Second Language Tutoring using Social Robots

Project No. 688014

L2TOR
Second Language Tutoring using Social Robots

Grant Agreement Type: Collaborative Project
Grant Agreement Number: 688014

D5.1 Interaction management for the number domain

Due Date: **31/09/2017**
Submission Date: **16/10/2017**

Start date of project: **01/01/2016** Duration: **36 months**

Organisation name of lead contractor for this deliverable: **Bielefeld University**

Responsible Person: **Stefan Kopp** Revision: **1.0**

Project co-funded by the European Commission within the H2020 Framework Programme		
Dissemination Level		
PU	Public	**PU**
PP	Restricted to other programme participants (including the Commission Service)	
RE	Restricted to a group specified by the consortium (including the Commission Service)	
CO	Confidential, only for members of the consortium (including the Commission Service)	

Contents

Executive Summary

This deliverable describes the development of the interaction manager including, but not restricted to the number domain. We present the overall architecture, communication between the modules and decisions taken towards each component. In addition, we provide insight gathered from qualitative and quantitative empirical work that was executed to inform the interaction manager about well-suited interaction patterns for child-robot tutoring, feedback mechanisms, motivational behaviors and cues of interaction engagement the robot should attend to. Furthermore, we describe difficulties we encountered during our pilot test and propose solutions to overcome these problems in the future.

Principal Contributors

The main authors of this deliverable are as follows (in alphabetical order):

Kirsten Bergmann, Bielefeld University
Laura Hoffmann, Bielefeld University
Stefan Kopp, Bielefeld University
Thorsten Schodde, Bielefeld University

Revision History

Version 1.0 (16-10-2017)
First Version.

1 Introduction

The goal of work package (WP) 5 is the development of an interaction management component which is responsible for planning and choosing the system reactions. The interaction manager receives its input from the input recognition and interpretation modules (WP4), which record and interpret the children's behavior during the tutoring interaction. Based on this input, the interaction manager outputs high level messages that specify semantic and pragmatic aspects and sends it to the multimodal output generation (WP6), which will transform this high level behaviour to actual verbal and non-verbal output realized on the tablet and robot (see Figure 1). The pragmatic output includes interaction management functions, socio-relational functions, and tutoring-related functions. Besides, a major goal of WP5 is the development and maintenance of a learner model. The model provides the central representation for capturing and joining cues extracted from the child's behaviour and forms the basis for planning system responses. To that end, the model will enable a form of mental perspective-taking in the sense that the system is able to reason about the understanding of the situation from the learner's perspective, and predict consequences of dialogue and tutoring moves on the state of the learner.

The project proposal specifies the content for Deliverable 5.1 as the interaction manager for the number domain. The lessons for this domain consist of number words, e.g. one, two and three, and knowledge about (pre-)mathematical concepts, such as weight, size and quantities. The child will learn these target words through game-like activities played on a tablet together with the humanoid robot Nao. Activities include for example counting animals in a zoo while dragging them into their cage (comparing quantities, number words) or creating of a bouquet in a flower shop (comparing and manipulating quantities, number words) and many more (cf. Deliverable 1.1). Although the settings and target words will change in the different lessons and domains, the underlying mechanism for the interaction management will not change much and can be defined in a more general way so that they can be used in all three domains. Hence, we focus on the tasks for WP5 as specified in the proposal from a more general view that is suitable to the number domain, but also for the other domains.

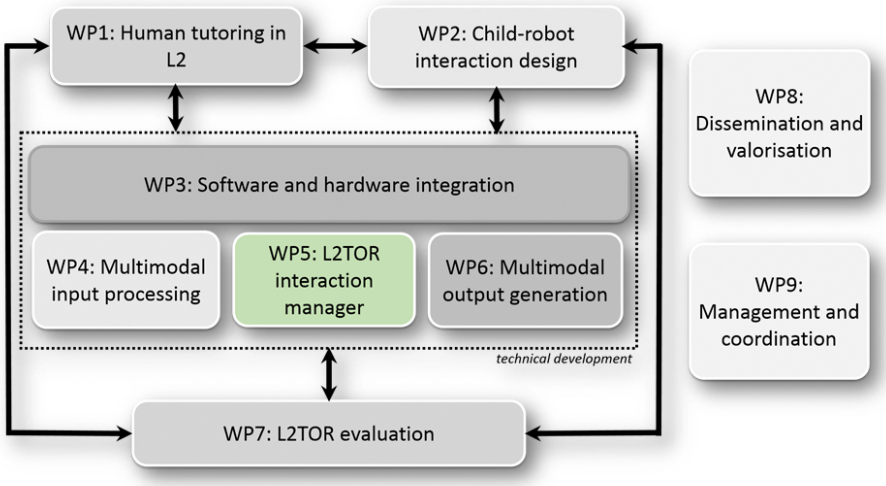

Figure 1: Workpackage overview and how they work together.

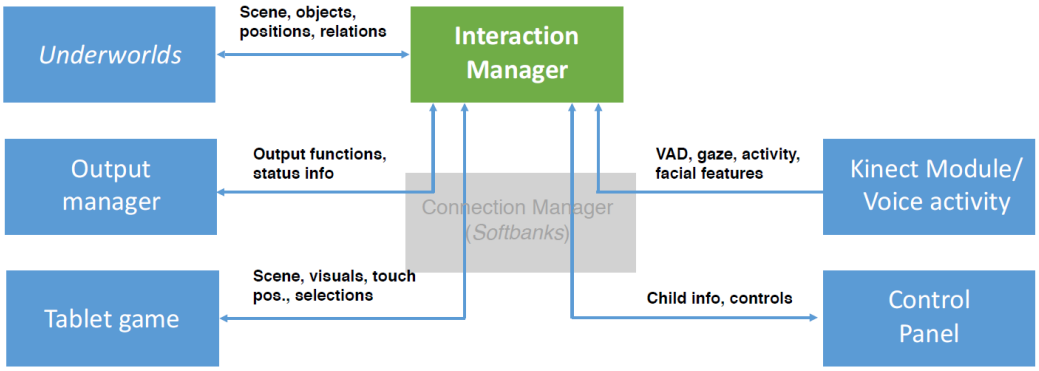

Figure 2: Communication interface structure of all modules used in the L2TOR System.

2 Tasks of the Interaction Manager

2.1 Input and Output Specification and Representation (T5.1)

2.1.1 Overall Interfaces between Interaction Manager and the Other Modules

Since the interaction manager has to receive, organize and distribute information between all other modules, and has to control the system flow, a communication protocol that can be used with different programming languages is needed. Therefore Softbank provides a middleware called “Connection Manager”, which is based on a TCP/IP connection. The communication flow between the interaction manager as te integral module and all other modules is illustrated in Figure 2 (for more general information about the project architecture see Deliverable 3.1).

As depicted in Figure 2, all modules, except for Underworlds, communicate via the *Connection Manager* and share information and events. For a detailed list of all functions with parameters please see Appendix A.

Underworlds
Some tasks during the learning interaction require the evaluation of spatial relations between the objects on the tablet, hence, a software module called “Underworlds” has been integrated that is capable to solve this problem. The module is initialized with the tablet scene and needs to be informed about all changes in the position of all objects to work properly. We decided to integrate Underworlds with a direct interface to the interaction manager, because of the delays while sending the messages between the modules (tablet game - connection manager - underworlds - connection manager - interaction manager and for feedback from interaction manager over connection manager to the output manager) added up to an unacceptable amount of time to provide a smooth interaction. Due to that the interaction manager will inform Underworlds about all object-position changes in the tablet game and can request the spatial relations for a specific object directly if needed.

Output manager
The output manager is the second module that allows the interaction manager to communicate with the learner. It will express requests and present the next task, request an answer, or give feedback. Later on, it will also be possible to provide help based on all task-related information stored in the interaction manager. To be able to control the flow, the interaction manager also needs to know when an action

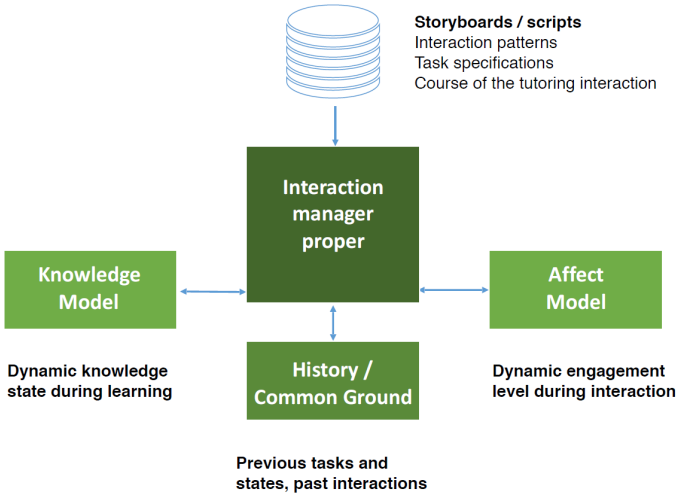

Figure 3: Internal structure of the interaction manager module.

(e.g., speech or gesturing) has been executed. Therefore, status events have been introduced, which are sent from the output manager to the interaction manager as a result of an executed request.

Tablet Game
The tablet game represents the graphical user interface for the child-robot interaction. Here the interaction manager takes care of which scene is displayed, and which objects have to be displayed, hidden, enabled for touch, or highlighted to ensure a smooth interaction. Since another important task of the interaction manager is the validation of the child's response, the tablet game sends the identifier (id) of the touched object and the touch position on the screen back to the interaction manager.

Control Panel
The control panel is capable of loading a child's interaction history into the interaction manager or to create a new one. Furthermore, the experimenter is able to control the interaction flow with start, stop, pause and resume commands. This commands are also send to the interaction manager, which executes the necessary steps and informs other modules if needed.

Kinect Module
The Kinect module provides all needed sensory data as voice activity, the gazed target, bodily activity and facial features to detect the affective state of the learner (see Section 2.6). Also basic situational information, e.g. if the child is still sitting in front of the tablet, are recognized by this module. All sensed information will be stored and made available for decisions making in the interaction manager.

2.1.2 Internal Architecture of the Interaction Manager

For a better understanding how the interaction manager works, the internal architecture is described in detail in the following. As can be seen in Figure 3, the interaction manager consists of four main elements. The knowledge model traces the skill mastery of the learner to allow an adaptive lesson planning. In the affect model the affective state of the learner, here engagement in the interaction, is inferred from a bunch of behavioral cues as bodily activity, gaze target of the learner, and facial expression (i.e., smiling). The interaction manager proper contains basic interaction patterns and

General Child Information	Description
Child-ID	Unique identifier for a particular child.
Name	The name of the child to be used by the robot to directly address the child in a verbal utterances.
General Task Information	
Task-ID	The ID of a particular task
Task-Type	The task-type, e.g., “select_object” or “move_object”.
Target words	The target words addressed in the task.
Answer right/wrong	The given answer and if it was right or wrong.
Response time	The time it took the child to answer after the task was given.
Feedback-Type	What type of feedback, or which feedback-parameters, were used.

Table 1: The different information which will be stored in the history model.

advanced flow control mechanism which are additionally informed by an external script containing more advanced interaction patterns and task specifications. It is responsible for controlling the flow of the interaction, maintaining the attention and engagement of the child, validating the task, requesting an answer if the child is not responding, and providing feedback. The feedback will be parametrized by the information stored in the knowledge, affect and history model. The history contains information about the general task history (e.g. correctness and response time), and further the type of feedback and knowledge that was provided and affective states of the past (see Table 1).

2.2 Model of the Child Learner’s States and Traits (T5.2)

The interaction manager needs to keep track of information about the child’s state in order to adapt to the individual needs of each learner. The general information about the child’s performance is stored in two modules: the “Knowledge Module” (see Section 2.4) and the “History/ Common Ground” module. The latter will store all information related to the interaction history to enable the system to establish a common ground (i.e., a shared history of experiences) between the system and the child. This will include information about the child, as his/her name and an identification number (Child-ID), as well as general task information, e.g., which tasks have been finished yet? were the answers given correctly? how long did the child take to answer? what was the knowledge state and the affective state at the moment? etc. It should be remarked that this module is not fully implemented yet. We are still running experiments and pilot tests to decide which information will be finally stored. A preliminary list can be found in Table 1. The history/common ground module can be regarded as a longterm memory, which enables the system to adapt to each child based on the actual interaction state, but also under consideration of earlier interactions. The system should hence be able to form expectations about each child’s feedback preferences and learning speed according to earlier experiences.

2.3 Basic Interaction Management (T5.3)

To set up the basic interaction management for the different domains, we decided to design the interaction on an empirical basis of human-to-human language tutoring data, to create a tutoring interaction that matches children’s needs. For that purpose, video recordings of language tutoring games were collected as they take place in kindergartens. Given that one-to-one interactions of educator

and child can hardly be realized in kindergartens, the games typically involve one educator and a small group of children. Data of four language tutoring games have been collected: reading a picture book together with children in an interactive manner; card game "I spy with my little eye"; card game "I'm giving you a present"; and a rhyming game. The collected data comprises round about 681 minutes of recorded video data from children between age four and six. These recordings have been transcribed and annotated with regard to the following categories:

- **Dialogue acts**: Utterances are classified due to the underlying intention based on the DAMSL annotation scheme [1].
- **Children's mistakes**: Types of language errors the children made, e.g. wrong plural form, missing articles, wrong syntax, etc.
- **Educator's speech repair**: Pedagogical acts used to correct the errors, e.g. reformulation, corrected repetition, etc.
- **Nonverbal behaviour**: Nods, smiles, gestures etc. used by the educators.

On the basis of these annotations, overall patterns were identified to inform the detailed design of the robot's behaviour. These fall basically into two categories, (i) overall interaction structure and (ii) feedback behaviour by the educators.

2.3.1 Overall Interaction Structure

A common pattern in all language tutoring games under investigation was the following basic course of actions:

1. **Opening:** Marks the beginning of the interaction and should mitigate the children's timidity as well as it should motivate the child.
2. **Game Setup:** This step is used to prepare the game by explaining the task and clarify the necessary terms.
3. **Test run:** A test run of the game is conducted to test whether the instructions have been understood and to practice the game flow.
4. **Game:** Here, the main interaction game takes place. Every move is accompanied by the educator's feedback and motivations to continue.
5. **Closing:** Marks the end of the learning interaction. Additionally, it is used to ensure motivation for future interactions by acknowledge the participation, joint singing a goodbye song and an outlook on what's going to happen next time.

2.3.2 Educator's Feedback Behaviour

In addition, the educators' behaviour when providing children with feedback was analysed. An important and common pattern is that language errors are almost never corrected explicitly. Instead, feedback is always provided in a positive way, falling into one of the following categories with the percentage of their occurrence given in squared brackets: (i) **praising the child** for a correct utterance whereby praise is often combined with a repetition of the correct word [13%] (ii) **implicit correction** in case of an error made by the child: repetition of the word as if correct (e.g. correct pronunciation,

with article, plural form, etc.) [54%], (iii) **correct recasting of a sentence**, e.g. after syntax errors [32%], (iv) **moving on to next task**, e.g. when the child's message is unclear due to incomprehensible pronunciation [1%]. All kinds of educators' feedback behaviour is typically accompanied by looking at the child, smiling and nodding. These findings are also supported by our observations in schools, where educators showed mainly positive feedback. Only a few times educators used negative feedback, which is rather implicitly given, see Deliverable 1.2 for more details.

2.3.3 Conclusion

Besides the observational studies, we also evaluated different dialog managers as OpenDial [2] and IrisTK [3], but the results gathered from observations mainly done in the realm of WP1 (see Deliverable 1.2) revealed that no full-fledged dialogue management is needed for our purpose, because large parts of the tutoring interaction structure can be pre-designed and determined. The lessons learned from the empirical studies have thus been used to build interaction scripts that ensure smooth and well-structured tutoring interactions. However, these scripts can still be adjusted to allow some adaptation at predefined points for intervention. For instance, the system can adapt the lesson to the learning speed of the children (see Section 2.4). In addition, it is planned that the interaction can be paused at any time, or at least after a task has been finished, to include motivational or relaxing activities for the children to recover until the learning interaction will go on (see Section 2.6).

