interaction with robots remains to be studied. Moreover, human adaptation to systems has so far been studied in one-shot, relatively short encounters. Persistence of adaptation across sessions has not been addressed.

In our work we investigate adaptation of children in face-to-face interaction with robots across multiple sessions. In a previous study we found that children adapt various aspects of their verbal and non-verbal behavior, including speech timing, speed and tone, verbal input formulation, nodding and gestures.[31]

In this article we investigate children's verbal turn-taking adaptation. We found that adaptation increases across multiple sessions. Moreover, we found that children adapt their verbal turn-taking behavior more readily when the robot gives explicit signals of familiarity with the child across sessions, for example by using their name or referring to previous experiences. The children who interact with a familiarity-displaying robot wait more with speaking, in order to avoid speech overlaps, and produce less turns that end up ignored by the robot. So overall, there is more adaptation in a condition with familiarity display in comparison to a condition where the robot's behavior is neutral in this respect. In Section 2 we present the experi-

ment method. Section 3 describes the approach to dialogue management and verbal output production implemented in the experiment system. Data analysis results are presented and discussed in Section 4, conclusions and outlook in Section 5.

This work is carried out in the larger context of the Aliz-E project.[a] The goal of Aliz-E is to develop the theory and practice behind cognitive robots capable of maintaining believable any-depth affective interactions with young users over an extended and (possibly) discontinuous period of time. Different strategies for achieving this goal (with children) are studied in the project.

2. Experiment Method

2.1. *Participants*

Participants were recruited by invitation letters sent to the members of the diabetes association connected to the San Rafaelle hospital in Milan and information brochures displayed at the hospital. Both contained a link to a website to make an appointment. The experiment took place on Saturdays in March – May 2012.

19 children participated in the experiment (Italian, 11 male, 8 female; age 5-12), but only 13 were able to participate in three sessions on different days as foreseen in the protocol. Table 1 shows demographic data of these 13 children (average age eight, SD=1.85). Although the experiment invitation was distributed through the diabetes association, more than half of the participants had no ailments.

Table 1. Demographic data distribution of the 13x children who completed 3 sessions.

Participant characteristics	Frequency
Gender	Male: 9; Female: 4
Education level	Preschool: 1; Elementary school: 11; Middle school: 1
Diabetes type I	Male: 5; Female: 1

2.2. *Procedure*

Upon arrival to their first session of the experiment, the child and the accompanying person were given an introduction comprising the following information:

General introduction We are building a robot to support hospitalized children. The child is there to test the current functionality and so help the development.

Experiment execution The child has three sessions on different days and every time it has one or more interactions with the robot. The robot is able to engage in three different activities: quiz, dance and imitation. The child

[a]http://www.aliz-e.org/

selects one of the activities as the main one, which will be performed in the first interaction in every session. If there is sufficient time for a second interaction within a session, the child can select a second activity freely in each session. There is no right or wrong behavior, the child should just behave naturally during the interactions. Questionnaires will be filled in before and after each interaction. These are important for us to understand what the child thinks of the robot.

Robot disclaimer The robot might be quite slow, make mistakes, or have technical problems (e.g., motors too hot, low battery, system failure). In case of technical problems, an adult will enter the room and manage the situation.

Activity-specific instructions A general introduction about each activity is given, to ensure the child knows what to expect and what to do.

After this introduction, a consent form is signed and pre-interaction questionnaires are filled in (demographic information, hobbies and use of technology, selection of the main activity). Then the first interaction starts, featuring the activity the child selected as main. The robot greets the child, introduces itself by name and asks for the child's name. Then it explains the activity, and asks the child whether it wants to play. The child can end the interaction at any point. At certain points during the activity (e.g., end of a phase, or a game round) the robot explicitly asks whether the child wants to continue. The interaction duration is not fixed: the child may quit the interaction, or continue as long as it wants, up to a limit of 30 minutes (unless the interaction has to be ended earlier for technical reasons). If the child continues playing for 30 minutes, the robot apologizes that it needs to end the interaction to take some rest. At the end of an interaction, the robot asks the child whether they liked to play, states that it enjoyed it and is hoping to play again, and gives the child good-bye. The child then fills in a post-interaction questionnaire for self-assessment of its engagement and relationship to the robot, and its opinions about the robot and the interaction.

Time permitting, the child can select a second activity and have another interaction with the robot, followed by filling in the post-interaction questionnaire again. The whole session is limited to one hour, including the questionnaire-filling time.

The second and third sessions take place on different days. The general introduction is not repeated, the child starts an interaction with the robot immediately with the main activity. The robot greets the child, but does not repeat the name- and activity-introductions. The rest of the process is the same as in the first session.

Table 2 shows activity selection distribution for the 13 children who completed all three sessions. About half of them were able to have two interactions in the first session (6/13). In the second and third session, most children had two interactions (11/13 and 9/13, respectively). 11 children selected quiz as their main activity. Quiz was therefore featured in more than half of all the interactions (37/65).

Table 2. Activity selection of the 13 children with 3 experiment sessions.

	Session 1		Session 2		Session 3		Total
Activity	Main	Second	Main	Second	Main	Second	
Imitation	1	-	-	6	1	4	12
Dance	1	4	2	3	1	5	16
Quiz	11	2	11	2	11	-	37
Total	13	6	13	11	13	9	65

2.3. *Activity Description*

The following activities were available (Figure 1 shows children and robot in action):

Quiz The child and the robot ask each other multiple-choice quiz questions from various domains (e.g., diabetes, nutrition, sports, geography, history, science). The asker provides correctness feedback. The asker can reveal the correct answer after two wrong attempts or upon request. The robot makes mistakes on purpose (with an answer error rate of about 30%), in order to avoid frustrating the child by a too good performance. At the end of each round the robot provides a summary of the number of correct and incorrect answers and a short evaluative comment. A round of quiz normally consists of four questions asked by the same asker, the child can however propose to switch roles at any time.

Dance The robot first explores various dance moves with the child and then teaches it the individual movements of a dance sequence chosen according to the child's abilities. After the child learns three movements, the robot plays music and they try together the movements learned so far. The robot provides encouraging feedback on the child's performance.

Imitation Either the child or the robot presents a sequence of simple arm poses (right/left arm up/down), and the other tries to memorize and imitate it. If there is a mistake, the imitator has one more attempt. Then they switch roles and the game goes on. The game starts with a one-pose sequence, and as it progresses, the sequences get longer by one pose every round. At the end of the game the robot provides a summary of the number of correct

Fig. 1. Children playing with the robot during the experiments. Left to right: quiz, dance, imitation.

and incorrect imitations and a short evaluative comment.

Besides activity-specific conversation, the interactions involve also a social component, such as greetings and introductions. When the robot provides performance feedback to the user during an activity, the social aspect requires careful handling of the evaluation process so as not to discourage the user with negative feedback. Preference is given to positive or encouraging comments on the child's performance. No comparison of the child's and the robot's performance is made in this version of the system, to avoid a focus on competition.

2.4. *Familiarity Display vs. Neutral Display Condition*

Long-term interaction involves series of encounters between the robot and a given user. As the robot interacts with a particular user, they become familiar with each other, i.e., they accumulate shared knowledge (shared history, personal common ground).[13] For example, they can know each other's name, performance on a game, ways of speaking or nonverbal behaviors. Familiarity increases over time.