2.4 Probabilistic State Estimation and Update (T5.4) & Decision-theoretic Dialogue Management (T5.5)

We developed a novel approach to personalize language tutoring in human-robot interaction [4]. This adaptive tutoring approach is based on a model of how tutors mentalize about learners – by keeping track of their knowledge state and by selecting the next tutoring actions based on their likely effects on the learner. This is realized in an approach that combines knowledge tracing (of what the learner learned) with tutoring actions in one causal probabilistic model. The combination allows the selection of skills and actions based on notions of optimality – here the desired learner's knowledge state as well as optimal task difficulty – to achieve a given skill.

2.4.1 Knowledge Tracing and Decision Making

The approach is based on Bayesian Knowledge Tracing (BKT) [5], a specific type of Dynamic Bayesian Networks (DBNs). The model consists of two types of variables, namely the *latent variables* representing the belief state of 'skills' to be acquired (e.g. whether a word has been learned or not) and the *observed variables* representing the observable information of the learning interaction (e.g. whether an answer was correct or not). In our proposed model, each latent variable can attain six discrete values, corresponding to six bins for the belief state (0%, 20%, 40%, 60%, 80%, 100%) representing whether a skill is mastered or not. That is, we reduce the complexity we would get through continuous latent variables but also attain more flexibility. The observed variables remain binary and still represent whether a learner's response is correct or not (see Figure 4). Moreover, the following update of the belief state of the skill, i.e. the skill-belief, at time $t+1$ is not only based on the previous skill-belief, but also on the chosen action and the previous observation at time t.

Based on this model, two types of decisions are made, (1) which skill would be the best to address next, and (2) the choice of action(s) to address that skill. Regarding the former, we employ a heuristic that maximizes the beliefs of all skills while balancing the single skill-beliefs among each other. This

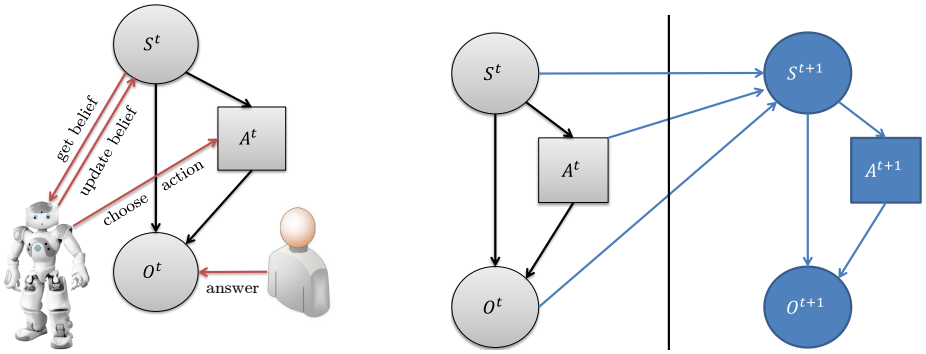

Figure 4: Dynamic Bayesian Network for BKT: With the current skill-belief the robot chooses the next skill S^t and action A^t for time step t (left). After observing an answer O^t from the learner, this observation together with action A^t and the previous skill-belief S^t are used to update the skill-belief S^{t+1} at time $t+1$ (right) [4].

strategy is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system [6]. Regarding the choice of actions, the model enables the simulation of the impact each action has on a particular skill. To keep the model simple, the action space of the model consists of three different task difficulties (easy, medium, hard). Consider an example where the skill-belief appears relatively high, such that the skill is nearly mastered by the learner. In this case, a less challenging task would only result in a relatively minor benefit for the training of that skill. In contrast, if we assume the skill-belief to be rather low and a very difficult task is given, the student would barely be able to solve the task, likewise resulting in a smaller (or non-existent) learning gain. Instead, a task of adequate difficulty, not too simple nor too complicated for the student to solve, will result in a higher learning gain [7]. This helps to position the robot as a capable instructor that uses these scaffolding techniques to help children acquire new skills beyond what they could have learned without help, by bringing them into the zone of proximal development (ZPD) [8].

2.4.2 Evaluation with Adults

When implemented in the robot language tutor, the model will enable the robot tutor to trace the learner's knowledge with respect to the words to be learned, to decide which skill (word) to teach next, and how to address the learning of this skill in a game-like tutoring interaction. To evaluate this model a study has been conducted (c.f. [4]), where participants (students) were asked to learn ten vocabulary items German – 'Vimmi' (Vimmi is an artificial language that was developed to avoid associations with other known words or languages for language-related experiments [9]). The items included colors, shapes and the words 'big' and 'small'. During the game, the robot introduced one of the Vimmi words. A tablet then displayed several images, one of which satisfied the Vimmi description (e.g. one object that is blue) and a number of distractors. The participant was then asked to select the image corresponding to the described item. Participants learned vocabulary items in one of two conditions, either in the condition with the adaptive model (20 participants) or in a non-adaptive (random) control condition (20 participants). In the adaptive condition, the skill to be taught and the action to address the skill were chosen by the model as described above. Participants' performance was assessed with two measures: (1) learners' response behavior was tracked over the course of the training to investigate the progress of learning, and (2) a post-test was conducted on the taught vocabulary in the form of both L1-to-L2 translations and L2-to-L1 translations to assess participants' state of knowledge following the intervention.

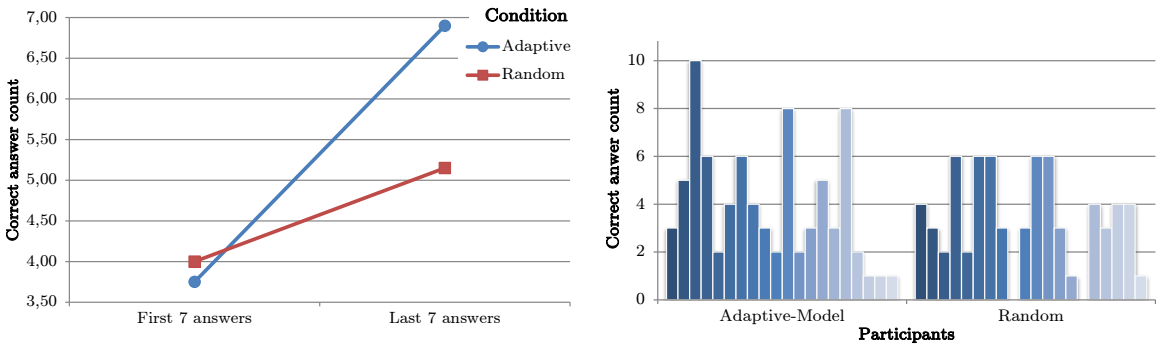

Figure 5: On the left, the mean numbers of correct answers at the beginning (first 7) and end (last 7) of the interaction in the different conditions and on the right, the participant-wise amount of corrects answers grouped by the different conditions for the German-to-Vimmi post-test.

	Adaptive (A)		**Control (C)**		**A, C**	
	M	**SD**	**M**	**SD**	**M**	**SD**
F7	3.75	1.37	4.00	1.17	3.88	1.27
L7	6.90	0.31	5.15	1.69	6.03	1.49
F7, L7	5.33	0.69	4.58	1.12		

Table 2: Means (M) and standard deviations (SD) of correct answers for the initial quarter of the training interaction (first seven items – F7) and the final quarter (last seven items – L7) in each condition, as well as the inter-model (A, C) and intra-model (F7, L7) means and standard deviations.

To analyze the participants' response behavior over the course of training, a mixed-design ANOVA with training phase (initial, final) as a within-subjects factor and training type (adaptive-model-based, control) as between-subjects factor has been conducted. Results are summarized in Table 2 and Figure 5 on the left. As expected, there was a main effect of training phase ($F(1, 38) = 66.85, p < .001, \eta^2 = .64$): Learners' performance was significantly better in the final phase as compared to the initial phase. In the first quarter of training, participants achieved a mean of 3.88 ($SD = 1.27$) correct responses, whereas in the final quarter, a mean of 6.03 ($SD = 1.49$) items was selected correctly. More interestingly, there was also a main effect of training type ($F(1, 38) = 6.52, p = .02, \eta^2 = .15$) such that participants who learned in the adaptive condition had a higher score of correct answers ($M = 5.33, SD = .69$) as compared to learners in the control condition (with an average of $M = 4.58$, $SD = 1.12$ correct answers). Finally, the interaction between training phase and training type was also significant ($F(1, 38) = 14.46, p = .001, \eta^2 = .28$) indicating that the benefit of adaptive-model-based training develops over time (see Figure 5). While participants' response behaviour in the first quarter of training was similar across conditions, a benefit of training with the adaptive model became evident in the final quarter. At this stage of training, participants in the adaptive model condition achieved a mean of $M = 6.9$ ($SD = .31$) correct responses, whereas participants in the control condition achieved a mean of $M = 5.15$ ($SD = 1.69$) correct responses.

The analysis of participants' response behaviour over the course of training has clearly shown that participants learned the L2 words during the human-robot interaction. Importantly, they learned more successfully with our adaptive model as compared to a randomized training. That is, the repeated trials addressing still unknown items as chosen by the adaptive model (until the belief state about these words equaled that of known items) outperformed the tutoring of the same material (same number of trials

	Adaptive (A)		Control (C)	
	M	**SD**	**M**	**SD**
German-to-Vimmi	3.95	2.56	3.35	1.98
Vimmi-to-German	7.05	2.56	6.85	2.48

Table 3: Results of both post-tests (German-to-Vimmi and Vimmi-to-German): Means (M) and standard deviation (SD) of correct answers grouped by the experimental conditions.

and items) but in randomized order. In the post-test, however, there was no significant difference across experimental conditions, despite a trend towards better performance in the adaptive model conditions over the controls.

In this post-test the participants' learning performance has been measured with two translation tests (L2-to-L1 and L1-to-L2). The results are summarized in Table 3. Paired-samples t-tests were conducted to compare the number of correctly recalled words after training with the adaptive model as compared to training in the control condition. For the German-to-Vimmi translation, there was no significant main effect ($T(38) = .25, p = .80$). Participants who trained with the adaptive-model recalled a mean of 3.95 ($SD = 2.56$) out of ten words correctly, while participants in the control condition recalled a mean of 3.35 ($SD = 1.98$) words. Likewise, there was no significant main effect ($T(38) = .83, p = .41$) for the Vimmi-to-German translation task. Participants' performance after learning with the adaptive model amounted to a mean of 7.05 ($SD = 2.56$) correct items, participants' performance in the control condition to a mean of 6.85 ($SD = 2.48$) correct items.

Although no main effect of training type emerged in the post test, some details might nevertheless be worth mentioning. In the German-to-Vimmi post test, a maximum of ten correct responses was achieved by participants in the adaptive-model condition, whereas the maximum of participants on the control condition were six correct answers. Moreover, there were two participants in the control condition who did not manage to perform any German-to-Vimmi translation correctly. In the adaptive-model condition, all participants achieved at least one correct response (see Figure 5 right).

Different explanations may account for this inconsistent finding. One potential explanation could be that the way how responses were prompted was not identical in training sessions and post test. In the training sessions, participants saw pictures reflecting the meaning of the to-be-learned words whereas in the post-test they just received a linguistic cue in form of a word they had to translate. It might be that repeated trials as they were particularly supported for difficult-to-remember items by the adaptive model, led to stronger associations between linguistic and imagistic materials. This might have caused a stronger decline of correct responses for participants who trained with the adaptive model as opposed to those in the control condition. An alternative explanation could be that test results measured immediately after the training session are subject to strong inter-individual differences among learners. This is the reason why studies on vocabulary learning usually range over repeated sessions spread over several days (cf. [10]). A typical pattern is that significant results emerge after two or three sessions/days and/or in the long-term (measured several weeks after training took place). So it might well be that further training sessions or delayed tests might result in a post test performance that matches the picture of the during-session performance.

Overall, results from the evaluation study are, at least, in parts very promising: learners' performance during training was significantly improved by personalized robot tutoring based on the adaptive model.

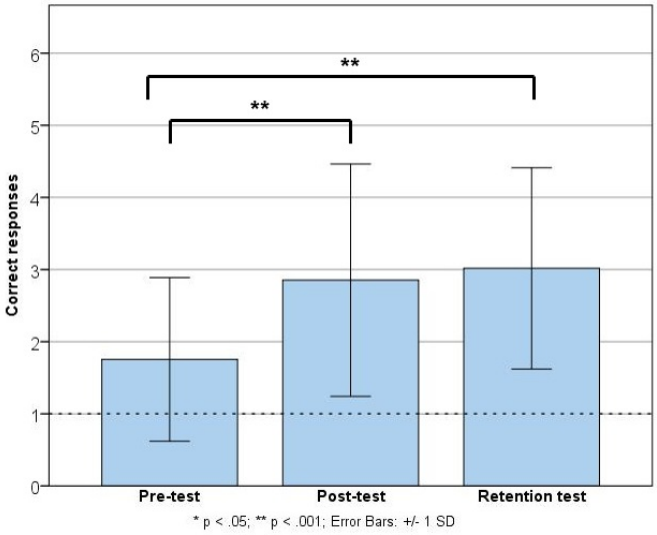

Figure 6: Average number of English words correct in Pre-test, Post-test and Retention test.

2.4.3 Evaluation with Children

Since the project aims for teaching a second language to children and not to young adults, another study has been conducted with our project partners from Tilburg (for more details please see [11], Deliverable 6.1). Therefore the system, especially the tablet game and the verbal output, has been adapted to be suitable for children. Additionally, the second language to be taught also has been changed to English and the words to be learned are English names of animals. Furthermore, the study should not only test the effect of the adaptive system but also the effect of iconic gestures on the learning gain of children. Hence, the study was conducted in a 2 (system: adaptive vs. random) x 2 (gestures: present vs. absent) between subjects design. For the testing of learned vocabulary, pictures slightly different from that of the teaching session were presented, so that it could also be tested whether the child learned a word-image mapping only, or if it learned the mapping of the word to the actual concept of the animal. Images of all animals were shown at the same time on a computer screen, while the computer asks for a specific animal to be clicked on. This way, the test was not a production test (cf. other study in Section 2.4.2). In addition, a retention test has been conducted a week after the teaching session, to check whether the learning effect in the post-test holds or even raises after time.

Results on learning gain

In total, 61 children between the age of four and six ($M = 62.2$ months, $SD = 6.9$ months) that were native Dutch speakers were included in the analyses. To test whether children managed to learn new words from the interaction regardless of strategy or the use of gestures, a paired-samples t-test was conducted to measure the difference between post-test and pre-test scores for all conditions combined (Figure 6). There was a significant difference between the scores on the pre-test ($M = 1.75$, $SD = 1.14$) and immediate post-test ($M = 2.85, SD = 1.61$), $t(60) = 5.23, p < .001$. Children on average recalled more words (animals) in the post-test compared to the pre-test. When considering the retention test (at least) one week after the initial test, the difference to the pre-test was still significant $t(60) = 6.81, p < .001$. The average amount of animals recognized one week after the test was even slightly higher ($M = 3.02, SD = 1.40$) than immediately after the child–robot interaction.

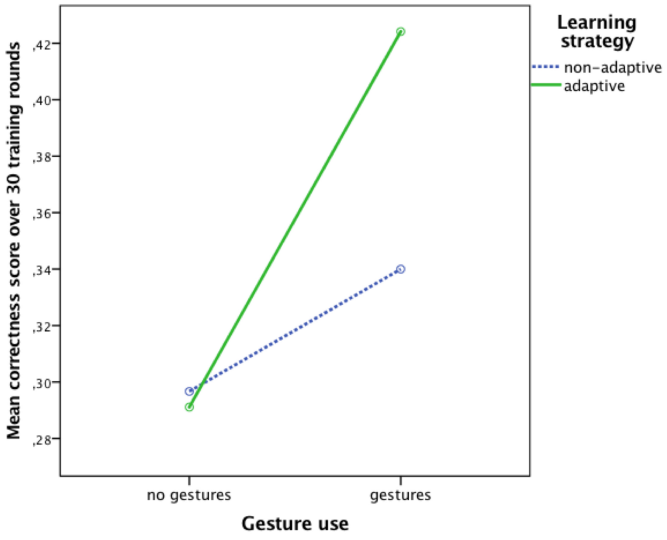

Figure 7: Interaction effects of gesture use and training strategy.

In summary, a learning gain was observable after the interaction with the robot regardless of the experimental condition.

To investigate the effects of the different conditions on learning performance during training, a two-way ANOVA was carried out with learning strategy (adaptive versus non-adaptive) and the use of gestures (gestures versus no gestures) as independent variables and performance (response accuracy) during training as the dependent variable.

For the 30 rounds of training there was a main effect of gesture use ($F(1, 57) = 18.23; p < .001, \eta^2 = .242$) such that training with gestures led to higher response accuracy than learning without gestures. The effect of learning strategy failed to reach significance ($F(1, 57) = 3.62; p = .062, \eta^2 = .060$), but there was at least a trend such that children in the adaptive condition achieved a higher response accuracy than children in the non-adaptive condition. In addition, there was a significant interaction effect between use of gestures and learning strategy ($F(1, 57) = 4.72; p = .034, \eta^2 = .076$), which has been visualized in Figure 7. Without gesture use, there was no difference whether children learned adaptively or not. In contrast, when gestures were present, children in the adaptive condition turned out to perform better than those in the non-adaptive condition. So children's learning outcome was best when gesture and adaptive training were combined.

To test whether the learning gain from pre-test to post-test was affected by the experimental conditions, another two-way ANOVA was carried out with the difference score between the tests as dependent variable. There was neither a significant main effect of learning strategy, $F(1, 57) = .00, p = .95, \eta^2 = .00$ (as in the adult-study, see Section 2.4.2), nor of gesture use, $F(1, 57) = 1.53, p = .22, \eta^2 = .026$, and also no interaction effect between both variables.

When considering the retention test results instead of the immediate post-test findings, there was still no effect of learning strategy on the learning gain after a week, $F(1, 57) = .36, p = .55, \eta^2 = .006$, However, there was a significant effect for the use of gestures, $F(1, 57) = 6.11, p = .02, \eta^2 = .097$ indicating that the learning gain between pre-test and retention test was greater when gestures were used during training ($M = 1.70, SD = 1.56$) than when no gestures were used ($M = .81, SD = 1.25$). No interaction effect was observable ($F(1, 57) = .04, p = .84, \eta^2 = .001$).

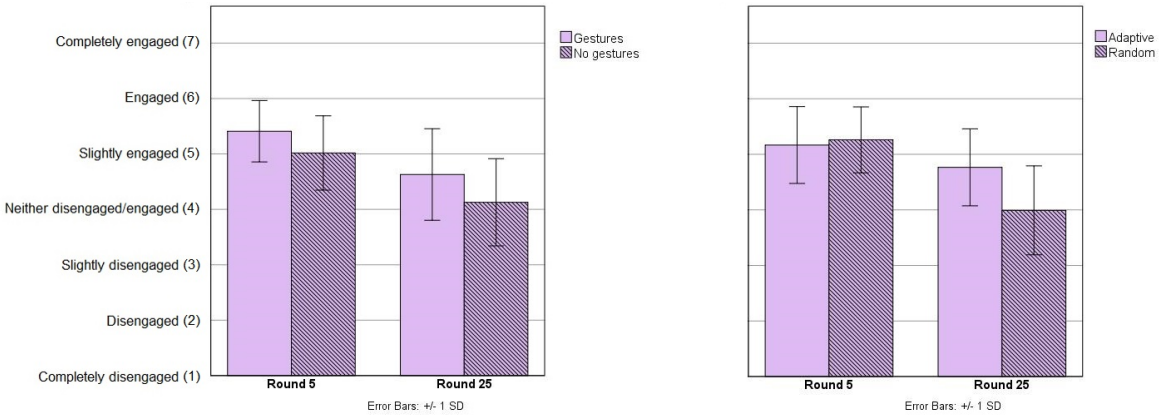

Figure 8: Rated engagement levels early and late in the training interaction for the gesture versus no gesture conditions (left) and the adaptive versus random conditions (right).

Results on Engagement

In addition to learning, the engagement of the children during the training stage with the robot was examined to find out whether children became bored and disengaged towards the end of the thirty rounds, and whether the application of an adaptive tutoring strategy and gestures help to keep children engaged. Therefore, we asked 18 adults without specific training in working with children, to evaluate the video-recordings from the experiment (without audio). For each child, one clip was taken from the fifth round of training and one clip from the twenty-fifth round, to get observations that are close to the beginning and end of the training, but far enough from these actual moments to avoid short bursts of engagement when children realize the experiment is starting or finishing. The clips start right after the robot finishes introducing the task, i.e. the point at which the turn switches to the child to provide an answer, and last five seconds.

Participants in the evaluation were asked to rate 122 clips (61 children, two clips each), in random order, on a scale from 1 (completely disengaged) to 7 (completely engaged) based on their own intuitions. As a practice round, two clips of a child that was not included in the main experiment were presented, where one example was clearly engaged and the other was clearly not engaged.

Figure 8 visualizes the data from the evaluation, in which the gathered ratings were compared between conditions using a paired-samples t-test. As a result, children's engagement significantly dropped from the fifth round ($M = 5.21, SD = .64$) to the twenty-fifth round ($M = 4.38, SD = .84$), $t(71) = -12.09, p < .001$. Furthermore, a two-way ANOVA with learning strategy and gesture use as independent factors showed no significant effect of gestures use, $F(1, 68) = 1.36, p = .25, \eta^2 = .02$, but of learning strategy, $F(1, 68) = 86.26, p < .001, \eta^2 = .559$, on engagement. The drop in engagement between round five and round twenty-five was lower in the adaptive strategy condition ($M = -.40, SD = .35$) compared to the non-adaptive (random) condition ($M = -1.27, SD = .44$). Also, no interaction effect was discovered ($F(1, 68) = .01, p = .93, \eta^2 = .00$).

When considering the average engagement level of the fifth and twenty-fifth rounds in combination, to get an idea of the overall engagement throughout the entire training session, a 2-factorial ANOVA with gesture use and learning strategy as fixed factors revealed that the overall level of engagement was significantly higher in the gesture condition ($M = 5.02, SD = .63$) than in the condition without gestures ($M = 4.57, SD = .68$), $F(1, 68) = 8.75, p = .004, \eta^2 = .114$. There was also a significantly higher engagement when an adaptive strategy was used ($M = 4.97, SD = .67$) as opposed to a random learning strategy ($M = 4.63, SD = .67$), $F(1, 68) = 5.10, p = .03, \eta^2 = .07$. No interaction effect

between the two factors emerged ($F(1, 68) = .08, p = .78, \eta^2 = .001$).

In summary, the results of this additional study conducted with children demonstrated that children actually gathered new knowledge, i.e. learned new words, through the interaction with robot and tablet. Most interestingly, it was observable that the use of iconic gestures had a positive influence on learning: training with gestures led to higher response accuracy than learning without gestures. In addition, the learning gain between pre-test and retention test was greater when gestures were used during training. Moreover, the inclusion of an adaptive learning strategy positively affected children's engagement during the interaction, although the use of an adaptive strategy had no significant impact on performance or learning gain during this (short) interaction period. It is, however, possible that adaptive strategies evolve over time and might develop a stronger impact over a longer course of learning.

Finally, an interaction effect between gesture use and learning strategy was observable with regard to the children's performance during the task: children performed equally well whether they learned adaptively or not when no gestures were present, but performed better in the adaptive condition when gestures were included. Hence, children's performance can be increased through the usage of both robot gestures and adaptive training.

Overall, the results revealed that the inclusion of gestures is beneficial for L2 word learning. Furthermore, findings with regard to children's engagement indicate that an personalized and adaptive learning interaction can hinder drops in engagement compared to a random strategy. In conclusion, we decided to integrate an engagement detector in our knowledge tracing network, so that it also can be taken into account during predictive decision making process (see Section 2.6), while the action space of the system could be extended by suitable gestures.

2.5 Modeling Interaction Patterns (T5.6)

In order to provide a smooth learning interaction for the children, different requirements have to be fulfilled. Besides a specific structure for the interaction (see Section 2.3), the chosen content as well as its presentation are crucial. Moreover, the tutoring interactions will include a tablet computer as well as a humanoid robot, which also have to be considered when designing interaction patterns.

2.5.1 Interaction Patterns gained through Empirical Studies

To get a better idea of how to design the lessons, several studies have been conducted. According to our findings from an observational study in a German Kindergarten, children prefer a specific structure included in their learning games (see Section 2.3), so that they can concentrate on the content itself. Regarding feedback we observed that educators prefer positive feedback, even if they have to correct errors (see Section 2.3.2). Similar results where found by a study conducted by partners working on WP1. They observed that educators primarily use positive feedback, but in a few cases also negative feedback was observed. However, negatuve feedback was rarly used and mainly given implicitly, e.g. by rephrasing an answer such as "It is a cat!" to "Yes, it is a dog!"One explanation might be, that the teachers do not want to demotivate young children. A robot tutor should therefore also use this kind of feedback strategy, although we plan to introduce the robot as a learning peer instead of a teacher. The role of a peer has been demonstrated to be beneficial in tutoring, not only with regard to teaching, but also with regard to the acceptance of a (not perfect) robot [12]. A peer is allowed to make mistakes and maybe say something wrong.

Regarding the content of the tutoring interactions itself, several factors are important. Pilot studies preceding experiments conducted by partners working on WP7 has shown, that the target words that should be learned have to be repeated at least 10 times to produce a good learning effect (see Deliverable

Figure 9: Screenshot from the zoo scene of the first session in the number domain in the 3D tablet game.

1.1), nevertheless the tasks should not be too repetitive. Therefore, the different lessons in each domain, including the recap lesson at the end, have different topics and different types of tasks, e.g. moving an object to its target location or selecting the right answer by touching static or moving objects in the scene. In addition, the child was required to repeat the target words, as the combination of productive and receptive learning tasks leads to higher (productive) learning gains than receptive learning tasks only [13].

To decide whether physical objects are favorable for learning, especially in the spatial relations domain, the usage of physical objects in comparison to virtual objects was investigated. Since children are used to play with physical objects in learning contexts as well as in their leisure time, it was assumed that physical objects are more suitable. What is not known is whether the manipulation of 3D models on a tablet is also sufficient. However, no significant difference were revealed regarding the words learned depending on whether physical objects or 3D graphics on the tablet computer were used (see [14]). In conclusion, virtual 3D objects are equally adequate for learning as physical objects, hence we decided to use virtual objects which are far more easy to track.

2.5.2 Interaction Patterns gained through Pilots

To test the system with the already implemented interaction patterns, we conducted two small pilots studies, in which the children were playing the first sessions of the tablet game, consisting of two different scenes in the zoo (see Figure 9). In this game, the robot and tablet guided the children trough the interaction in which they were asked to move particular objects to their target location, select an object or repeat the L2 target words. But during these first pilots, several problems emerged on a more technical level.

If the child did not listen to the robot attentively, no answer was given by her/him and (s)he was waiting for further instructions while simply looking at the robot or trying to play with the tablet. To make sure the children understand the task, it will be rephrased and presented again after a specified amount of time, e.g. after 5 seconds. If a target word has been presented the first time, and the task requires the interaction with and between several objects, e.g., moving the monkey into the cage, all relevant objects get highlighted until they were touched the first time. This way also a visual hint is given, which might help to understand the task, which is partially told in L2. Furthermore, very attentive and concentrated children often try to answer very quickly, even before hearing the full task description, as soon as all necessary information has been presented. They become impatient if the system is not fast enough for them and this resulted in boredom and decreasing motivation to play the learning game. To handle this, the system will now accept that the child starts to answer a task,

when all necessary task information has been given (e.g.: "*Now, I think there's a very important task for us! <tablet(on)> The monkey is loose and we have to put it in its cage! <**accept_answer**> Put <pointAt(tablet)> <Gaze(tablet)> the monkey in its cage.*") and all output has been made interruptible. This way the interaction can be speed up for fast children and can stay the same as before for slower children. Additionally, to hinder the children from moving the objects around when the game proceeds too slow, or when they simply want to play with the objects, we lock all movable objects until they are needed. Another problem which has shown up during the pilots was that some children were focused too much on the tablet game, so that they did not look at the robot while it was speaking and gesturing. To allow the focus of the child to switch back to the robot, the tablet screen is now turned off during important verbal or non-verbal output. Finally, some children seem to struggle with dragging an item and dropping it at the target location. Instead they loose touch on the way and hence drop it (unintentionally) in-between. This resulted in negative feedback, which nearly instantly gets interrupted, because the child did take up the item again to complete the task. To overcome this problem, we plan to provide a warm-up session for the children beforehand, so that they can familiarize themselves with the tablet and the shown 3D environment. This hopefully will reduce the frequency of losing objects during the actual teaching interaction. Furthermore, a small delay of 500ms between the answer validation and the actual feedback will be introduced, but it still has to be tested whether this delay is suitable.

Most of the mentioned issues that were discovered in initial studies and pilot tests have also been considered in our so called "Storyboard" (see Appendix B and Deliverable 2.1). The storyboard is based on scripts developed by WP1 (see Deliverable 1.1), which have been transferred to a more technical representation which is still editable from non-technical people. Thanks to that, the interaction experts in our project can create this storyboards, which can in turn be automatically translated into a machine-readable format to be used in the interaction manager and also in the output manager, later on.

2.6 Motivational-relational Strategies (T5.7)

Another important task for the interaction management is to use motivational and relational strategies in an adaptive way to maintain engagement of the child [15]. Since not only a bad task performance can influence the motivation of the child, but also tasks that are too repetitive or too boring, it is important to track the affective state of each child during the interaction, since individual differences and preferences can be expected.

To get an idea of which affective states occur and are important during child-robot tutoring, and how they can be detected based on the observation, a qualitative approach was chosen. We used video recordings from a previous study in kindergarten (see [16] for further details) and interviewed five preschool teachers on their perception and interpretation of the childrens behavior during child-robot interactions [17].

The experts were instructed that they should judge the behavior and related affective state of each child in the 4 shown videos. After each video (one video relates to one child, see Figure 10) the interviewer asked how the experts would react to negative changes in the childs state, e.g., if they recognize a lack of attention, and how this could be realized with a robot.

Our major research questions were:

- RQ1: How do experts interpret the cognitive and emotional state of children during the robot-child tutoring lessons?
- RQ2: To which behavioral cues do they refer when they remark changes (e.g., in the child's level of attention)?

Figure 10: Screenshot from one of the videos shown to the experts during the interview. The learning interaction is displayed from two perspectives.

- RQ3: How would the experts react to changes in the children's engagement from the perspective of the robot?

2.6.1 Extracted Behavioral Cues

According to the experts descriptions of the children's states, categories of states were derived, see Table 4. The analyses revealed that the childrens states can be classified into states of engagement, disengagement, and negative engagement, on a meta level (RQ1, for a more comprehensive overview see [17], Appendix D). Engagement is composed of concentration and thinking, activity and involvement, as well as expressiveness. If a child kept eye contact with the robot and tablet, and sit still, the experts interpreted their behavior as concentrated and engaged. If they mimicked the gestures the robot made, or answered verbally or nonverbally (e.g., nodding, head-shaking), they were also described as involved and thus engaged in the interaction. Likewise, expressive behaviors such as smiling, or showing a thumb up were interpreted as a sign of engagement by the experts. On the other hand, behaviors that were interpreted as signs of inattentiveness and distraction, or boredom were regarded as indicators of disengagement. For instance, rubbing eyes, gazing away, or frequent changes of the seating position were interpreted as inattentiveness. Additionally, supporting ones head with the hands, undirected tapping with the fingers, and gazing away, were (among others, cf. Table 4) named as remarkable behaviors that demonstrate boredom and disengagement. Finally, the category negative engagement contains states like skepticism and averseness. These states were related to frowning, lowering mouth corners, and head-tilt (RQ2).