In the experiment we compared two versions of the system in a between-subjects design: In the *familiarity-display* (FD) condition, the robot tries to foster a sense of persistence and familiarity. It uses verbalizations explicitly acknowledging and refering to the shared history with a given user, thus making it explicit that it is familiar with the user and remembers the previous encounters. Such verbal moves are accompanied by nonverbal behaviors showing familiarity, e.g., nodding, higher excitement. In the *neutral display* (ND) condition, the system only uses verbalizations that are neutral with respect to familiarity, i.e., they do not signal familiarity. Examples of verbalizations from both conditions are shown in Table 3.

Table 3. Examples of verbalizations that signal familiarity (used in the FD condition) or are neutral in this respect (used in the ND condition).

Familiarity display	Neutral display
Use of user's name:	
So, which answer do you choose, *Marco*?	So, which answer do you choose?
References to previous encounters and play experiences:	
I am happy to see you *again*.	I am happy to see you.
It was nice playing with you *last time*.	–
References to previous performance in an activity:	
Are you ready to play quiz *again*?	Are you ready to play quiz?
Today you were *again really good* at quiz.	Today you were really good at quiz.
Well done, you've done *better than last time*	Well done.
Reference to familiarity of a quiz question or a dance move:	
The next question should sound familiar.	The next question.
Let's try *again* this move: the spring step.	Let's try this move: the spring step.
Reference to familiarity of activity rules:	
Remember the magical pose?	Now the magical pose.

2.5. *System*

We used Nao, a humanoid robot from Aldebaran Robotics.[b] Nao is 57 cm tall, weighs 5.2 kg and its body has 25 degrees of freedom. It has a cartoon-like appearance which is considered especially suitable for use with children, although it has no capability for facial articulation.

The experiment was carried out using the human-robot interaction system developed in the ALIZ-E project. The system integrates components for speech and gesture capture and interpretation, activity and interaction management, user modeling, speech and gesture production and robot motor control (Figure 2). We use components developed within the project as well as off-the-shelf technologies such as Julius and HTK for speech recognition, OpenCV for gesture recognition, Acapela and MARY for speech synthesis, OpenCCG for language parsing and generation, Weka and JavaBayes for maintaining a probabilistic personalized user profile.[24,43,8,1,19,32,42,14] To bring all components together within a concurrent execution approach we use the Urbi middleware.[2] More details on system implementation have been published elsewhere.[27,28].

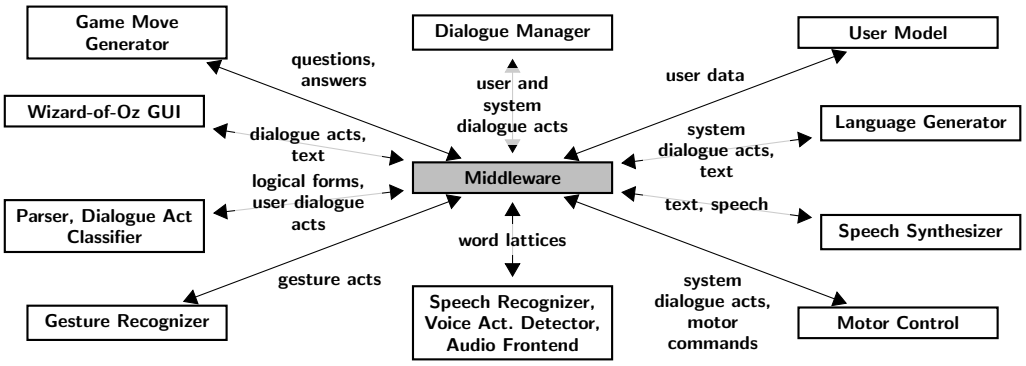

Fig. 2. High-level architecture of the integrated system.

Although the system was presented as fully automatic to the children, we relied on a human Wizard to simulate the recognition and interpretation of the user's speech and gestures. The Wizard first selects an interpretation of the user's input in a GUI (Figure 3 shows an instance of the GUI during a quiz interaction). Then the next system action is selected by the Dialogue Manager component (Section 3.1), while the Wizard has the possibility to override the automatic selection if needed. Only in several quiz interactions early in the experiment the Wizard simulated the next system action selection entirely, due to technical issues.

The dialogue act corresponding to the selected next system action is verbalized

[b] *www.aldebaran-robotics.com*

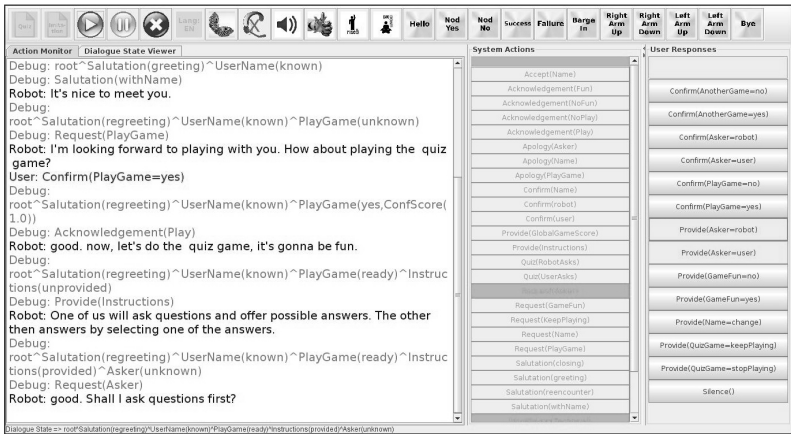

Fig. 3. The Wizard's GUI. The buttons at the top allow the Wizard to start, pause and abort the interaction, toggle automatic dialogue management, speech recognition and synthesis, and trigger arm poses for imitation as well as several communicative gestures. The large pane on the left shows a transcript of the interaction, the middle pane lists the possible next actions of the system, and the right-most pane lists possible interpretations of user input. Those options that the system expects at a given point are highlighted as suggestions for the Wizard, to make their task easier.

automatically by the natural language generation component which produces text for the speech synthesizer. The generation process is divided into the two classical steps of *utterance planning* and *lexical realization*:[36] The utterance planner takes as input the dialogue act and any additional relevant information from the DM and constructs a linguistically motivated semantic structure, which in turn serves as input to a grammar-based lexical realizer to produce the corresponding utterance verbalization. The utterance planner, which was inspired by the KPML system[5], is implemented as a general rule-based graph rewriting engine. For lexical realization we use the open source natural language processing library OpenCCG,[32] which provides parsing and realization services based on Multimodal Combinatory Categorial Grammar (CCG).[3,4] However, in the system used in the experiment we employed a *canned-text approach* to generation, where the utterance planning and lexical realization steps are combined: The utterance planner component turns the input it gets from the dialogue manager directly into one or more utterances.[c] We will talk more about the verbalization process in section 3.2, especially about verbalization variablity.

Nonverbal behavior planning and motor control are also automatic and include dance movements and imitation poses, communicative gestures assigned to specific types of dialogue acts (e.g., greetings, requests) and static key poses displaying emotions namely anger, sadness, fear, happiness, excitement and pride.[6]

[c]The reason for using the canned-text approach is development speed: We were able to include alternative verbalizations, without first having to ensure grammar coverage for their realization.

To summarize, these features describe the system used in the experiment:

- Speech and gesture recognition simulated by a Wizard
- System action selection automatic with the possibility of Wizard override
- User barge-in: Interruption of the robot's speech by an early child response
- Automatically produced verbal output in Italian with many variations and expressive speech synthesis distinguishing sad, happy and neutral state
- Automatically produced head and body poses and gestures
- Persistent user-specific interaction profile

During the experiment the robot was standing or kneeling on a table, the child is sitting (for quiz) or standing (for dance and imitation) in front of the table (Figure 1). An additional camera recorded the interaction and its video and audio signal was transmitted to the Wizard in another room.

2.6. *Collected Data*

The data collected in the experiment consists of the pre- and post-interaction questionnaires, video and audio recordings of the interactions and system logfiles.

3. Automatic System Action Selection and Verbalization

In this section we provide more details on our approach to dialogue management and verbal output production for the quiz interactions, as these components are responsible for the system behavior that is most relevant for the analysis of user adaptation presented in this paper.

3.1. *Dialogue Management*

The dialogue manager (DM) component carries the primary responsibility for controlling the robot's conversational behaviour in our system.[16] It keeps track of the interaction state, and integrates the interpretations of the user's input/actions with respect to this state. In addition, it queries and updates the game move generator and user model components, and selects the next action of the system as a transition to another state, making progress towards a goal. The next system action is selected according to a set of policies that specify a mapping from dialogue states describing situations in the interaction, to (communicative) actions. The dialogue policies are learnt offline from a simulated environment partially estimated from real interaction data.

The role of dialogue policy learning is important to optimize dialogue behaviours rather than using purely hand-coded dialogue policies. Since hand-coding dialogue behaviours for complex speech-based or multimodal systems is a daunting task, researchers have turned their attention to machine learning dialogue systems to support adaptive interactions. The reinforcement learning framework has been a

promising direction.[30] Unfortunately, applying such a framework to complex dialogue systems (i.e. systems with large state-action spaces) is not a trivial task. The need for machine dialogues that support flexible, complex, optimal and robust interactions is still a long standing problem that deserves to be further investigated. Previous work in human-robot interaction does not optimize dialogue control or optimizes it with flat learning.[7],[41] We apply the Hierarchical Reinforcement Learning (HRL) approach described below that aims to overcome some of these problems.

At the core of our framework, dialogue management is cast as a discrete Semi-Markov Decision Process (SMDP) in order to address the problem of scalable dialogue optimization. Such a discrete-time SMDP $M = <S, A, T, R, L>$ is characterized by the following elements: (a) a finite set of states S; (b) a finite set of actions A; (c) a stochastic state transition function $T(s', \tau|s, a)$ that specifies the next state s' given the current state s and action a, τ denotes the number of time-steps taken to execute action a in state s; (d) a reward function $R(s', \tau|s, a)$ that specifies the reward given to the agent for choosing action a when the environment makes a transition from state s to state s'; and (e) L is a language that provides the mechanism to express tree-based state representations. We describe L as a context-free grammar to represent formulas constructed from predicates, functions, variables, constants and connectives.[37]

We distinguish two types of actions: (a) single-step actions roughly corresponding to dialogue acts or actions such as 'greeting' or 'ask question', and (b) multi-step actions corresponding to sub-dialogues or contractions of single-step actions such as 'robot asks' or 'user asks'. We treat each multi-step action as a separate SMDP. [17],[15] In this way, an MDP can be decomposed into multiple SMDPs which are hierarchically organized into X levels and Y models per level, denoted as $\mu = \{M_j^i\}$, where $j \in \{0, ..., X-1\}$ and $i \in \{0, ..., Y-1\}$. The indexes i and j only identify a subtask (i.e. SMDP) in a unique way in the hierarchy, they do not specify the execution sequence of subtasks because that is learnt by the reinforcement learning agent. Thus, a given SMDP in the hierarchy is denoted as $M_j^i = <S_j^i, A_j^i, T_j^i, R_j^i, L_j^i>$. The solution to a Semi-Markov decision process is an optimal policy π^*, which is a mapping from environment states $s \in S$ to single- or multi-step actions $a \in A$. In other words, the goal of an SMDP is to find a function denoted as $\pi^*(s)$ that maximizes the cumulative reward of each visited state. The optimal action-value function $Q^*(s, a)$ specifies this cumulative reward for executing action a in state s and then following policy π^*. The optimal policy for each learning agent in the hierarchy is defined by $\pi_j^{*i}(s) = \arg\max_{a \in A_j^i} Q_j^{*i}(s, a)$. We use the HSMQ-Learning algorithm to induce a hierarchy of policies.[20] More recently we have extended such an algorithm with more flexible interactions by relaxing the strict hierarchical control.[16]

We use the hierarchy of dialogue agents shown in Figure 4. Table 4 shows the set of state variables for our system, each one modelled as a discrete probability distribution with predefined parameters. Dialogue and game features are included to inform the agent of situations in the interaction. Our action set consists of meaningful

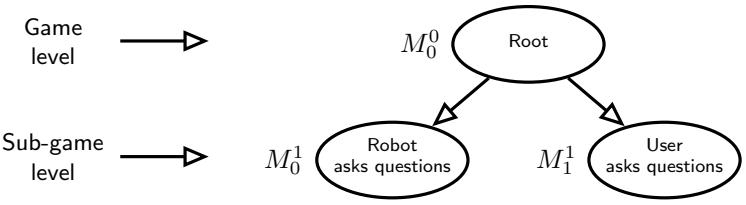

Fig. 4. Hierarchy of dialogue agents for our robot in the Quiz domain.

Table 4. State variables for modeling the quiz interactions, where combinations of variable-value pairs define situations in the interaction used by the DM for action-selection.

State Variable	Values
Salutation	none, greeting, withName, regreeting, closing
UserName	unknown, filled, known
ConfScore	null, 0.1, 0.2, 0.3, 0.4, 0.5, ... , 0.9, 1.0
Confirmed	null, no, yes
PlayGame	unknown, no, yes, ready
Instructions	unprovided, provided
Asker	unknown, robot, user
QuizGame	unplayed, playing, semiplayed, played, interrupted, keepPlaying, stopPlaying
GameFun	unknown, no, yes
GameOver	no, yes
GameInstructions	unprovided, provided
QuestionState	null, unknown, unasked, askedWithAnswers, askedWithoutAnswers, reaskedWithAnswers, reaskedWithoutAnswers, confirmed
AnswerState	unanswered, unclassified, correct, incorrect, unknown
MaxQuestions	no, yes
GameScore	unknown, good, bad
GlobalGameScore	null, unprovided, provided
ExpressedScore	no, yes

combinations of dialogue act types[d] and the associated parameters[e]. We constrained the actions per state based on the CFGs L_j^i, i.e. only a subset of sensible actions was allowed per dialogue state. While our HRL agent with tree-based states has 10^4 state-actions, a static, propositional representation (enumerating all variables and values) has 10^{12} state-action pairs. This makes the hierarchical tree-based representation scalable to larger sets of state variables and actions. The reward function addressed efficient and effective interactions by encouraging to play and get the right answers as much as possible. It is defined by the following rewards for choosing action a in state s: +10 for reaching a terminal state or answering a question

[d]Dialogue act types: Salutation, Request, Apology, Confirm, Accept, SwitchRole, Acknowledgement, Provide, Stop, Feedback. Express, Classify, Retrieve, Provide.
[e]Parameters: Greeting, Closing, Name, PlayGame, Asker, KeepPlaying, GameFun, StopPlaying, Play, NoPlay, Fun, NoFun, GameInstructions, StartGame, Question, Answers, CorrectAnswer, IncorrectAnswer, GamePerformance, Answer, Success, Failure, GlobalGameScore, ContinuePlaying.

correctly, −10 for remaining in the same state (i.e. $s_{t+1} = s_t$ or $s_{t+1} = s_{t-1}$), and 0 otherwise.[f] The DM learnt its behaviour offline by interacting with a stochastic simulated user. The simulated user acts were estimated using bigram language models $P(a^{usr}|a^{sys})$ with Witten-Bell discounting from pilot interactions. A sample dialogue together with dialogue act lables is shown in Table 5.