When considering the frequencies with which each behavior was displayed by the children the results indicate that eye contact ($n = 4$ children), smiling ($n = 4$), and self-touches to the head ($n = 3$) were interpreted as a sign of engagement for multiple children in the video recordings. Regarding disengagement, making grimaces ($n = 4$), gazing away ($n = 7$), turning away ($n = 4$), moving the position ($n = 2$), and finger tapping ($n = 3$) were observed across several children. As a sign of negative engagement, head tilt was for several children ($n = 3$) interpreted as showing skepticism. Instead, giving verbal answers, nodding, head-shake, eye rub, frowning, and lowered mouth corners were only addressed for one child, respectively, and appear hence less informative. Note that the counts refer to the spontaneous mention of the cue per child and that the cues were overall mentioned repeatedly over the course of the interaction.

Meta-level State	**State Interpretation**	**Behavioral Cue**	**n**	**additional analyses** (non-expert coders)
Engagement	Concentration/ Thinking	eye contact	5 (4)	8/8
		sit still	2 (2)	7/8
		hand to head	4 (3)	3/8
	Involvement/ Activity	mimic robots gestures	2 (2)	1/8
		answer verbally	1 (1)	8/8
		nodding	1 (1)	7/8
		head-shaking	1 (1)	3/8
	Expressive/ Proud	smiling	7 (4)	8/8
		thumb up	1 (1)	0/8
		raise fist	1 (1)	0/8
Disengagement	Inattentiveness/ Distraction	rub eyes	2 (1)	0/8
		grimace	4 (4)	0/8
		gaze away	7 (4)	5/8
		turn away (whole body)	10(4)	0/8
		move position (stand up, lay down)	2 (2)	7/8
	Boredom/ Impatience	support the head with hand(s)	3 (2)	4/8
		move the head from left to right	2 (2)	1/8
		undirected finger tapping	4 (3)	4/8
		gaze away	2 (1)	*
		move position (stand up, lay down)	6 (4)	*
Negative Engagement	Skepticism	tilt head	3 (3)	-
	Disinterest	frown	1 (1)	-
	Averseness	lower mouth corners	1 (1)	-

Table 4: Children's States and Related Cues, where n is the frequency of reference to a cue; the amount of children for which the cue was observed is noted in parentheses

Preventive actions	**Paraphrases**	**n***
Include verbal input	It would be more motivating for the child if it should talk to the robot (expert 2, video 2)	3
Heighten robot's activity (e.g., move head)	The interaction would be more engaging if the robot moves. (expert 2, video 2)	3
Repair actions		
React to the child's behavior/ give feedback	The robot should react to the behavior of the child, e.g., tell him/her to sit down again. (expert 5, video 1)	4
Change task difficulty	The task should increase in difficulty to get the childs attention back. (expert 1, video 3)	1
Include alternative activities (e.g., play a game; stand up)	The robot could ask the child to stand up and move around, so that he/she is ready to listen again afterwards. (expert 3, video 2)	4
Allow a break	A break or a continuation at another day could be helpful to get the attention back (expert 2, video 1)	2

Table 5: Possible actions mentioned by the experts, where *n is the amount of experts out of the 5 experts that mentioned the strategy.

Additional Video-analyses by Non-Experts

To ensure that the extracted cues are not a result of the specific child-robot interaction displayed in the videos, we had additional child-robot interaction recordings analysed by project partners. For that purpose, our partners from Utrecht, coded eight additional videos of child-robot-teacher interactions, and checked whether the cues mentioned before are recognizable in these interactions, too. The video recordings included material from four girls and four boys interacting with the robot, supervised by a human teacher. The coders only noted whether a specific cue was displayed by the child during the interaction (present/absent), regardless of how often a cue was visible. The right column of Table 4 shows for how many children out of eight the cues were recognizable as signs of engagement or disengagement by non-expert observers. In the cases where one cue is listed twice due to multiple interpretations (e.g. gaze away as inattentiveness/distraction and boredom/impatience), an asterisk is inserted when the cues is listed for the second time. Regarding cues for engagement, all cues except of "thumb up" and "raise fist", were also displayed in the additional material. Especially, eye contact, sit still, answer verbally, nodding, and smiling was more or less observable in all interactions (at least 7/8). For the disengagement cues, rub eyes, grimaces and turn away with the whole body was not observed in the video recordings. These cues hence appear to be too specific, and hence can be neglected for future tracking of cues. In addition to the cues, which were extracted from the expert interviews, non-expert coders further remarked that the children in the video-recordings also pointed to the robot (7/8), and leaned forward (3/8), which was also interpreted as a sign of engagement in the interaction. However, when the child leaned on the tablet (4/8), yawned (2/8), or started playing with the tablet before s/he was allowed to (4/8) this was assumed to be a sign of boredom and thus disengagement.

In conclusion, the additional analyses support the behavioral cues of engagement and disengagement that have been extracted from the expert interviews. Furthermore, the results demonstrate that the unveiled behaviors are typically observable from children during child-robot interactions. With

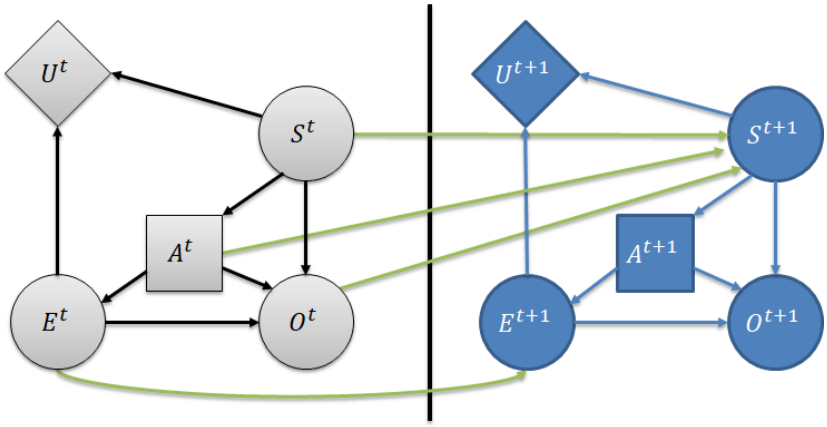

Figure 11: With the current skill-beliefs and engagement level E^t the system chooses the next skill S^t and/or action A^t for time step t to balance between maintaining engagement and raising the learning gain. If a teaching action has been carried out and after an answer O^t from the learner has been observed, this observation together with action A^t, engagement level E^t and the previous skill-belief S^t are used to update the skill-belief S^{t+1} at time $t + 1$.

respect to the interaction management it was also of interest how the robot should react if it recognizes a sign of dis-engagement. Thus, we asked the experts in the interviews how they would intervene to keep children engaged from the robots point of view (RQ3). Their suggestions were summarized into categories of potential actions to re-engage children in the tutoring with the robot (see Table 5).

2.6.2 Intervention Strategies

Parts of the experts suggestions can be regarded as preventive strategies that can be employed in the interaction from the outset. These are general strategies to keep children engaged in an interaction as allowing multi-modal interactions (here: add speech) or more expressive robot behavior (e.g., gestures, movements). Beyond that, actions were mentioned that can be useful to re-engage children in an ongoing interaction after their engagement was lowered (repair actions, see Table 5). The robot could for example suggest alternative activities to get the child's attention back (e.g., play a game). In some cases, it will even be necessary to stop the tutoring for a break according to the expert's opinions. Moreover, it was suggested that the difficulty of the task should be increased if signs of disengagement are recognizable.

Additional Video-analyses by Non-Experts

When re-considering the video-recordings of the eight additional child-robot-teacher interactions (see above), it has been observed that the teachers intervened in the interaction at some points when one of the cues was present, namely:

- the child gazes away → teacher leans forward to attract attention
- the child shows heightened activity → teacher says "pay attention"
- the child starts playing on tablet before introduction by robot is finished → teacher says "first you have to listen"

In conclusion, the analyses of the expert interviews as well as the additional analyses by non-experts demonstrated that specific behavioral cues are displayed by children during child-robot interactions, that can be categorized as signs of engagement and disengagement. Fortunately, the majority of these cues can be easily recorded via existing technologies like Microsoft Kinect (e.g. activity, gaze).

Hence, our next step will be the usage of the information from the expert interviews to build a classifier for "interaction engagement". As first approach it is planned to use a Naive Bayesian Classifier to check whether all cues (gaze direction, smiling and activity) could be tracked and combined into a model and how strong the influence of each cue on the interaction engagement is. Since sensor data can be noisy, we also will include cues like accuracy, correctness of an answer and the response time.

After the classifier has been evaluated with annotated videos which are accompanied by recorded Kinect- and interaction data, the classified "interaction engagement" can be included into the "Adaptive-BKT" model (see Figure 11).

This enables the system to choose not only the next skill and action from which the learner will most likely benefit but also to balance teaching by maintaining the "interaction engagement". Therefore the action space will be extended. Some possible actions have been mentioned in the expert interviews and can be seen in table 5. In fact, actions like allowing a break or include alternative activities already have been shown to improve concentration during learning sessions [18].

3 Conclusion

In summary, the goal of WP5 is the development of an interaction manager for the L2TOR system, which is responsible for planning the course of an second language learning interaction by choosing appropriate actions based on its internal Knowledge-, Affect- and History-Models as well as pre-designed interaction patterns. Therefore, this component has to (1) receive/send multimodal input/output, (2) interpret the input, and finally (3) decide which action should be performed as a reaction to the actual state. In the following we briefly highlight what has been accomplished with regard to the steps 1-3 so far, and what still needs to be addressed in the future.

1. **Input/Output (Task 5.1)**
   Interfaces to all other modules have been defined and implemented, which will be extended to fit the changing requirements of new developed interaction mechanism (e.g., new types of tasks or new robot behaviors as allowing the robot to move an object from position A to position B) and patterns (e.g., allow the robot to help if the child get stuck in the interaction).

2. **Interpretation (Tasks 5.2, 5.4, 5.7)**
   For the interpretation of the input information during the learning interaction three models are important: (1) a model to keep track of the knowledge state of the child, from which a first version already has been developed and evaluated (T5.4), (2) a model to keep track of the affective state of the child, which is currently under development based on expertise from teachers of German kindergartens (T5.7), and (3) a model to keep track of the longer-term interaction-, knowledge- and affect-history to build up a common ground and inform the short-term tracing models (1) and (2) to allow for further adaptation to the child (e.g., slow learner vs. fast learner). The latter is currently under development and we are collecting information about which data will be needed to enable long-term adaptation as well as an evaluation of the full system (T5.2).

3. **Decision Making (Task 5.5; based on Tasks 5.3, 5.4, 5.6, 5.7)**
   A basic framework has been implemented to combine all collected and inferred information, to

plan the next steps of the tutoring interaction, based in them. This will be our main focus in the final year, the main challenge being to combine flexible decision-making with the designed interaction patterns (T5.3, T5.6) to provide a smooth interaction, while maintaining positive engagement and maximizing the learning gain. A first approach has been implemented, following a predefined interaction structure (T5.3), which already takes decisions to maximize the learning gain based on a predicted knowledge state of the learner (T5.5). As a next step we will extend this approach by including the *Engagement* of the learner and actions to regulate this engagement, which will enable the system to maintain a positive engagement (T5.7).

References

[1] Mark G Core and James Allen. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56. Boston, MA, 1997.

[2] P. Lison and C. Kennington. Opendial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Demonstrations)*, pages 67–72, Berlin, Germany, 2016. Association for Computational Linguistics.

[3] Gabriel Skantze and Samer Al Moubayed. Iristk: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 69–76. ACM, 2012.

[4] Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. In *Proceedings of ACM/IEEE HRI 2017*, pages 128–136. ACM Press, 2017.

[5] Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1994.

[6] Sebastian Leitner. *So lernt man lernen. Der Weg zum Erfolg.* Herder, Freiburg, 1972.

[7] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. Affect and learning: An exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, 29(3):241–250, 2004.

[8] Lev Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA, 1978.

[9] Manuela Macedonia, Karsten Müller, and Angela D. Friederici. The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping*, 32(6):982–998, 2011.

[10] Kirsten Bergmann and Manuela Macedonia. *A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance*, pages 139–148. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[11] Jan de Wit, Thorsten Schodde, Kirsten Bergmann, Emiel Krahmer, and Stefan Kopp. The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies. 2018. Submitted to ACM/IEEE HRI 2018.

[12] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, 19(1):61–84, 2004.

[13] Jan-Arjen Mondria and Boukje Wiersma. Receptive, productive, and receptive+ productive l2 vocabulary learning: What difference does it make. *Vocabulary in a second language: Selection, acquisition, and testing*, 15(1):79–100, 2004.

[14] Rianne Vlaar, Josje Verhagen, Ora Oudgenoeg-Paz, and Paul Leseman. Comparing L2 Word Learning through a Tablet or Real Objects: What Benefits Learning Most? In *Proceedings of the Workshop R4L at ACM/IEEE HRI 2017*, 2017.

[15] Tony Belpaeme, Paul E Baxter, Robin Read, Rachel Wood, Heriberto Cuayáhuitl, Bernd Kiefer, Stefania Racioppa, Ivana Kruijff-Korbayová, Georgios Athanasopoulos, Valentin Enescu, et al. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

[16] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. Exploring the Effect of Gestures and Adaptive Tutoring on Childrens Comprehension of L2 Vocabularies. In *Proceedings of the Workshop R4L at ACM/IEEE HRI 2017*, 2017.

[17] Thorsten Schodde, Laura Hoffmann, and Stefan Kopp. How to Manage Affective State in Child-Robot Tutoring Interactions? 2017.

[18] Aditi Ramachandran, Chien-Ming Huang, and Brian Scassellati. Give me a break!: Personalized timing strategies to promote learning in robot-child tutoring. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 146–155. ACM, 2017.