3.2. *Verbal Output Production*

The verbal output of the system is produced by the natural language generation (NLG) and Text-To-Speech Synthesis (TTS) components. The task of the NLG component is to produce an utterance that verbalizes the dialogue act corresponding to the next system action selected by the DM.

To avoid repetitive verbalizations, we invested considerable effort to implement a large range of verbal output variation. Selection among variants is either random or controlled by selection criteria. Some selection criteria refer to characteristics of the *content* to be conveyed, e.g., how many answer options a quiz question has and whether they are short or long. Other selection criteria refer to various parameters of the *context*, e.g., the user's gender, how many quiz questions have already been asked, who is the current asker, etc. An important selection criterion is the familiarity display condition: only when the robot is to explicitly display familiarity, verbalization variants are used that include the child's name and/or explicitly refer to the interaction history (Table 3). In this case utterance planning uses the information whether the current user interacts with the system for the first time or it is a subsequent encounter, whether they have already played the current game or it is new, whether the user's previous performance was good or not, etc.

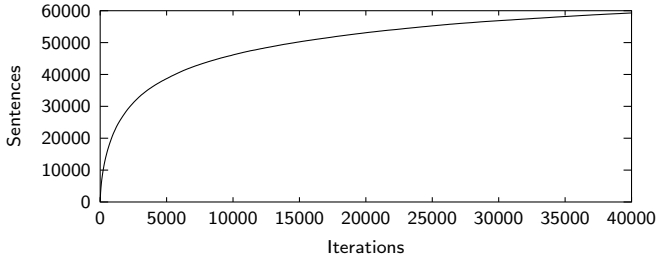

Fig. 5. Unique sentence count per iteration of the full utterance planning rule set

We assessed the output verbalization variability by letting the utterance planner run for many iterations using a test corpus of 530 input structures that covered the 59 dialogue acts implemented in the experiment system with a variety of parameter

[f] A reward function induced automatically (e.g. using correlation and linear regresssion analyses[18]) from richly-annotated multimodal interactions is left as future work.

Table 5. Sample quiz interaction (user utterances in italics). Utterance groups correspond to subtasks in the hierarchy of dialogue agents.The dialogue acts with the star symbol omit parameter values for space reasons.

Agent	Dialogue Act	Utterance
Rob	Salutation(Greeting)	Hi, my name is NAO, and you?
Usr	Provide(Name=Danilo)	*I am Danilo*
Rob	Salutation(withName, Name=Danilo)	Nice to meet you Danilo.
Rob	Request(PlayGame, Game=Quiz)	Do you want to Play a Quiz game?
Usr	Confirm(PlayGame=yes)	*Sure*
Rob	Provide(Instructions)	One of us asks questions and offers possible answers. The other answers by selecting one of those answers. Do you want to ask me questions or shall I ask you?
Usr	Provide(Asker=robot)	*You ask me*
Rob	SwitchRole(Asker=Robot)	Okay, I'll ask you.
	Retrieve(Question)+Provide(Question*)	The question is. What is the typical magic instrument?
	Provide(Answers*)	Your choices are: 1, guitar. 2, drums. 3, violin. 4, flute.
	Request(Answer)	What is your choice?
Usr	Provide(Answer=drums)	*Number two*
Rob	Feedback(IncorrectAnswer)	That is not quite right.
	Request(AnswerAgain)	Do you have another choice?
Usr	Provide(Answer=violin)	*May be number three*
Rob	Provide(CorrectAnswer=flute)	The correct answer is number four, flute.
	Retrieve(Question)+Provide(Question*)	The next question is. Which one of these is an important artificial channel of the central America?
	Provide(Answers*)	Your choices are: 1, channel five. 2, Panama canal. 3, Suez canal. 4, English channel.
	Request(Answer)	What is your choice?
Usr	Provide(Answer=Panama canal)	*Panama canal*
Rob	Feedback(CorrectAnswer)	Yes, correct.
	Retrieve(Question)+Provide(Question*)	The next one is. Which are the elements of the football goal?
	Provide(Answers*)	Your choices are: 1, poles, cross and net. 2, cross, net and arbiter. 3, poles, ball and grass.
	Request(Answer)	What is your choice?
Usr	Provide(Answer=poles, cross and net)	*The first one*
Rob	Feedback(CorrectAnswer)	Yes, correct.
Rob	Provide(GamePerformance*)	Well done, you got right two questions out of three.
Rob	SwitchRoleTo(user)	Okay, let's switch roles. Now you ask me.
Rob	Request(Question)	What is the first question.
Usr	Provide(Question*)	*The first question is: What does water become at the temperature of zero degrees?*
Rob	Request(Answers)	What are my choices?
Usr	Provide(Answers)	*Your choices are: 1, condensation. 2, ice. 3, steam.*
Rob	Provide(Answer=condensation)	I choose condensation.
Usr	Feedback(IncorrectAnswer)	*That is incorrect.*
Rob	SelfFeedback(IncorrectAnswer)	Oh no.
Rob	Request(CorrectAnswer)	So what is the right answer?
Usr	Provide(CorrectAnswer=ice)	*The correct answer is number two.*
Rob	Request(Question)	Okay, ask me another question.
Usr	Provide(Question*)	*The next question is. Through which breach the Italian army entered in Rome in 1870?*
Rob	Request(Answers)	What are my choices?
Usr	Provide(Answers)	*Your choices are: 1, porta venezia. 2, porta pia. 3, porta pazienza. 4, porta nuova.*
	Request(Answer)	*What is your choice?*
Rob	Provide(Answer=Porta pia)	Porta pia
Usr	Feedback(CorrectAnswer)	*Yes, that's right.*
	Provide(Question*)	*The next question is. What is the plural of I?*
Usr	Request(Answer)	*Your choices are: 1, they. 2, you. 3, we.*
	Request(Answer)	*What is your choice?*
Rob	Provide(Answer=we)	I think is we.
Usr	Feedback(CorrectAnswer)	*Yes, correct.*
Rob	Provide(GamePerformance*)	Not bad, I got right two questions out of three.
	Request(KeepPlaying)	Shall we keep playing?
Usr	Provide(QuizGame=stopPlaying)	*I want to stop.*
Rob	Provide(GlobalGameScore*)	In this session you got right 2 out of 3 questions.
	Request(GameFun)	Did you like playing the Quiz Game?
Usr	Provide(GameFun=yes)	*I did.*
Rob	Acknowledgement(Fun)	I am glad to hear that.
Rob	Salutation(closing)	It was nice playing with you, see you soon. Bye!

values. As Figure 5 shows, convergence was achieved with a sequence of 40,000 iterations, which produced 59,296 unique sentences (not excluding substrings).

The number of alternative verbalizations varies greatly between dialogue acts. While those dialogue acts that only appear once, such as name introduction or activity explanation, have just a single verbalization, the often occurring ones have tens to hundreds, some even thousands of variants, ensuring that the users are not being exposed to repetitive system output.

For speech synthesis we used the open source Mary TTS platform[19,38] with an Italian voice built using the Mary TTS voice creation toolkit.[39,27] In order to contextualize the system speech we implemented the following prosody modifications, using the support Mary TTS provides for controlling the prosody of HMM-based voices with symbolic markup:[34] (a) *Prosodic prominence modification (stress)*: The NLG component labels focus words.[40] The TTS component then modifies the prosodic realization by decreasing the speech rate and raising the pitch contour on the focus words. (b) *Emotional prosody modification*: The dialogue manager decides when the system output should be rendered with (non-neutral) emotional colouring, either "sad" or "happy". The TTS component then increases/decreases the speech rate and the pitch contour, respectively.