A Interface Specification

A.1 Output

Module	Function	Description	Parameters	Used?
ControlPanel	memoryLoaded	Indicator-event that the memory has been successfully created and/or loaded	ID of the memory which has been loaded	X
TabletGame	loadScene	Sends the content of the predefined scene-file to be loaded on the tablet	scene specification in json format	X
	hintObject/hintRObject	Add/Remove a glow to highlight objects	List of object-IDs in json format	X
	showMap	Go back to the town-map	-	X
	showObject/hideObject	Make objects visible or invisible	List of object-IDs in json format	X
	makeMovable/makeStatic	Make objects movable for the user or static again	List of object-IDs in json format	X
	enableObject	Enable objects for clicking/touching	List of object-IDs in json format	X
Underworlds	loadScene	Sends the content of the predefined scene-file to be loaded on the tablet	scene specification in json format	X
OutputManager	load_session	Sends the ID of the next lesson to be taught	lesson-ID	X
	set_child_name	Send the name and ID of the child to be used in the verbal output and the log-files	Name and ID in json format	X
	give_task	Next task should be given	Task-ID	X
	give_feedback	Give feedback to the child	Valid and task-type in json format (not fully specified yet)	X
	request_answer	Request an answer from the child	Task-type and task related information like the objects to be used in the task in json format	X
	interrupt_output	Interrupt any kind of output	-	X
	give_break	Give a break or a break filling activity like a puzzle or a dance	not yet specified	
	give_help	Provide some verbal or non-verbal help	not yet specified	
	grab_attention	Try to grab the attention back if the child is inattentive	not yet specified	

A.2 Input

Module	Function	Description	Parameters	Used?
ControlPanel	loadMemory	Load memory of child with ID X	Filename	X
	createMemory	Create a new memory for the child	ID, Name	X
	CPinit	Init message from Controlpanel to load all lesson files, send them around and start the interaction	Name of lesson file	X
KinectModule	vadStart/vadStop	Start/End of detected voice	-	X
	setTrackingData	Childrens' gaze direction (looking at robot, tablet or somewhere else), a normalized value how active (moving) the child is and a value of the happiness detection	gaze-direction, activity, happiness	
TabletGame	touchDown/touchUp	Start/End of touch (including the position and touched object id)	Object-ID, 3D-Position	X
Underworlds	updtSpRel	Update of the spatial relations	Relevant spatial relations from all object to all others in json format	X
OutputManager	give_task_completed	Send when the robot finished the task description	-	X
	feedback_completed	Send when the robot finished giving feedback	-	X
	request_answer_completed	Send when the robot finished requesting an answer from the child	-	X

B Clipping of the Storyboard for the First Session of the First Domain

	A	B	C	G	H	I	J	K
1			Robot		Tablet			
2	#	Input (touch and speech)	Text L1 English	Text L2	Scene	Objects	Say	Comment
29	16		<tablet(on)><Gaze(tablet)> Cool, elephants! Touch them <pointAt(tablet)> and we'll <Gaze(child)> hear the English word for elephant.					
30	17	objectTouched, voiceActivity, support	<giveResponseToSelectObject(elephant)>				elephant	did the child select the correct object?
31	18		<Gaze(child)><tablet(off)>Ah, an elephant is in English an	elephant				
32			Can you also say	elephant				
33	19	objectTouched, voiceActivity, support	<giveResponseOnSpeech(elephant)>					did the child say something?
34	20		<tablet(on)>let's see <Gaze(tablet)> what we have to do now.					
35			<Gaze(child)> The elephants are loose and we have to put them in their cage! Put	one elephant				
36		probabilityGesture	in its cage. <pointAt(tablet)> <Gaze(tablet)> .			highlight elephant_1 and cage_2		
37	21	objectTouched, voiceActivity, support	<giveResponseToMoveObject(elephant, in, cage_2)>			display stars	<happy_sound>	did the child move the correct object? Elephants go in their cage elephant and the elephand makes a happy elephant sound, child and robot receive a star
38	22		<Gaze(child)> there is still	one elephant		remove stars		
39		probabilityGesture	outside of the cage. Add it to the cage <Gaze(tablet)> and we <Gaze(child)> will hear what "add" is					
40	23	objectTouched, voiceActivity, support, target	<giveResponseToMoveObject (elephant, in, cage_2)>				add	did the child select the correct object?
41	24		<tablet(off)>	add				
42			say	add				
		objectTouched, voiceActivity, support						did the child say

Chart1 | in

C HRI 2017 Paper

Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making

Thorsten Schodde
CITEC, Bielefeld University
Bielefeld, Germany
tschodde@techfak.uni-bielefeld.de

Kirsten Bergmann
CITEC, Bielefeld University
Bielefeld, Germany
kirsten.bergmann@uni-bielefeld.de

Stefan Kopp
CITEC, Bielefeld University
Bielefeld, Germany
skopp@techfak.uni-bielefeld.de

ABSTRACT

In this paper, we present an approach to adaptive language tutoring in child-robot interaction. The approach is based on a dynamic probabilistic model that represents the interrelations between the learner's skills, her observed behaviour in tutoring interaction, and the tutoring action taken by the system. Being implemented in a robot language tutor, the model enables the robot tutor to trace the learner's knowledge and to decide which skill to teach next and how to address it in a game-like tutoring interaction. Results of an evaluation study are discussed demonstrating how participants in the adaptive tutoring condition successfully learned foreign language words.

CCS Concepts

•Computing methodologies → Probabilistic reasoning; Cognitive robotics; •Applied computing → Interactive learning environments; •Human-centered computing → *Empirical studies in HCI;*

Keywords

Language tutoring; Education; Assistive robotics; Bayesian Knowledge Tracing; Decision making

1. INTRODUCTION

The use of robots for educational purposes has increasingly moved into focus in recent years. This is due to two major developments. First, robots became cheaper and more robust so that applications in everyday environments are now conceivable. In particular, technology has matured up to a point where intuitive interaction using natural language or gesture has become feasible. Second, the need for second language learning becomes increasingly important, and empirical evidence has demonstrated that learning with and from a physically present, interactive robot can be more effective than learning from classical on-screen media [14, 15, 20, 22]. In fact, recent research showed that performance can increase up to 50% [17]. It can, hence, be assumed that tutoring using social robots is qualitatively different from alternative digital tutoring technologies. Nowadays, first practical applications can be found, e.g. in nursery where toy robots teach the alphabet to kids in a very simple way. More generally, findings from a variety of settings seem to suggest that robots can help small children to develop in an educational setting [10, 18, 24, 27].

In the L2TOR project[1], we investigate in how far a social robot can support children at pre-school age with respect to second language learning. Learning a language is a very complex task. It involves not only acquiring vocabulary, but also learning prosodic features, syntactical structures, semantic meanings as well as situation-dependent language use. Yet, it has been argued that social robots can create the interactive environment and motivational experience needed to learn languages [19].

One of the most important aspects in tutoring is the robot's ability to keep track of the knowledge state, i.e. the learned and not-yet-learned skills, of the child interacting with it. This information is indispensable to enable a personalized tutoring interaction and to optimize the learning experience for the child [27]. The tutor has to structure the tutoring interaction, choose the skills to be trained, adjust the difficulty of the learning tasks appropriately and has to adapt its verbal and non-verbal behaviour.

The importance of personalized adjustments in the robot's behaviour has been substantiated in recent research showing that participants who received personalized lessons from a robot (based on heuristic skill assessment) outperformed others who received a non-personalized training [22]. Suboptimal robot behaviour (e.g. too much, too distracting, mismatching or in other ways inappropriate) can even hamper learning [17]. In this paper we present an integrated approach for tracing the knowledge of the learner during a L2 learning interaction together with a strategic adaptation of tutoring actions.

In the following, we discuss related work in Section 2. In Section 3 an extension of Bayesian Knowledge Tracing is presented as well as a model to select the next tutoring actions based on the predicted effects they may have on the learner's knowledge state. This model has been implemented in a robot that provides language tutoring in a game-like fashion. Section 4 introduces the empirical basis for this scenario and observational studies on language tutoring in kindergarten. Section 5 presents an evaluation study carried out with this robot and Section 6 discusses the results.

[1] http://www.l2tor.eu

HRI '17, March 06 - 09, 2017, Vienna, Austria

ISBN 978-1-4503-4336-7/17/03…$15.00

DOI: http://dx.doi.org/10.1145/2909824.3020222

2. RELATED WORK

Numerous studies have investigated the effects of social robots in tutoring scenarios. Empirical evidence demonstrates that learning with and from a physically present, interactive robot can be more effective than learning from classical on-screen media [14, 15, 20], and that robots can help children to develop in educational settings [10, 18, 24, 27]. However, at the same time, it is found that suboptimal behaviour of the robot can hamper learning [17]. Thus, a crucial ingredient for successful robot tutoring is the ability to provide personalized lessons [22] and to adapt in appropriate ways to the needs of the learner. The key question is when and how to adapt robot tutoring, according to which adaptation strategies, and based on what features of the state of learner or the tutoring interaction.

2.1 Approaches to Adaptive Tutoring

In the realm of Intelligent Tutoring System (ITS), dedicated *pedagogical modules* are employed for planning an optimal path through the curriculum by using an internal model of the learner's present knowledge state (cf. [8]). Cakmak and Lopes [3], for example, proposed a teaching algorithm that selects the most informative demonstrations for the learner. This learning agent makes use of Inverse Reinforcement Learning (IRL) to reduce the learner's hypothesis space of possible reward functions as fast as possible. In an evaluation, the authors showed that a learner trained with non-optimal selected expert demonstrations require significantly more demonstrations to achieve a similar performance as the optimally taught learner. This system, however, is designed for a sequential decision task in which no uncertainty about the learner's knowledge/skill exists. This assumption does not hold for the domain of L2 learning where the learner's current state of knowledge can, at best, be inferred from observed behaviour. Another important limitation of this approach is a lack of flexibility as no adaptation towards students' individual needs is considered.

Addressing especially the issue of adaptation towards students' individual needs, Partial Observable Markov Decision Processes (POMDPs) have been employed as basis for the pedagogical module of an ITS. Rafferty et al. [25], for instance, proposed different algorithms for planning an action-policy based on a POMDP and compared these against two different random and a maximum information gain (MIG) choice. They showed that even a simple action-policy based on a POMDP can achieve a significant faster skill learning than choosing actions randomly. But compared to the simple MIG algorithm, no significant difference was observed. Only with increasing skill space the MIG algorithm seems not to be sufficient anymore. A likely explanation for this finding is that the knowledge tracing model is insufficient. In addition, finding a good policy based on a POMDP is often computational intractable.

Clement et al. [4] compared two algorithms choosing the next skill and action in a tutoring interaction against a lesson given by a human expert. Both algorithms based on prior knowledge, e.g. the impact of actions on the learning gain or the difficulty of different types of tasks, which had been annotated by experts beforehand. The algorithms differed with regard to the adaptation method and the amount of additional knowledge stored besides the prior. The authors showed that even if the ITS does not make use of an internal model to store beliefs about the child's knowledge state regarding a specific skill, the use of their algorithm can lead to a higher learning gain compared to an expert lesson. Furthermore their second proposed algorithm, which additionally stores information about the knowledge state of the child, performed even better. Clement et al. concluded that extending their system with a more complex model for tracing the knowledge state of a student might lead to a higher learning.

An often criticized issue in this line of research is the lack of an effective *knowledge-tracing* method in the pedagogical module of an ITS that could be profitable for the learning interaction, e.g. by increasing the students' learning gain. Hence, we review research on knowledge tracing methods in the following.

2.2 Knowledge Tracing

Knowledge tracing aims to model learners' mastery of the knowledge being tutored. An often used approach is *Bayesian Knowledge Tracing* (BKT). BKT is a specific type of *Dynamic Bayesian Networks* (DBN), or more precisely, of Hidden Markov Models consisting of observed and latent variables. The latent variables represent the 'skills' and are classically assumed to be binary. That is, a skill is represented to be mastered or not. Generally, separate BKT networks are used for each skill to be learned [5]. Belief update is based on the observation of an answer to a given task testing a specific skill. The observed answer is binary too. Further, BKT models have two types of parameters: The emission probability and the transition probability. The emission probabilities are given by the 'guess probability' $p(guess)$, the probability of answering correctly without knowing the skill, and the 'slip probability' $p(slip)$ of answering wrongly although knowing the skill. In contrast, the transition probabilities are given by $p(t)$, the skill transition from unknown to known, and $p(f)$ the probability of forgetting a previously known skill. Often $p(f)$ is assumed to be zero.

Spaulding et al. [29] recently adopted BKT to trace the language-reading skill of children in robot-based language tutoring. They proposed the 'Affective BKT model', which is characterized by two further observable variables called 'smile' and 'engagement' to take into account the affective state of the child. This model structure allows emotions to influence the belief-state of each skill as they are included in every belief-update. The authors showed that the affective state of the children can be successfully integrated into BKT and that this approach outperforms traditional models for tracing the knowledge state in learning situations [29].

Another modification of BKT was published by Käser et al. [16]. Instead of using a dedicated BKT for every skill, they defined one comprehensive DBN to trace the knowledge on all skills to be learned. This enables to trace the knowledge on each skill individually and, in addition, to represent and reason with skill inter-dependencies. This allows for searching some kind of order in which skills may be learned best. The authors could demonstrate that this more detailed model outperforms other traditional models of knowledge tracing, including the normal BKT, with regard to the accuracy of the skill belief [16].

Finally, Gordon et al. [11] recently presented a so-called 'active learner model' to trace the word-reading skills of small children. This model does not work on the basis of BKT but employs a simple distance metric to approximate the conditional probability $p(w_2|w_1)$ of whether the child

can read a word w_2 if it already knows the word w_1. Their evaluation showed that their system is able to adapt to users of different age and to trace their reading knowledge up to a certain extent [11].

In this paper we present an expandable model based on BKT for knowledge tracing that, in contrast to the systems reviewed above [16, 29], allows for the simulation of actions and decision-making in teaching interactions.

3. ADAPTIVE LANGUAGE TUTORING

As a basis for our approach to adaptive language tutoring, we adopt the Bayesian Knowledge Tracing model [5] which has been successfully employed in other work and was shown to be easily extensible. However, we modify and extend the BKT model in order to enable predictive decision-making based on the represented beliefs about the learner's knowledge state. In this section, we first introduce our version of BKT and then present the approach for decision-making.

3.1 Bayesian Knowledge Tracing

The traditional approach to BKT uses only one latent variable S to represent the skill belief and one observable variable O for the user's answer. This suffices to represent if a skill is mastered or not, and how probable it would be that the user will answer correctly. Also, this information can be used to choose the next skill to learn, e.g. the skill which has the lowest belief probability of having been mastered. However, this model does not include information about how a skill can be addressed for teaching. In consequence, there is no possibility to take possible actions and their influence on the update of skill beliefs into account. We thus add a decision node A for actions to the Bayesian network (see Figure 1). This node not only influences the possible observation but also the belief update in the next time step. Further, we use a latent variable S that can attain six discrete values for each skill, corresponding to six bins for the belief state (0%, 20%, 40%, 60%, 80%, 100%). This allows for a more detailed model of the impact of tutoring actions on the possible observations and skills. Moreover, it becomes possible to better quantify the robot's uncertainty about the learner's skill.

With these changes, especially the conditional probability table $p(O^t|S^t)$ and the additional influence of the action A^t on the observable (now $p(O^t|S^t, A^t)$), the classical BKT update function, which was based on simple assumptions about guessing $p(guess)$ and slipping $p(slip)$ during the answer process, cannot be applied anymore. Instead, we apply a normal Bayesian update rule for the conditioning of skill beliefs including a transition probability $p(S_i^{t+1}|s_k, O^t, A^t)$ where s_k identifies a bin of the skill S_i^t. As a simplification we substitute this probability with $p(S^{t+1}|s_k)$:

$$\begin{aligned} p(S_i^{t+1}) &:= p(S_i^{t+1}|O^t, A^t) \\ &= \sum_{s_k \in S_i^t} [p(S_i^{t+1}|s_k, O^t, A^t) \cdot p(s_k|O^t, A^t)] \\ &\approx \sum_{s_k \in S_i^t} [\frac{p(O^t|s_k, A^t) \cdot p(A^t|s_k) \cdot p(s_k)}{p(O^t, A^t)} \cdot p(S_i^{t+1}|s_k)] \end{aligned}$$

3.2 Predictive Decision-Making

The extended BKT model is used to decide which tutoring action the robot should take next. At first, the skill to address with the next tutoring action is chosen. For this, the Kullback-Leibler divergence (KLD) between the current skill belief and the desired skill belief is used, the latter being a maximally certain belief in a maximally high skill of the learner:

$$next\_skill = \underset{\forall S_i^t \in \mathbb{S}}{\operatorname{argmin}}[\alpha(S_i^t) \cdot KLD(p(S_i^t), p(S_{opt}))]$$

$\mathbb{S}$ represents the set of all skills that can be addressed, which consists of all words to be taught to the user. $p(S_{opt})$ is the desired belief for each skill, which means 99.999% of probability mass in the last bin (100%). The factor $\alpha(S_i^t)$ has been added for each skill to regulate the skill occurrence frequency. It is decreased each time the skill is addressed, and it is increased if another skill is being practised. In this way, the skill-selection algorithm takes care of the maximization of each skill belief as well as the balancing of all skills.

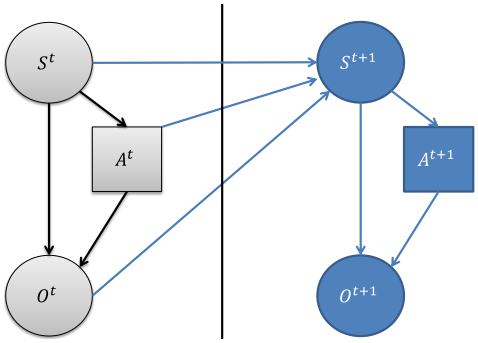

Figure 1: Dynamic Bayesian Network for BKT: The action node A^t predicts the observation O^t and influences the belief update of S^t for the next time step $t+1$.

After the skill has been chosen, the next step is to decide with which tutoring action this should be done. Here, we consider abstract tasks as tutoring actions. These tasks will have to be mapped onto concrete exercises or pedagogical acts at a later stage in the robot control architecture (see Section 4). For simplicity, we distinguish between tutoring actions according to the difficulty (easy, medium or hard) of the task that addresses the corresponding skill. Finding the best action a_l for a given skill S_i^t is thus a minimization problem of the following form:

$$next\_action = \underset{\forall a_l \in A^t}{\operatorname{argmin}}[\alpha(a_l) \cdot KLD(p(S_i^{t+1}), p(S_{opt}))]$$

where

$$\begin{aligned} p(S_i^{t+1}) &:= p(S_i^{t+1}|a_l) \\ &= \sum_{s_k \in S_i^t} \sum_{o_j \in O^t} p(S_i^{t+1}|o_j, s_k, a_l) \cdot p(o_j, s_k|a_l) \\ &\approx \sum_{s_k \in S_i^t} p(s_k|a_l) \sum_{o_j \in O^t} p(S_i^{t+1}|s_k) \cdot p(o_j|s_k, a_l) \end{aligned}$$

with

$$p(o_j, s_k|a_l) = p(o_j|s_k, a_l) \cdot p(s_k|a_l)$$

Here, $p(S_i^{t+1})$ could be seen as predicting the effect of applying the current action a_l to the skill S_i, where we again

substitute the transition probability $p(S_i^{t+1}|o_j, s_k, a_l)$ with $p(S_i^{t+1}|s_k)$ regarding simplicity. In addition, here again the skill belief is compared with $p(S_{opt})$ which represents the desired tutor belief state for each skill. The factor $\alpha(a_l)$ provides a more detailed selection of the "best" action. This way, the model will select an easy task if the skill is believed to be low, a hard task if it is high, and medium in-between. The goal of this strategy is to create a feeling of flow which can lead to better learning results [2, 7, 12]. Thus, it strives not to overburden the learner with too difficult tasks nor to bore him with too easy tasks, both of which may lead to frustration and thus hamper the learning [9, 13].

4. ROBOT LANGUAGE TUTORING

The adaptive model as described in the previous section has been brought to application in a child-robot second language (L2) tutoring game on the basis of empirical data from adult-child language tutoring interactions.

4.1 Empirical Basis

To design a tutoring interaction that matches children's needs, we decided to design the interaction on an empirical basis of language tutoring data. We collected video recordings of language tutoring games as they take place in kindergartens. Given that 1:1 interactions of educator and child can hardly be realized in kindergartens, the games typically involve one educator and a small group of children. Data of four language tutoring games have been collected: reading a picture book together with children in an interactive manner; card game "I spy with my little eye"; card game "I'm giving you a present"; and a rhyming game. The children were between four and six years of age. The data collected comprises round about 681 min of video data. These video data have been transcribed and annotated with regard to the following categories:

- **Dialogue acts**: Utterances are classified due to the underlying intention based on the DAMSL annotation scheme [6].
- **Children's mistakes**: Types of language errors the children made, e.g. wrong plural form, missing articles, wrong syntax, etc.
- **Educator's speech repair**: Pedagogical acts used to correct the errors, e.g. reformulation, corrected repetition, etc.
- **Nonverbal behaviour**: Nods, smiles, gestures etc. used by the educators.

On the basis of these annotations, we identified some overall patterns to inform the detailed design of the robot's behaviour. These fall basically into two categories, (i) overall interaction structure and (ii) feedback behaviour by the educators.

4.1.1 Overall Interaction Structure

A common pattern in all language tutoring games under investigation was the following basic course of actions:

1. **Opening:** Marks the beginning of the interaction and should mitigate the children's timidity as well as it should motivate the child.
2. **Game Setup:** This step is used to prepare the game by explaining the task and clarify the necessary terms.
3. **Test run:** A test run of the game is conducted to test whether the instructions have been understood and to practice the game flow.
4. **Game:** Here, the main interaction game takes place. Every move is accompanied by the educator's feedback and motivations to continue.
5. **Closing:** Marks the end of the learning interaction. Additionally, it is used to ensure motivation for future interactions by acknowledge the participation, joint singing a goodbye song and an outlook on what's going to happen next time.

4.1.2 Educator's Feedback Behaviour

In addition, we analysed the educators' behaviour when providing children with feedback. An important and common pattern is that language errors are almost never corrected explicitly. Instead, feedback is always provided in a positive way, falling into one of the following categories with the percentage of their occurrence given in squared brackets: (i) **praising the child** for a correct utterance whereby praise is often combined with a repetition of the correct word [13%] (ii) **implicit correction** in case of an error made by the child: repetition of the word as if correct (e.g. correct pronunciation, with article, plural form, etc.) [54%], (iii) **correct recasting of a sentence**, e.g. after syntax errors [32%], (iv) **moving on to next task**, e.g. when the child's message is unclear due to incomprehensible pronunciation [1%]. All kinds of educators' feedback behaviour is typically accompanied by looking at the child, smiling and nodding.

4.2 Game Setup

We have chosen the game "I spy with my little eye..." as a paradigm for our child-robot language tutoring game. The robot – in the role of a tutor, assisting the child in learning novel L2 vocabularies – is acting as 'the spy'. The child-robot setting is further enriched with a tablet PC on which objects are displayed (see Figure 2). In addition, the tablet's touch-screen displays three buttons to enable further user input in terms of 'yes' and 'no' answers as well as the option to let the robot repeat its previous statement.

A basic move of the game is structured as follows: It starts with a set of objects being displayed on the tablet screen and the robot saying "I spy with my little eye, something that is ...", followed by a foreign language word that refers to a property of one of the items on the screen. The child's task is now to respond by selecting the object referred to via touch input on the tablet. The robot's feedback behaviour in response to a correct or false answer is realized on the basis of our empirical data (see Section 4.1.2). That is, the robot responds to correct answers by praising the learner as well as repeating the L2 word and the corresponding L1 translation. In case of a false guess by the child, the robot explains the correct meaning of the to-be-learned word one more time. In addition, the wrongly chosen object as well as the actually correct object are both displayed on the tablet screen and the child is asked to select the correct object. The overall game structure is framed by the other elements making up typical language tutoring games in adult-child interaction (see Section 4.1.1).

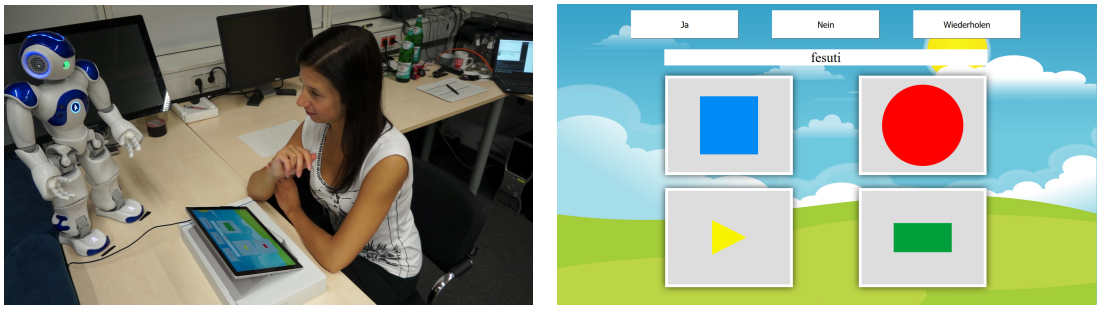

Figure 2: Experimental setup (left) with a participant sitting in front of a tablet displaying the graphical user interface (right). The robot *Nao* stands next to the tablet slightly rotated towards the user.

4.3 Technical Realization

We employed the Nao robot[2] for our language tutoring game. It is standing in a bit more than 90 degrees rotated, to the right of the participant. In addition a Microsoft Surface Pro 4[3] tablet PC is used to catch the user input and to display the graphical user interface realized via a HTML website. For the implementation of the interaction and dialogue structure, the state-chart based dialogue-manager IrisTK has been used [28]. NAOqi[4] has been applied as middleware between the robot, the graphical user interface, the dialogue manager, and our developed adaptive tutoring model. NAOqi is shipped with each Nao robot and allows to communicate via a simple event system between various programming-languages (Python, Java, C++, JScript).

5. EVALUATION STUDY

To assess the effects of our adaptive model on L2 word learning, we set up an evaluation study based on the language tutoring game described in the previous section. The major objective behind this study was to evaluate the effects of the adaptive model on learners' performance. We used the Nao robot to deliver all task information and direct feedback to the learner. This enables us to test the model within the desired final setting, including the effects of a robot's presence in the tutoring interaction. Given that children show a high degree of inter-individual variation and might further need child-specific adaptations of, for instance, synthesized speech to enable them to understand what the robot says, we decided to conduct this first study with adult learners.

We employed a between-subjects design with a manipulation of training type: Participants learned L2 vocabulary items either with the fully adaptive model, or in a random control condition. In the adaptive condition, the skill to be taught and the action to address the skill were chosen by the model as described in Section 3. In our language tutoring game, skill relates to the foreign language words and action refers to the specific task used in the game (target word, objects displayed). The difficulty of the actions/tasks in this study were implemented by using less or more distractor objects that were shown together with the correct object on the screen. For instance, an easy task consisted of two distractor objects, whereas a hard task had four distractors. Distractors were chosen with respect to the skill beliefs of the user, with the set of objects mainly consisting of items for which the L2 words were still/mostly unknown by the learner.

As shown by Craig et al. [7], better learning performance is to be expected when learners have to expend the right amount of cognitive effort (i.e. not too hard or too easy tasks). Accordingly, while learning with our model in the adaptive condition, no hard tasks are shown until the system believes the user to have basic knowledge on all skills. Then, the system will increase task difficulty (as determined by the adaptive tutoring model) by adding distractor objects. Note, however, that at a certain point the user will know too many skills/words so that finding a distractor set (i.e. task difficulty) that cannot be sorted out by exclusion becomes impossible. In the control condition, all skills are taught in a random order and always with 'medium' task difficulty.

Participants' performance was assessed with two measures: (1) we tracked learners' response behaviour over the course of the training to investigate the progress of learning, (2) we conducted a post-test on the taught vocabulary in the form of both L1-to-L2 translations and L2-to-L1 translations to assess participants' state of knowledge subsequent to the intervention.

5.1 Materials

The training materials for the study comprised German–'Vimmi' word pairs. Vimmi is an artificial language created for experimental purposes [23] that aims to avoid associations with other known words or languages. The Vimmi items are created according to Italian phonotactic rules. Ten items have been chosen: four colour terms, four shape-encoding terms and two terms describing size (see Table 1).

5.2 Procedure

Upon entering the lab, participants were randomly assigned to one of the two experimental conditions. They were informed that they take part in an experiment on foreign language learning and were asked to sign an informed consent form. They also filled out a questionnaire that covered personal information like age and nationality as well as a personal estimation of language learning skills in general and memorization ability for L2 vocabulary.

[2]https://www.ald.softbankrobotics.com/en/cool-robots/nao
[3]https://www.microsoft.com/surface/en-gb/devices/surface-pro-4
[4]http://doc.aldebaran.com/2-1/naoqi/

N	German	Vimmi	English translation
1	blau	bati	blue
2	grün	uteli	green
3	gelb	dirube	yellow
4	rot	fesuti	red
5	rund	beropuga	round
6	dreieckig	pewo	triangular
7	quadratisch	tanedila	square
8	rechteckig	paltra	rectangular
9	klein	kiale	small
10	groß	ilado	big

Table 1: The 10 words from Vimmi to be learned in the evaluation study with its corresponding translation in German as well as English for comprehension purposes.

Next, a list of the to-be-learned Vimmi items were presented to the participants for 30 seconds. This was to enable participants to practice the items right from the first game interaction on. Then, the learning interaction with the Nao robot began. After introducing itself, the robot explained the "I spy with my little eye"-game and started a test-run with the participants. Once this test run was finished and the participants agreed that (s)he understood the game, the main interaction consisting of a total of 30 trials (game moves) began. Each trial addressed one vocabulary item as described in Section 4.2. That is, the robot asked for one of the objects displayed on the tablet screen, whereby the question was in L1 (German) for the most part, except for the referring, to-be-learned word in L2 (Vimmi). After 30 trials, the game was finished, the Nao robot thanked the participants and said goodbye.

Subsequent to the interaction with the robot, participants' learning performance was assessed with a post-test. In an interview with the experimenter, they had to translate the ten to-be-learned vocabulary items from German to Vimmi and likewise from Vimmi to German (both in randomized order). The whole interaction and the vocabulary-post-test at the end of the study were recorded with an external camera. Also the system decisions taken during the interaction and the probability distributions for each updated skill belief were logged.

5.3 Participants

A total of 40 participants (20 per condition) with an average age of 24.13 ($SD = 3.82$) took part in this study (16 males and 24 females). All participants had very good command of the German language and normal or corrected sight. All of them were paid or received credits for their participation.

5.4 Results

5.4.1 *Learning Progress During Training*

In order to assess the learners' progress during training, we compared the number of correct responses addressing the initial quarter of the tutoring game (first seven items) against the final quarter (last seven items). When an item occurred repeatedly within the initial quarter, the first occurrence has been taken into account. When an item occurred repeatedly within the final quarter, the last occurrence has been considered.

	Adaptive (A)		Control (C)		A, C	
	M	**SD**	**M**	**SD**	**M**	**SD**
F7	3.75	1.37	4.00	1.17	3.88	1.27
L7	6.90	0.31	5.15	1.69	6.03	1.49
F7, L7	5.33	0.69	4.58	1.12		

Table 2: Means (M) and standard deviations (SD) of correct answers for the initial quarter of the training interaction (first seven items – F7) and the final quarter (last seven items – L7) in each condition, as well as the inter-model (A, C) and intra-model (F7, L7) means and standard deviations.

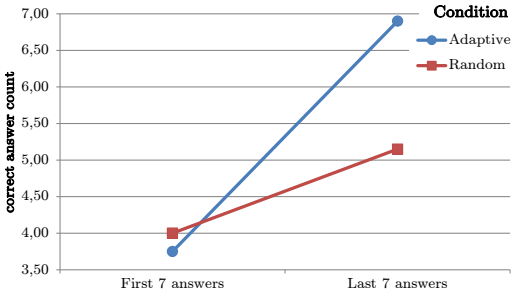

Figure 3: Mean numbers of correct answers at the beginning (first 7) and end (last 7) of the interaction in the different conditions.

A mixed-design ANOVA with training phase (initial, final) as a within-subjects factor and training type (adaptive-model-based, control) as between-subjects factor has been conducted. Results are summarized in Table 2 and Figure 3. Not surprisingly, there was a main effect of training phase at a significant level ($F(1, 38) = 66.85, p < .001, \eta^2 = .64$): Learners' performance was significantly better in the final phase as compared to the initial phase. In the first quarter of training, participants achieved a mean of 3.88 ($SD = 1.27$) correct responses, whereas in the final quarter, a mean of 6.03 ($SD = 1.49$) items was selected correctly. More interestingly, there was also a main effect of training type ($F(1, 38) = 6.52, p = .02, \eta^2 = .15$) such that participants who learned in the adaptive condition had a higher score of correct answers ($M = 5.33, SD = .69$) as compared to learners in the control condition with an average of $M = 4.58$ ($SD = 1.12$) correct answers. Finally, the interaction between training phase and training type was also significant ($F(1, 38) = 14.46, p = .001, \eta^2 = .28$) indicating that the benefit of adaptive-model-based training develops over time (see Figure 3). While participants' response behaviour in the first quarter of training was similar across conditions, a benefit of training with the adaptive model became evident in the final quarter. At this stage of training, participants in the adaptive model condition achieved a mean of $M = 6.9$ ($SD = .31$) correct responses, whereas participants in the control condition achieved a mean of $M = 5.15$ ($SD = 1.69$) correct responses.