4. Analysis and results

4.1. *Engagement Analysis From Self-Assessment Questionnaires*

The post-interaction questionnaires administered to the children asked them to rate their happiness, relaxedness and amusement during the interaction, in order to study their engagement. We used self-assessment questionnaires even though we are aware that such self-reports from children often suffer from a ceiling effect, and therefore are not a very reliable tool when used with children.

A first analysis of the questionnaire answers shows the following results: (a) a smaller decline in happiness throughout the three sessions in the Familiarity Display (FD) condition than in the Neutral Display (ND) condition ($t(10)=2.70$, $p=0.02$); (b) increase in relaxation in both conditions (although this cannot be attributed only to the robot's effect, as the children were of course becoming more acquainted also with the experiment environment and the personnel); (c) no statistically significant effect or trend for amusement. Other ways of analysing the interaction data in order to assess the children's engagement are under way.

4.2. *Analysis of Children's Turn-Taking Adaptation From Video*

It was noted informally and investigated in a preliminary analysis using video data from three children, that children seem to adapt various aspects of their verbal and non-verbal behavior, including speech timing, speed and tone, verbal input formulation, nodding and gestures.[31] Following up on these preliminary results, we carried out a systematic analysis of turn-taking behavior using video data from all

children who completed three quiz interactions. In this case we were also able to study the effect of the FD vs. ND condition.

We only considered quiz interactions for the present analysis, for several reasons. One, they constitute more than half of the collected data, and there are three quiz interactions in three different sessions for each child that chose quiz as the main activity (Table 2 in Section 2). Two, the quiz data is more consistently verbal, whereas in dance and imitation, where physical movement is part of the activity, a large part of the interaction is nonverbal. Three, quiz interactions are a good starting point for studying face-to-face conversation because of their verbal character: Although both the robot and the children use nonverbal behavior in quiz, it is to accompany verbal communication, but almost never alone.

Data from N=10 children (equally distributed over FD/ND condition) were included in this analysis, a total of 9.5 hours of video material.[g] Table 6 shows their age and gender distribution.

Table 6. Age and gender distribution for the 10 children whose data were included in the analysis.

Age	Male	Female	Total
7	2	1	3
8	-	-	-
9	4	1	5
10	1	-	1
11	-	1	1
Total	7	3	10

4.2.1. *Data Coding*

The units that were coded were *child speech segments* (CSS). Any occurrence of child speech was considered a CSS. A CSS could contain silence between stretches of child speech, as long as there is no robot's speech in between. It could be a single complete utterance or a sequence of utterances (e.g., a quiz question followed by listing the answer options), but also just an utterance fragment or a short acknowledgement or feedback. It could also be a sequence of repetitions. A CSS could be the realization of one or more dialogue moves (e.g., a quiz question plus a reguest for answer, or an acknowledgment plus the next dialogue move, etc.). Since CSSs were not available in the system logs, they were identified manually by the coders.

The following attributes were coded for each CSS:

Start time The CSS onset time relative to the beginning of the quiz interation.

[g]Of the 13 children who completed three interactions 11 selected quiz as the main activity. One however experienced technical problems in the second session and we thus had to discard the data.

Timing An abstract characterization of the timing of the CSS w.r.t. the robot's speech. This attribute has three possible values:

- **Overlap** The child and the robot speak simultaneously (at some point) during the given CSS. Overlaps are coded irrespective of which interlocutor started speaking first.
- **Forced** The child clearly waits with its speech until the robot finishes speaking, or even until the robot produces a particular prompt, for example a request for the next quiz question. The child waits, even though it does not have to, since it knows what to say next, and it could barge in. Only clear cases of the child obviously delaying its speech are coded with this value.
- **Timely** The CSS comes in a timely fashion, resulting in smooth turn-taking (without an overlap or forced waiting). It might be that the child waits a little with their speech, but not obviously so.

Robot's reaction Whether the robot appears to take the CSS into account for its next action. This attribute has two possible values:

- **Ignore** The CSS has no or only a partial effect on the next action of the robot, the robot carries on with the interaction as if (a part of) the CSS did not occur. This often leads to the child repeating (part of) their speech. An example of a partial effect is when the child presents the next quiz question along with the answer options, but the robot still asks for the latter. The ignore is not decided by the Wizard, most of time it is caused by delays in the system (from the moment when the wizard sends the command to the actual execution in the robot), by the child's barge, or by the child switching to another dialogue state to which the robot is not prepared.
- **Not-ignore** The robot's next action is a coherent continuation of the interaction given the CSS. The robot either immediately responds to the CSS (e.g., answers a quiz question), or it moves on to an appropriate next step.

Alignment Whether the child's verbal behavior aligns with the robot's expectations (i.e., the implemented strategy), in other words, whether the child adheres to the foreseen interaction script. This is an attribute derived on the basis of the other two, in order to see their combined effect. It has two possible values:

- **Not-aligned** The CSS has problems either in timing (overlap) or in the robot's reaction (ignore), or both.
- **Aligned** The CSS has no overlap and is not ignored by the robot.

Table 7 summarizes the coding scheme. The 9.5 hours of quiz interaction videos were coded by two independent coders (two of the authors). Inter-annotator agreement was checked for the preliminary analysis reported elsewhere:[31] the two coders

Table 7. Coding scheme summary

Item	Label	Value
Start Time	Start Time	Time of the the CSS event
Timing	Overlap	There is overlap between CSS and robot's speech
	Forced	The child is clearly waiting for the robot to finish before speaking
	Timely	The is no overlap in the speech
Robot's reaction	Ignore	The CSS has no or only a partial effect on the next action of the robot
	Not ignore	The robot's next action is coherent with the dialogue script
Alignment	Not-aligned	The CSS had either an overlap or an ignore
	Aligned	No overlaps or ignore in by the robot

coded independently the same 36 minutes of video of the same child to identify overlap- and ignore-CSSs, and reached Cohen's κ of 0.94, indicating very good reliability. For the analysis presented in this paper the data was divided between the two coders, and merged afterwards.

4.2.2. *Results*

Since the next action selection was simulated by a Wizard in a few quiz interactions in the first three weeks of the experiment and automatic later (Section 2.5), we first checked whether the data from these interactions can be combined for analysis. We found no statistically significant differences in either interaction speed (measured in the number of CSS per minute) or in the timing and robot's reaction factors. We therefore feel justified to combine them for the analysis presented below.

Tables 8 – 11 show the distributions of the values coded in the data for the factors of timing, robot's reaction and alignment. We report the mean of each factor averaged over the 10 children, separated per session and per condition. For the analysis of the effect of growing familiarity across the three sessions and of the FD/ND condition we use two-way Analysis of variance (ANOVA), where each factor is a dependent variable and the session number and FD/ND condition are independent variables.

Timing The results clearly show that the relative number of CSSs with forced waiting is increasing over the three sessions ($F(2, 29)=5.185$, $p=0.032$), and it is increasing more in the FD condition ($F(1, 29)=4.570$, $p=0.021$). Furthermore, the children in the FD condition tend to force themselves to wait at least twice as much as children in the ND condition. Figure 6 shows the trend of the CSSs with forced waiting over the three sessions for the FD and ND condition.