	Adaptive (A)		Control (C)	
	M	**SD**	**M**	**SD**
German-to-Vimmi	3.95	2.56	3.35	1.98
Vimmi-to-German	7.05	2.56	6.85	2.48

Table 3: Results of both post-tests (German-to-Vimmi and Vimmi-to-German): Means (M) and standard deviation (SD) of correct answers grouped by the experimental conditions.

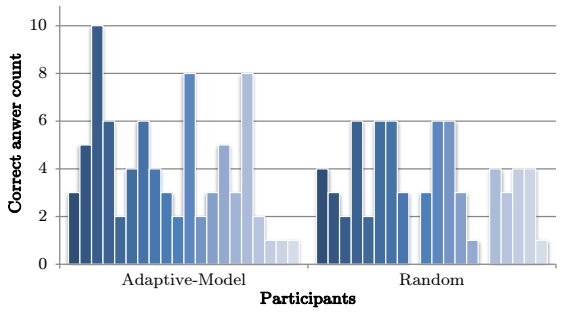

Figure 4: Participant-wise amount of correct answers grouped by the different conditions for the German-to-Vimmi post-test.

5.4.2 Post-Test

Participants' learning performance subsequent to the intervention has been measured with two translation tests (L2-to-L1 and L1-to-L2). Results are summarized in Table 3. Paired-samples t-tests were conducted to compare the number of correctly recalled words after training with the adaptive model as compared to training in the control condition. For the German-to-Vimmi translation, there was no significant main effect ($T(38) = .25, p = .80$). Participants who trained with the adaptive-model recalled a mean of 3.95 ($SD = 2.56$) out of ten words correctly, while participants in the control condition recalled a mean of 3.35 ($SD = 1.98$) words. Likewise, there was no significant main effect ($T(38) = .83, p = .41$) for the Vimmi-to-German translation task. Participants' performance after learning with the adaptive model amounted to a mean of 7.05 ($SD = 2.56$) correct items, participants' performance in the control condition to a mean of 6.85 ($SD = 2.48$) correct items.

Although no main effect of training type emerged in the post-test, some details might nevertheless be worth mentioning. In the German-to-Vimmi post-test, a maximum of ten correct responses was achieved by participants in the adaptive-model condition, whereas the maximum on the control condition were six correct answers. Moreover, there were two participants in the control condition who did not manage to perform any German-to-Vimmi translation correctly. In the adaptive-model condition, all participants achieved at least one correct response (see Figure 4).

6. CONCLUSION

In this paper we have presented a novel approach to personalize language tutoring in human-robot interaction. This adaptive tutoring is enabled through a model of how tutors mentalize about learners – by keeping track of their knowledge state and by selecting the next tutoring actions based on their likely effects on the learner. This is realized via an extended model that combines Bayesian Knowledge Tracing (of the learned) with tutoring actions (of the tutor) in one causal probabilistic model. This allows, for selecting skills and actions based on notions of optimality – here the desired learner's knowledge state as well as optimal task difficulty – to achieve this for a given skill. This model has been implemented into a robot language tutoring game and tested in a first evaluation study.

The analysis of participants' response behaviour over the course of training has clearly shown that participants learned the L2 words during the human-robot interaction. Importantly, they learned more successfully with our adaptive model as compared to a randomized training. That is, the repeated trials addressing still unknown items as chosen by the adaptive model (until the belief state about these words equalled that of known items) outperformed the tutoring of the same material (same number of trials and items) but in randomized order. In the post-test, however, there was no significant difference across experimental conditions, despite a trend towards better performance in the adaptive model conditions over the controls.

Different explanations may account for this inconsistent finding. One potential explanation could be that the way how responses were prompted was not identical in training sessions and post-test. In the training sessions, participants saw pictures reflecting the meaning of the to-be-learned words whereas in the post-test they just received a linguistic cue in form of a word they had to translate. It might be that repeated trials as they were particularly supported for difficult-to-remember items by the adaptive model, led to stronger associations between linguistic and imagistic materials. This might have caused a stronger decline of correct responses for participants who trained with the adaptive model as opposed to those in the control condition. An alternative explanation could be that test results measured immediately after the training session are subject to strong inter-individual differences among learners. This is the reason why studies on vocabulary learning usually range over repeated sessions spread over several days (cf. 1). A typical pattern is that significant results emerge after two or three sessions/days and/or in the long-term (measured several weeks after training took place). So it might well be that further training sessions or delayed tests might result in a post-test performance that matches the picture of the during-session performance.

One might argue that the performance of our adaptive model is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system [21]. In this system flashcards are sorted into groups according to how well the learner knows each one. Learners try to recall items written on a flashcard. If they succeed, the card is sent to the next group. If they fail, the card is sent back to the first group. Each succeeding group has a longer period of time before the learner is required to revisit the cards. This way all items, that are hard to remember for the learner will be repeated more often. In contrast to such spaced repetition systems, our model is more flexible as it can vary the difficulty of the tasks by providing more or less distractor items. In addition, we plan a more comprehensive action space of the model to account for motivating actions

where necessary or adaptations in the robot's verbal or non-verbal behaviour.

Overall, results from the evaluation study are, at least, in parts very promising: learners' performance during training was significantly improved by personalized robot tutoring based on the adaptive model. Nevertheless, the fact that this positive effect did not hold in the post-test, inter alia, marks a starting point for further refinements of the model: Training stimuli should be designed such that they match the way language learners need to apply them best possible. That is, when the aim is to enable people to translate words from one language to another, training stimuli should provide cues for this process of mapping linguistic materials on each other. Moreover, a further study with more learning sessions (e.g. over several days as common in many vocabulary studies) should be conducted. Regarding the model itself, we plan to incorporate skill-interdependencies as well as to take the affective user state into account, too. Both kind of extensions have been shown to improve learning [16, 29]. Additionally, the model can (and is meant to) provide a basis for exploiting the full potential of an embodied tutoring agent. Regarding this, we plan to advance the model such that the robot's verbal and non-verbal communicative behaviour is adapted to the learner's state of knowledge and progress. Specifically, we aim to enable dynamic adaption of (i) embodied behaviour such as iconic gesture use to be known to support vocabulary acquisition as a function of individual differences across children (cf. [26]); (ii) the robot's synthetic voice to enhance comprehensibility and prosodic focusing of content when needed; and (iii) the robot's socio-emotional behaviour depending on the learners' current level of motivation or engagement. Further, as the long-term goal of our work is to enable robot-supported language learning for pre-school children, another important goal is to make children-specific adaptations to the language game and test it in child-robot interaction studies.

7. ACKNOWLEDGEMENTS

This work was supported by the L2TOR (www.l2tor.eu) project supported by the EU Horizon 2020 Program, grant number: 688014, and by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

8. REFERENCES

[1] K. Bergmann and M. Macedonia. *A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance*, pages 139–148. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[2] D. E. Berlyne. Conflict, arousal, and curiosity. 1960.

[3] M. Cakmak and M. Lopes. Algorithmic and human teaching of sequential decision tasks. In *AAAI Conference on Artificial Intelligence (AAAI-12)*, 2012.

[4] B. Clement, D. Roy, P.-Y. Oudeyer, and M. Lopes. Multi-armed bandits for intelligent tutoring systems. *arXiv preprint arXiv:1310.3174*, 2013.

[5] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[6] M. G. Core and J. Allen. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56. Boston, MA, 1997.

[7] S. Craig, A. Graesser, J. Sullins, and B. Gholson. Affect and learning: An exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, 29(3):241–250, 2004.

[8] C. Dede. A review and synthesis of recent research in intelligent computer-assisted instruction. *International Journal of Man-Machine Studies*, 24(4):329–353, 1986.

[9] S. Engeser and F. Rheinberg. Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion*, 32(3):158–172, 2008.

[10] M. Fridin. Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Comput. Educ.*, 70:53–64, Jan 2014.

[11] G. Gordon and C. Breazeal. Bayesian active learning-based robot tutor for children's word-reading skills. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 1343–1349. AAAI Press, 2015.

[12] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593, 2013.

[13] M. J. Habgood and S. E. Ainsworth. Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *The Journal of the Learning Sciences*, 20(2):169–206, 2011.

[14] J. Han, M. Jo, S. Park, and S. Kim. The educational use of home robots for children. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 378–383, Aug 2005.

[15] E. ja Hyun, S. yeon Kim, S. Jang, and S. Park. Comparative study of effects of language instruction program using intelligence robot and multimedia on linguistic ability of young children. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pages 187–192, Aug 2008.

[16] T. Käser, S. Klingler, A. G. Schwing, and M. Gross. *Beyond Knowledge Tracing: Modeling Skill Topologies with Bayesian Networks*, pages 188–198. Springer International Publishing, Cham, 2014.

[17] J. Kennedy, P. Baxter, and T. Belpaeme. The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 67–74, New York, NY, USA, 2015. ACM.

[18] J. Kory and C. Breazeal. Storytelling with robots: Learning companions for preschool children's language development. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 643–648, Aug 2014.

[19] J. Kory Westlund, G. Gordon, S. Spaulding, J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Learning a second language with a socially assistive robot. In *The 1st International Conference on Social Robots in Therapy and Education*, 2015.

[20] H. Kose-Bagci, E. Ferrari, K. Dautenhahn, D. S. Syrdal, and C. L. Nehaniv. Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*, 23(14):1951–1996, 2009.

[21] S. Leitner. So lernt man lernen: Der weg zum erfolg [learning to learn: The road to success]. *Freiburg: Herder*, 1972.

[22] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati. The physical presence of a robot tutor increases cognitive learning gains. In *CogSci*. Citeseer, 2012.

[23] M. Macedonia, K. Müller, and A. D. Friederici. Neural correlates of high performance in foreign language vocabulary learning. *Mind, Brain, and Education*, 4(3):125–134, 2010.

[24] J. R. Movellan, M. Eckhardt, M. Virnes, and A. Rodriguez. Sociable robot improves toddler vocabulary skills. In *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*, pages 307–308, March 2009.

[25] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching via pomdp planning. *Cognitive Science*, 2015.

[26] M. L. Rowe, R. D. Silverman, and B. E. Mullan. The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemporary Educational Psychology*, 38(2):109–117, 2013.

[27] M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse. Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1613–1622, New York, NY, USA, 2010. ACM.

[28] G. Skantze and S. Al Moubayed. Iristk: A statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 69–76, New York, NY, USA, 2012. ACM.

[29] S. Spaulding, G. Gordon, and C. Breazeal. Affect-aware student models for robot tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, pages 864–872, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems.

D ICCT 2017 Paper

How to Manage Affective State in Child-Robot Tutoring Interactions?

Thorsten Schodde
CITEC, Bielefeld University
Bielefeld, Germany
tschodde@techfak.uni-bielefeld.de

Laura Hoffmann
CITEC, Bielefeld University
Bielefeld, Germany
lahoffmann@techfak.uni-bielefeld.de

Stefan Kopp
CITEC, Bielefeld University
Bielefeld, Germany
skopp@techfak.uni-bielefeld.de

***Abstract*—Social robots represent a fruitful enhancement of intelligent tutoring systems that can be used for one-to-one tutoring. The role of affective states during learning has so far only scarcely been considered in such systems, because it is unclear which cues should be tracked, how they should be interpreted, and how the system should react to them. Therefore, we conducted expert interviews with preschool teachers, and based on these results suggest a conceptual model for tracing and managing the affective state of preschool children during robot-child tutoring.**

I. Introduction

The use of robots for educational purposes has increasingly moved into focus in recent years. One rationale is to enable individually adapted one-to-one teaching for weaker students, which can hardly be provided in regular classrooms. This idea already underlay educational on-screen applications like intelligent tutoring systems (ITSs). Physically present social robots are expected to bring an additional quality to the learning interactions, similar to co-present teacher-child or child-child interaction, which can make the tutoring experience more effective. Indeed, a recent study showed that students' learning performance increases up to 50% if a social robot was included compared to a classical on-screen media learning [1].

One of the main challenges for robot tutors is to identify the learner's internal states, e.g., whether she is following, distracted, or losing motivation. Yet, recognizing and reacting to these cognitive and affective states is vital to keep the learner engaged and to foster learning. In previous work, we developed an approach to dynamically adapt robot tutoring to the changing pedagogical state of the learner [2]. There, the skill mastery of the student is kept track of inferentially using Bayesian Knowledge Tracing, which enables the robotic tutor to choose the to-be-addressed skill and difficulty of the next task accordingly. This way the model works to keep the child in the "zone of proximal development" [3], which can lead to a feeling of flow, motivation and better learning [4], [5].

However, this approach lacks "emotional intelligence" [6]. Successful human teachers not only teach the curriculum according to the learner's knowledge state, but also manage the affective states of children. Studies have shown that affective states like curiosity, interest, flow, joy, boredom, frustration and surprise can influence learner's problem-solving abilities, and affect task engagement and learning motivation [7]. Further, such states are found to influence cognitive processes like long-term memorizing, attention, understanding, remembering, reasoning, decision-making and the application of knowledge in task solving [4], [8]. It is thus not surprising that good human tutors are sensitive to learners' vocal (e.g., intonation) or nonvocal behavior (e.g., facial expression, body language) [9]. Technical systems are also increasingly able to recognize most of these cues - albeit sometimes in a quite rudimentary way. However, little attention has been paid to the question how a robot should interpret and respond to the affective state of a learner during tutoring with the needed flexibility and adaptiveness [10], [11].

In this paper we present steps towards a model for tracing and managing the affective state of preschool children in second language tutoring interactions with a robot tutor. This model is based on pedagogical knowledge about children's affective states during actual robot-child tutoring gathered through expert interviews with preschool teachers. This knowledge comprises information about which affective states are relevant, from which features they can be tracked and, finally, how to react to them appropriately as a tutor. It lends itself to a decision-theoretic affective state tracing model that can be combined with our previously developed adaptive knowledge tracing approach. The following section discusses previous work on affect detection and affective tutoring systems. Afterwards we present the procedure and results of the conducted expert interviews. Finally, we discuss how these findings can be incorporated into a conceptual model that enables the recognition of and reaction to changes in children's affective states.

II. Related Work

A. Affect Detection

A lot of work has been done on affect recognition based on different modalities. One widely used approach is the analysis of facial expressions to detect the affective state of a user [12]. Often, classifiers are trained on "very expressive and played" emotions, making their applicability to real-world interactions questionable. In fact, the accuracy of emotion detection based on facial features is often low in real-world applications. Furthermore, the recognition rate is strongly dependent on the expressiveness of each target.

An alternative approach is the detection of affect from the user's voice [13]. Classifiers based on voice analysis are

trained on datasets of spontaneous speech, so that they are more suitable for real-world applications. With regard to robot-child tutoring, affect detection through speech analysis is, however, difficult because speech input is often not included as speech recognition for children has a low accuracy [14]. Other attempts have been made to detect the affective state through analyzing written text [15]. This approach includes, for instance, analyzing the usage of adjectives and adverbs. But in most natural interactions humans do not write text, and preschool children are usually not able to read and write.

A broader approach for affective state detection is the tracking of the whole body posture and movements by using a body pressure mat laying on a seat [16], or using a Microsoft Kinect [17]. A limitation is that the use of a body pressure mat assumes that the user remains on a seat and cannot move around. The Kinect, however, allows the user to move around, but may have problems in detecting smaller events like small postural shifts. Also, approaches based on human physiology have been adopted. In this realm, measures such as ECG, EEG, EMG [18], [19], and brain imaging [20] have been applied to "read" the affective state from the user's body. The results of these methods are promising, however the applicability of such obtrusive approaches (e.g., wires and patches on the body) in tutoring interactions with children is clearly limited.

In sum, all of these approaches have their field of use, but also their limitations. In contrast, multi-modal approaches have been studied to overcome these limitations and to increase accuracy of the detection. A lot of combinations exist, e.g., facial expressions and voice [21], facial expression, voice and body posture [22], facial expressions, body postures and context dependent activity logs [23], or speech and text [24]. Such systems demonstrated that a multi-modal approach to detect affective states results in higher accuracy rates.