The relative number of CSSs with overlaps appears to decrease across the three sessions from 14.15% to 7.63% in the FD condition, and from 19.93% to 12.82% in

Table 8. Distribution of CSS timing values across sessions and conditions (averaged over 10 children).

	CSS timing								
	Forced (%)			Timely (%)			Overlap (%)		
Session	1	2	3	1	2	3	1	2	3
FD cond.	04.17	10.98	15.90	81.68	77.50	76.47	14.15	11.52	07.63
ND cond.	00.94	05.03	07.83	79.13	79.07	79.36	19.93	15.91	12.82

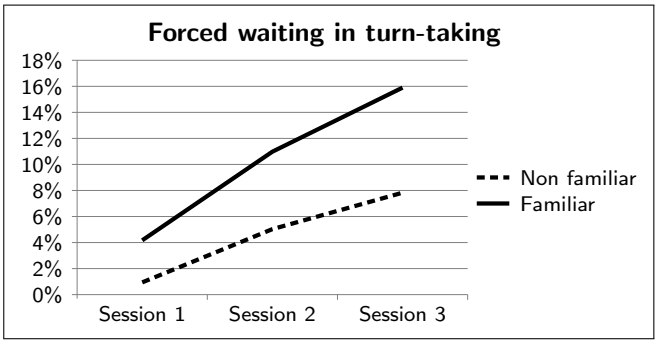

Fig. 6. Change in the relative number of CSSs with forced timing (waiting) across sessions. The x axis: session number. The y axis: the mean number of CSSs with forced waiting relative to the total number of CSSs per interaction. The two conditions are shown in different line styles.

the ND condition. While the statistical significance of this improvement between sessions is only weak ($F(2,29)=2.586$, $p=0.096$), the difference between the FD and ND condition shows higher statistical significance ($F(1, 29)=4.375$, $p=0.047$).

Robot's reaction The relative number of CSSs ignored by the robot drops across the three sessions: from 23.05% to 9.05% in the FD condition and from 28.2% to 12.89% in the ND condition. This time there is statistical significance in both the improvement across sessions ($F(1, 29)=10.608$, $p=0.001$) and the difference between the FD and ND condition ($F(1, 29)=5.121$, $p=0.033$).

Alignment Combining the above aspects, the relative number of CSSs that are aligned with the foreseen interaction script increases across the three sessions from 68.78% to 85.95% in the FD condition and from 62.16% to 79.44% in the ND condition (Figure 8). Also these improvements show statistical significance in both the improvement across sessions ($F(1, 29)=9.436$, $p=0.001$) and the difference between the FD and ND condition ($F(1, 29)=5.514$, $p=0.029$).

It is also interesting to look at the improvements in alignment between the first and the second session, the first and the third session, and the second and third session. We performed both the Tukey-Kramer test for differences between means and

Table 9. Data on robot's reactions to child's speech, extracted from the video coding analysis.

Robot's reaction						
	Ignore (%)			Not-ignore (%)		
Sessions	1	2	3	1	2	3
Familiar	23.05	09.67	09.03	76.95	90.33	90.97
Non-Familiar	28.20	19.13	12.89	71.80	81.87	87.11

Note: Mean of the percentage of the child speech acts which were ignored or answered by the robot.

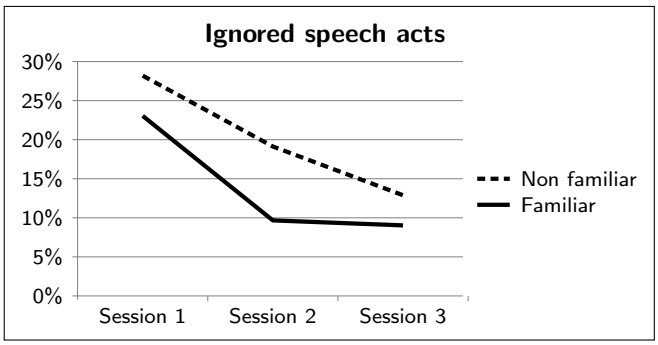

Fig. 7. Speech acts ignored by the robot. The x axis: the three days the children came to the hospital. The y axis: the mean of the different children's percentage of ignored speech acts over the total number of speech acts. The two conditions are shown in two different line styles.

Table 10. Data on speech alignment, extracted from the video coding analysis.

Alignment with the dialogue managed by the robot						
	Aligned (%)			Not-aligned (%)		
Sessions	1	2	3	1	2	3
Familiar	68.78	83.40	85.95	31.22	16.60	14.05
Non-Familiar	62.16	73.07	79.44	37.84	26.93	20.56

Note: Mean of the percentage of the speech acts aligned (i.e., nor overlapped with the robot speech, nor ignored by the robot) and not-aligned.

the Scheffe test for contrasts among pairs of means, using an α=0.05 for both tests, and the result was the same: There is a significant difference between the first and the second session (Sheffe statistic 3.10, critical value 2.59; Tukey-Kramer statistic 4.384, p=0.0131), and between the first and the third session (Sheffe statistic 4.19, critical value 2.59; Tukey-Kramer statistic 5.919, p=0.0010), but not between the

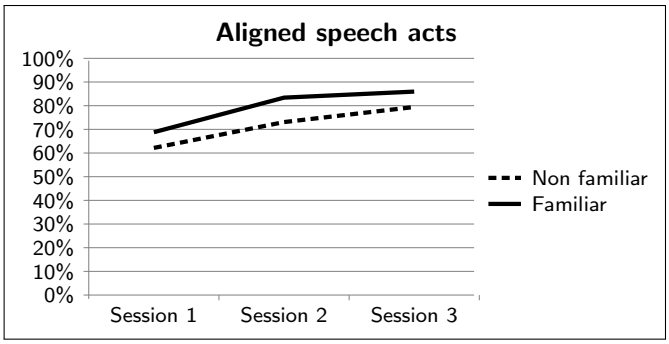

Fig. 8. Aligned speech acts in the dialogue. The x axis represents the three days the children came to the hospital, the y axis represents mean of the different children's percentage of aligned speech acts over the total number of speech acts. The two conditions are shown in two different line styles.

second and the third session (Sheffe statistic 1.09, critical value 2.59; Tukey-Kramer statistic 1.535, p=0.5322).

Table 11. Verbalization rate, calculated as the number of speech acts in a session, divided by the duration of the session. Table reports the mean for all the children (N=10).

Verbalization rate			
Sessions	1	2	3
Familiar	3.13	2.89	2.78
Non-Familiar	4.71	3.85	3.50

CSS Rates The number of CSSs per minute appears to be decreasing from 3.13 to 2.78 in the FD condition, and from 4.71 to 3.5 in the ND condition. Again, the decrease across the sessions is itself not significant ($F(2, 29)=2.272$, $p=0.125$), but the difference between the FD and the ND condition is ($F(1, 29)=12.511$, $p=0.002$). Apparently, children in the ND condition produce almost 40% more CSSs than children in the FD condition.

4.3. *Discussion*

There is a clear change in the children's speech timing and their adherence to the interaction script. Whereas many synchronization problems occur in their first session with the robot, the second and third session are smoother. In particular, the children are often waiting for the robot to talk, even if they know how to continue without the robot's prompt. For example, in the first interactions, having asked a multiple-choice question, the children often go on to read the list of possible answers,

thus causing the robot to barge-in with the possible answers request, while in the subsequent interactions, they wait for the request from the robot before reading the list. Conversely, when the robot asks a question, children might answer straightaway in their initial interaction, again causing the robot to barge-in, whereas in the later interactions they wait for a prompt from the robot. To summarize, children seem to adapt the timing of their speech to the robot's non-adaptive dialogue strategies, so as to avoid speech overlaps. Similar channel exclusion phenomena have been observed in another study of human turn-taking in HRI: subjects waited for the robot to finish speaking before they spoke and tended to avoid simultaneous speaking after a simultaneous start.[12] While other researchers have also studied user speech timing adaptation, they focused on different aspects, e.g., user response latency decrease with practice during a single session,[12] or user response latency adaptation to the systems extrovert/introvert style.[33]

The results show an effect of the familiarity vs. neutral display condition. There are fewer overlaps between child and robot speech in the familarity display condition, and forced waiting of the children for the robot to speak is twice as frequent in the familiarity display condition. These children are apparently more lenient with the robot when it makes mistakes (in particular speech timing mistakes). They adapt their behavior more, for the sake of smooth turn-taking.

The children also adhere more to the foreseen interaction script in the familiarity display condition, as shown by a lower relative number of speech segments to which the robot does not react. A child's speech segment is ignored either because it is out of the currently implemented domain of interaction (e.g., the child confides about belly ache to the robot), or because the child "runs ahead" of the implemented script, and provides information that the robot did not prompt for yet in a situation where the robot is not flexible enough to react to this. The children in the familiarity display condition seem more committed to respect the robot's expectations concerning the interaction script, once they understand them.

The children in the neutral display condition appear to produce more speech segments. What our analysis does not make clear is whether this is a difference in the amount of speech or only in the number of speech chunks. The latter could be a consequence of there being more speech overlaps in the neutral display condition, and thus the children's speech is more fragmented, and we therefore count more child's speech segments. Moreover, since these children tend to deviate more from the foreseen interaction script, resulting in a higher number of speech segments to which the robot does not react, they may (have to) repeat their input more often (until the appropriate system prompt appears).

What it is that leads to these effects is not clear yet. Data from the post-questionnaires concerning the children's relationship to the robot indicate that all the children felt a strong connection with the robot and perceived it as a peer. Even if the familiarity display condition appears to have no effect on these self-assesments (with the caveat of the ceiling effect), the use of the child's name in the familiarity display condition seems to catch their attention and might result in

more concentration on the interaction with the robot. We speculate that this might, consciously or not, lead to the children's higher commitment to the (efficiency of) the interaction.

The fact that the change in alignment is larger between the first and the second session than between the second and third session seems to indicate that the children adapt their behavior to the interaction with the robot quite fast, and this level of adaptation persists. It is not clear whether the children's adaptation is just a consequence of becoming trained in "the rules of the (interaction) game", or it could be linked to social aspects of the interaction, and particularly the children's perception of and interaction with the robot as a social partner. The effect of the familiarity display condition on the adaptation seems to corroborate the latter.

5. Conlusions and Outlook

Our research focuses on verbal behavior adaptation of children in face-to-face interaction with a robot across multiple sessions. We built an HRI system using the humanoid robot Nao and set up an experiment in which children interacted with the robot in three sessions on different days, engaging in a quiz, dance or imitation activity, in one of two conditions: the robot either gave explicit verbal and non-verbal signals of being familiar with the user from previous interactions, or it did not. The experiment system relied on a human wizard to interpret the user's speech and gesture input, but the rest of the system behavior was automatic, notably dialogue management and system output behavior production.

We observed informally and in a preliminary analysis on a small subset of the interactions that the children adapted various aspects of their verbal and non-verbal conversational behavior to the robot, including speech timing, speed and tone, verbal input formulation, nodding and gestures, just as humans generally adapt to their conversational interlocutors. We therefore carried out a follow-up systematic analysis of all quiz interactions focusing on the children's verbal turn-taking behavior. In particular, we analyzed the timing of the children's speech, and whether or not the robot reacted to a child's turn. We found that adaptation increases across multiple sessions. Moreover, we found that children adapt their verbal turn-taking behavior more readily when the robot gives explicit signals of familiarity with the child across sessions, for example by using their name or referring to previous experiences. The children who interact with a familiarity-displaying robot force themselves more to wait with speaking, in order to avoid speech overlaps, and produce less turns that end up ignored by the robot. So overall, there is more adaptation in the condition with familiarity display in comparison to the condition where the robot's behavior is neutral in this respect.

One practical upshot of these results is that a robot explicitly displaying familiarity can elicit more cooperation from a (young) user leading to a smoother communication. There might also be more tolerance towards such a system, despite its inevitably imperfect interaction capabilities.

In immediate future work we plan to extend the analysis to other aspects of verbal behavior, especially verbal input formulation, and to the other activities performed in the experiment.

References

1. Acapela Development Team. The Acapela Text-To-Speech Synthesis System. http://www.acapela-group.com/, (accessed 1.5.2012).
2. J. Baillie. Urbi: Towards a universal robotic low-level programming language. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3219–3224. IEEE, 2005.
3. J. Baldridge and G.J.M. Kruijff. Coupling CCG and hybrid logic dependency semantics. In *Proc. ACL 2002*, pages 319–326, Philadelphia, PA, 2002.
4. Jason Baldridge and Geert-Jan Kruijff. Multi-modal combinatory categorial grammar. In *Proceedings of 10th Annual Meeting of the European Association for Computational Linguistics*, Budapest, Hungary, 2003.
5. John A. Bateman. Enabling technology for multilingual natural language generation: the KPML environment. *Natural Language Engineering*, 1(1), 1997.
6. A. Beck, L. Cañamero, and K.A. Bard. Towards an affect space for robots to display emotional body language. In *Proceedings of the 19th IEEE international symposium on robot and human interactive communication*, Ro-Man 2010, pages 464–469. IEEE, 2010.
7. M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke. Integrating vision and speech for conversations with multiple persons. In *IEEE/RSJ IROS*, pages 2523–2528, Aug 2005.
8. G. Bradsky and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly, 2008.
9. S. Brennan. Lexical entrainment in spontaneous dialogue. In *Proceedings of the International Symposium on Spoken Dialogue (ISSD-96)*, pages 41–44, 1996.
10. S. Brennan and J.O. Ohaeri. Effects of message style on user's attribution toward agents. In *Proceedings of CHI'94 Conference Companion Human Factors in Computing Systems*, pages 281–282. ACM Press, 1994.
11. Judee K. Burgoon, Lesa A. Stern, and Leesa Dillman. *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge Univ. Press, 1995.
12. Crystal Chao, Jinhan Lee, Momotaz Begum, and Andrea L. Thomaz. Simon plays simon says: The timing of turn-taking in an imitation game. In *Proceedings of the 20th IEEE International Symposium on Robotics - RO-MAN*, 2011.
13. Herbert Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.
14. F. G. Cozman. Generalizing Variable Elimination in Bayesian Networks. In *IBERAMIA/SBIA, Workshop on Probabilistic Reasoning in Artificial Intelligence*, pages 27–32, 2000.
15. H. Cuayáhuitl. Learning dialogue agents with Bayesian relational state representations. In *Proceedings of the IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems (IJCAI-KRPDS), Barcelona, Spain*, pages 9–15, Jul 2011.
16. H. Cuayáhuitl and I. Kruijff-Korbayová. An interactive humanoid robot exhibiting flexible sub-dialogues. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Montreal, Canada*, pages 17–20, Jun 2012.
17. Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. Evaluation of a hierarchical reinforcement learning spoken dialogue system. *Computer Speech*

and Language, 24(2):395–429, 2010.
18. Nina Dethlefs, Heriberto Cuayáhuitl, Kai-Florian Richter, Elena Andonova, and John Bateman. Evaluating task success in a dialogue system for indoor navigation. In *Proc. of the 14th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, pages 143–146, 2010.
19. MARY Development Team. The MARY Text-To-Speech System. http://mary.dfki.de/, (accessed 1.5.2012).
20. T. Dietterich. An overview of MAXQ hierarchical reinforcement learning. In *Symposium on Abstraction, Reformulation, and Approximation (SARA)*, pages 26–44, 2000.
21. Elizabeth Z. Ford. How to get people to say and type what computers can understand. *Int. J. Man-Mach. Stud.*, 34(4):527–547, April 1991.
22. Howard Giles, Anthony Mulac, James J. Bradac, and Patricia Johnson. Speech accommodation theory: The first decade and beyond. In Margaret L. McLaughlin, editor, *Communication Yearbook 10*, pages 13–48. SAGE Publications, 1987.
23. R. Gockey, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz, and J. Wang. Designing robots for long-term social interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems -IROS'05*, pages 1338–1343, 2005.
24. Julius Development Team. Open-Source Large Vocabulary CSR Engine Julius. http://julius.sourceforge.jp/, (accessed 1.5.2012).
25. T. Kanda, T. Hirano, and E. Daniel. Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, 19:61–84, 2004.
26. Cory D Kidd. *Designing for Long-Term Human-Robot Interaction and Application to Weight Loss*. Ph.d. dissertation, Massachusetts Institute of Technology, 2008.
27. Ivana Kruijff-Korbayová, Heriberto Cuayáhuitl, Bernd Kiefer, Marc Schröder, Piero Cosi, Giulio Paci, Giacomo Sommavilla, Fabio Tesser, Hichem Sahli, Georgios Athanasopoulos, Weiyi Wang, Valentin Enescu, and Werner Verhelst. Spoken language processing in a conversational system for child-robot interaction. In *Workshop on Child-Computer Interaction*, 2012.
28. Ivana Kruijff-Korbayová, Heriberto Cuayáhuitl, Bernd Kiefer, Marc Schröder, Piero Csi, Giulio Paci, Giacomo Sommavilla, Fabio Tesser, Hichem Sahli, Georgios Athanasopoulos, Weiyi Wang, Valentin Enescu, and Werner Verhelst. A conversational system for multi-session child-robot interaction with several games. In *German Conference on Artificial Intelligence (KI)*, 2012. system demonstration description.
29. Iolanda Leite, Carlos Martinho, André Pereira, and Ana Paiva. As time goes by: Long-term evaluation of social presence in robotic companions. In *Proceedings of the 18th IEEE International Symposium on Robotics - RO-MAN*. IEEE Computer Society, 2009.
30. O. Lemon and O. Pietquin. Machine learning for spoken dialogue systems. In *INTERSPEECH*, pages 2685–2688, 2007.
31. Marco Nalin, Ilaria Baroni, Ivana Kruijff-Korbayová, Lola Cañamero, Matthew Lewis, Aryel Beck, Heriberto Cuayáhuitl, and Alberto Sanna. Children's adaptation in multi-session interaction with a humanoid robot. In *Proceedings of the Ro-Man Conference*, Paris, France, 2012.
32. OpenCCG Development Team. OpenCCG: The OpenNLP CCG Library. http://openccg.sourceforge.net/, (accessed 1.5.2012).
33. Sharon Oviatt, Courtney Darves, and Rachel Coulston. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Trans. Comput.-Hum. Interact.*, 11(3):300–328, September 2004.
34. Sathish Pammi. Prosody control in HMM-based speech synthesis. Technical report,

DFKI. Aliz-E Project, 2011.
35. Byron Reeves and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, 1996.
36. Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.
37. S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2003.
38. M. Schröder and J. Trouvain. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377, 2003.
39. Marc Schröder, Marcela Charfuelan, Sathish Pammi, and Ingmar Steiner. Open source voice creation toolkit for the MARY TTS platform. In *Proc. Interspeech*, Florence, Italy, 2011.
40. Mark Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31(4):649–689, 2000.
41. R. Stiefelhagen, H.K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel. Enabling multimodal human-robot interaction for the Karlsruhe humanoid robot. *IEEE Transactions on Robotics: Special issue on human-robot interaction*, 23(85):840–851, 2007.
42. H. I. Witten, E. Frank, and A. M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2011.
43. S. Young. *The HTK Book*. Cambridge University Engineering Department, 2007.

6. Authors' bios

Ivana Kruijff-Korbayová graduated in 1992 from the Faculty of Electrotechnical Engineering, Czech Technical University, Prague. In 1998 she obtained a PhD in mathematical linguistics at the Faculty of Mathematics and Physics, Charles University, Prague. In 1999-2000 she held a British Academy Visiting Fellowship and a Royal Society / NATO Postdoctoral Fellowship for research at the University of Edinburgh. In 2001 she took up a research position at the Department of Computational Linguistics and Phonetics, Saarland University, where she has been also regularly teaching. She has joined the DFKI as a senior scientist in 2008. She has been a project leader on several national and international research projects. Her research interests cover various areas of discourse and dialogue processing.

Heriberto Cuayáhuitl received a PhD degree in Informatics from the University of Edinburgh in 2009. He has been a postdoctoral researcher at the University of Bremen (2009-2010), and the German Research Center for Artificial Intelligence (DFKI) in Saarbrücken (2011-2012), both in Germany. His research interests lie in machine learning methods for speech-based and multimodal interactive systems exhibiting adaptive behaviour in unknown and uncertain environments.

Bernd Kiefer received his diploma in Computer Science from the Saarland University in 1989. Since 1990, he works at DFKI in Saarbrücken on various aspects of natural language processing. One main research focus is the analysis of language by means of constraint-based grammars, and the relation and transformation of the many existing formalisms in this area. Furthermore, he investigates the use and development of advanced data structures and algorithms for NL processing in general.

Marco Nalin is program manager in the telemedicine company Telbios S.p.A., based in Milan. He received his M.Sc. degree in electronic engineering at the University of Padova, Italy, in 2005. Since 2004 he has been working as research scientist at San Raffaele Hospital, in Milan, in the e-Services for Life and Health department, cooperating on several projects funded by the EU Commission. In January 2013, he joined the telemedince company Telbios S.p.A., cooperating and supervising R&D projects. His research interests include personal health systems, mobile health management applications for personal wellbeing and disease

prevention, telemedicine, cognitive robotics and edutainment, surgery robotics, energy management systems, cloud computing, privacy and security.

Ilaria Baroni is a research scientist at the San Raffaele Hospital, Milan. She graduated at Politecnico di Milano in computer science engineering (specialization in robotics). For her thesis, she worked on hardware and software aspects of humanoid robots. Since 2010, she has been working at San Raffaele in the e-Services for Life and Health department, cooperating on two projects funded by the European Commission (cloud computing and cognitive robotics), contributing both to scientific research and technical development activities. Her fields of investigation include personal health systems, self-management systems, patient monitoring systems, cognitive robotics and edutainment, cloud computing, privacy and security.

Alberto Sanna graduated in Nuclear Engineering at Politecnico di Milano. He has been in charge of healthcare process re-engineering projects at the Scientific Institute of San Raffaele since 1989, leading highly innovative Information Technology and Automation & Robotics-driven clinical projects in Nuclear Medicine, Clinical Lab, Hospital Pharmacy, Ward and Surgical Room. Since 1999 he is director of the e-Services for Life and Health Research Unit (www.eservices4life.org) and is successfully managing many international R&D projects.