B. *Affective Tutoring Systems (ATSs)*

Since the technical progress yields new possibilities to make use of the affective state in tutoring interactions, a lot of systems have been extended with such a module. Shen et al. [25], for instance, used physiological signals for affect detection and then guided the learning interaction by different affective strategies. Their results demonstrated the superiority of an emotion-aware over a non-emotion-aware system with a performance increase of 91% .

Alexander et al. [26] developed an affect-detecting ITS including a virtual agent for primary school students. The affective state is detected by analyzing the facial expressions of the student and serves as the basis for a case-based selection of the next tutoring actions. The case-based rules have been informed by an observational study of human tutors. In a study conducted in a primary school, where children had to solve mathematical equations, the use of their affective system showed a significant increase of the students' performance as compared to a control group without affective support.

The "Affective AutoTutor" system [27] can automatically detect boredom, confusion, frustration and neutral affect by monitoring conversational cues and discourse features along with gross body language and facial features. Cues provided by each channel are combined to select a single affective state, based on which AutoTutor responds with empathic, motivational, or encouraging dialog-moves and emotional displays. Evaluations showed that this systems is able to support learners not only in acquiring knowledge, but also in using it in transfer tasks later on. Recently, Goren et al. [28] incorporated affect detection via facial expressions in robot-child tutoring. In a study with preschool children they showed that their system personalized its policy over the course of training, and that children who interacted with the personalized robot showed increased long-term positive valence as compared to a control group without personalization.

Taken together, the findings from earlier approaches suggest the inclusion of affect detection in robot-child tutoring. Most affect detectors are trained on specifically annotated data to identify the important cues for each affective state. For example, the emotion classifier "Affectiva Affdex" [29] is trained on more than 5 million human faces to classify facial expressions. Strategies for how to respond to those states are usually based on observational studies of the reactions of a human tutor to the behavior of a student [30]. We adopt this approach here, too, with the aim of building a model that enables a robot to detect changes in children's learning-relevant affective states and to react to these changes appropriately. For this, child-robot interaction specific knowledge is necessary that could be best gathered from experts in reading and managing the affective states of young children in tutoring interactions, namely, preschool teachers.

III. Empirical Basis

With the aim of answering the questions, which affective states occur and are important during robot-child tutoring, and how they can be detected based on the observation of a child, a qualitative approach was chosen. We used video recordings from a previous study in kindergarten and interviewed five preschool teachers on their perception and interpretation of the children's behavior.

A. *Participants*

A total of five female preschool teachers were invited and interviewed as experts. They were between 36 - 61 years old ($M = 48.6; SD = 8.16$) and had a working experience from 16 to 42 years ($M = 29; SD = 8.88$).

B. *Materials*

With the objective of allowing the experts to observe children during robot-child tutoring in a controlled manner, video recordings from an interaction study were used. They were presented and discussed during face-to-face interviews with one interviewer. In total, video recordings of eight different children (4 female, 4 male), which varied in their level of activity and expressiveness when facing the robot, were chosen. The decision was taken to ensure that individual difference are considered in spite of the small samples. The recordings were taken in the realm of a separate study in Dutch

Fig. 1. Screenshot from one of the videos shown to the experts during the interview. The learning interaction is displayed from two perspectives.

preschools were children were tutored to learn animal names in a foreign language by means of a "I spy with my little eye..." game with a Nao robot. Here, up to four images of animals were displayed on a tablet screen, while the robot is referring to one of them using a Dutch description and the English name of the animal [31]. To choose the animal the robot mentioned, the children had to tap on the picture on the tablet. Two camera perspectives were recorded and presented to the experts to allow a frontal view on the child, but also a landscape view from the side on the whole experimental setup which includes the robot, the tablet and the child (see Fig. 1).

C. Procedure

At the beginning of each interview session, the participants were informed about the purpose and the procedure of the interview and signed an informed consent that their voice was recorded. They were instructed that they should judge the behavior and related affective state of children, which are presented in video recordings. First, a small example video was presented, which had to be commented by the experts to make sure the task was clear. Then, the interviewer started the video on a laptop and asked the expert to comment on the child's behavior and state. After each video (one video relates to one child) the interviewer asked how the experts would react to negative changes in the child's state, e.g., if they recognize a lack of attention, and how this could be realized with a robot. At each point in time, the interviewees were allowed to pause the video and go back to review a scene. Each expert discussed a total of four videos with the interviewer. Afterwards they were thanked for their participation and dismissed.

D. Analyses and Results

The whole interview session were recorded by means of a computer microphone, and a screen capture tool to synchronize the comments with the video recording that was played at the time. The recordings were afterwards transcribed to enable detailed content analyses of the experts' comments. The transcripts were then analyzed regarding the following research questions:

TABLE I
CHILDREN'S STATES AND RELATED CUES

Meta-level State	State Interpretation	Behavioral Cue	n*
Engagement	Concentration/ Thinking	eye contact	5 (4)
		sit still	2 (2)
		hand to head	4 (3)
	Involvement/ Activity	mimic robots gestures	2 (2)
		answer verbally	1 (1)
		nodding	1 (1)
		head-shaking	1 (1)
	Expressive/Proud	smiling	7 (4)
		thumb up	1 (1)
		raise fist	1 (1)
Disengagement	Inattentiveness/ Distraction	rub eyes	2 (1)
		grimace	4 (4)
		gaze away	7 (4)
		turn away (whole body)	10(4)
		move position (stand up, lay down)	2 (2)
	Boredom/ Impatience	support the head with hand(s)	3 (2)
		move the head from left to right	2 (2)
		undirected finger tapping	4 (3)
		gaze away	2 (1)
		move position (stand up, lay down)	6 (4)
Negative Engagement	Skepticism	tilt head	3 (3)
	Disinterest	frown	1 (1)
	Averseness	lower mouth corners	1 (1)

*n is the frequency of reference to a cue; the amount of children for which the cue was observed is noted in parentheses.

- RQ1: How do experts interpret the cognitive and emotional state of children during the robot-child tutoring lessons?
- RQ2: To which behavioral cues do they refer when they remark changes (e.g., in the childs level of attention)?
- RQ3: How would the experts react to changes in the children's engagement from the perspective of the robot?

According to the experts descriptions of the children's states, categories of states were derived. As listed in Table I, the childrens states can be classified into states of engagement, disengagement, and negative engagement, on a meta level (RQ1). Engagement is composed of concentration and thinking, activity and involvement, as well as expressiveness. If a child kept eye contact with the robot and tablet, and sit still, the experts interpreted their behavior as concentrated and engaged. If they mimicked the gestures the robot made, or answered verbally or nonverbally (e.g., nodding, head-shaking), they were also described as involved and thus engaged in the interaction. Likewise, expressive behaviors as smiling, or showing a thumb up were interpreted as a sign of engagement by the experts. On the other hand, behaviors that were interpreted

as signs of inattentiveness and distraction, or boredom, were regarded as indicators of disengagement. For instance, rubbing eyes, gazing away, or frequent changes of the seating position were interpreted as inattentiveness. Additionally, supporting ones head with the hands, undirected tapping with the fingers, and gazing away, were (among others, cf. Table I) named as remarkable behaviors that demonstrate boredom and disengagement. Finally, the category negative engagement contains negative states like skepticism and averseness. These states were related to frowning, lowering mouth corners, and head-tilt (RQ2).

Each interaction with the robot varied according to individual differences of the children (e.g., age, self-confidence). Hence, we counted for each behavioral cue, how many times it was mentioned by different experts for different children. If two experts observed a cue for one child as relevant it was counted as two; but if one expert mentioned one cue for one child several times it was counted as one. To reflect on the occurrence of the cues over different children, it was further listed for how many different children the cue was observed (see Table I numbers in parentheses).

The results indicate that eye contact ($n = 4$ children), smiling ($n = 4$), and self-touches to the head ($n = 3$) were interpreted as a sign of engagement for multiple children in the video recordings. Regarding disengagement, making grimaces ($n = 4$), gazing away ($n = 7$), turning away ($n = 4$), moving the position ($n = 2$), and finger tapping ($n = 3$) were observed across several children. As a sign of negative engagement, head tilt was for several children ($n = 3$) interpreted as showing skepticism. Instead, giving verbal answers, nodding, head-shake, eye rub, frowning, and lowered mouth corners were only addressed for one child, respectively, and appear hence less informative. Note that the counts refer to the spontaneous mention of the cue per child and that the cues were overall mentioned repeatedly over the course of the interaction.

Furthermore, we asked the experts how they would intervene to keep children engaged in the interaction from the robots point of view (RQ3). Their suggestions were summarized into categories of potential actions to re-engage children in the tutoring with the robot (Table II).

Parts of the experts suggestions can be regarded as preventive strategies that can be employed in the interaction from the outset. These are general strategies to keep children engaged in an interaction as allowing multi-modal interactions (here: add speech) or more expressive robot behavior (e.g., gestures, movements). Beyond that, actions were mentioned that can be useful to re-engage children in an ongoing interaction after their engagement was lowered (repair actions, see Table II). The robot could for example suggest alternative activities to get the child's attention back (e.g., play a game). In some cases, it will even be necessary to stop the tutoring for a break according to the expert's opinions. Moreover, it was suggested that the difficulty of the task should be increased if signs of disengagement are recognizable.

TABLE II
POSSIBLE ACTIONS MENTIONED BY THE EXPERTS

Preventive actions	**Paraphrases**	**n***
Include verbal input	It would be more motivating for the child if it should talk to the robot (expert 2, video 2)	3
Heighten robot's activity (e.g., move head)	The interaction would be more engaging if the robot moves. (expert 2, video 2)	3
Repair actions		
React to the child's behavior/ give feedback	The robot should react to the behavior of the child, e.g., tell him/her to sit down again. (expert 5, video 1)	4
Change task difficulty	The task should increase in difficulty to get the childs attention back. (expert 1, video 3)	1
Include alternative activities (e.g., play a game; stand up)	The robot could ask the child to stand up and move around, so that he/she is ready to listen again afterwards. (expert 3, video 2)	4
Allow a break	A break or a continuation at another day could be helpful to get the attention back (expert 2, video 1)	2

*n is the amount of experts out of the 5 experts that mentioned the strategy.

E. Discussion

In summary, the analyses of the expert interviews revealed that preschool teachers agree on the interpretation of several child behaviors as signs of (dis-)engagement. The behavioral cues that were identified during robot-child tutoring were changes in gaze direction (eye contact versus gaze away), body posture (turn away, stand up, lay down), or facial expressions (smiling). These cues that have been identified can be used to narrow down the feature space in affective state recognition. We note, though, that the small amount of video samples restricts the significance of our findings. However, a frequent, independent naming of the most relevant cues by different experts for different children points to the importance of these cues for detecting the affective state of children. Interestingly, the majority of these cues can be recorded by means of non-obtrusive technologies (e.g., video cameras, Microsoft Kinect) and can be extracted using existing tools (e.g., Affdex, see above). Building on this, the following section lays out a conceptual approach to interpret and respond to changes in the child's state during robot-child tutoring interactions.

IV. Affective State Management Model

A. Tracing the Affective State

The first step is to combine the different cues mentioned in Section III into higher-level states and to trace them over time. As a first approach, this can be achieved using a naive Bayesian classifier that determines the hidden internal state E that is assumed to independently cause cues $C_1, C_2, ..., C_n$. Since cues need to be integrated into coherent belief updates over time, the corresponding belief must be updated every time step according to a dynamic Bayesian model $P(E^{t+1}|C_i^{t+1}, E_t)$.

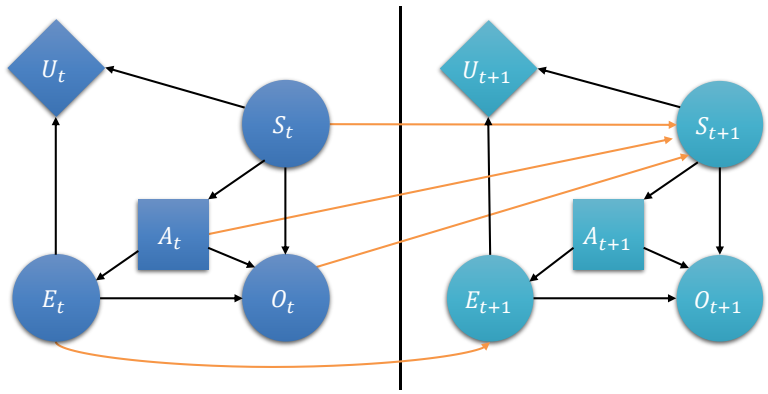

Fig. 2. Here the adaptive Bayesian Knowledge Tracing model is shown, consisting of the belief regarding the mastery of a skill S_t, the observation (response) O_t to an action A_t, the affective state E_t of the learner and the expected value U_t of a chosen chain of actions.

Variables E and C_i are directly based on the results of the expert interviews. We focus on the most reliable and explicit cues that can be tracked with current technology. Thus we base the model on those cues that were frequently mentioned for several children (cf. Table I). Since most cues from the negative engagement group were only mentioned once, and "head tilt" is difficult to track due to the danger of mixing it up with moving the head from side to side (from the disengagement group), we focus on signs of engagement and disengagement in the first stage of the model's development. Engagement and disengagement can be regarded as opposing poles on a continuum of engagement. Hence, we combine them into the meta state variable E_t that is called *interaction engagement*. Cues that were identified as indicating engagement will have a positive effect on this state, while all cues related to disengagement will have a negative impact.

B. Managing the Affective State

After computing the belief update for interaction engagement, the next step is to determine whether and how the robot tutor should act. To this end, we include the belief variable E into our previously developed approach based on Bayesian Knowledge Tracing [2] (see Fig. 2). According to this model, the belief over the learners mastery of a certain skill S_t explains the observed answer O_t to a given teaching-task A_t selected to address this skill S_t. We add the state variable E_t as well as an utility value U_t, which represents the expected value of a chosen chain of pedagogical and affective actions. E_t is assumed to influence the students answer to a task, e.g. if the student is disengaged there may be a higher probability of observing a wrong answer as she may not have understood the task description. This information will also affect the belief update for the currently addressed skill, so that a wrong answer will have a lower impact when the student is disengaged.

Although experts' agreed on the identification of the behavioral cues, the interpretation of these cues should be regarded carefully since one behavior could have distinct meanings depending on the situation and the specific child. For the realization of a general model, the expert information is useful to determine which cues are relevant to look at as a starting point. A final system must, however, be able to adapt to specific variations in the child and the situation.

Next, we need to extend the action space of A_t to actions that manage the affective state, in addition to the already present actions of addressing a certain skill with a particular task. This allows evaluating and weighing both options, teaching a skill or managing the affective state of a student. Still, the main goal is to find an action (or action sequence) from which the child will learn the most. Since the model is a Dynamic Bayesian Decision Network, this evaluation can be carried out across several time steps, where each additional time step lowers the utility gained on the basis of the increase of the skill belief. Hence, the system can decide whether it is more beneficial to first raise interaction engagement, before teaching the next skill, or the other way around.

Again, we based our selection of actions to manage affective state on the results of the expert interviews (cf. Table II). We consider only the repair actions here, out of which the change of task difficulty is already implemented in the model. Three other actions remain, which could be useful to re-engage a child in the interaction: First, directly addressing the child's behavior, e.g., urge to sit down again or ask for attention; secondly, using alternative tasks or activities to provide a more variable interaction, e.g., ask to move around or to play a game; finally, if the interaction engagement drops significantly, the robot can propose a break and the interaction can be resumed later. All of these behaviors can be immediately included in the model as well as the robot's behavior repertoire. Note, however, that the conditional probabilities $P(E|A)$ ans $P(O|A,E,S)$ need to be defined heuristically as long as sufficient interaction data is not available.

V. Summary

The present paper addressed the importance of coping with a learner's affective state during preschool child-robot tutoring. While the automatic recognition of cues seems to

be within reach with today's technology, we are still lacking a model of which affective states are most relevant in such learning interactions, how they can be recognized, and how they should be responded to by the robot tutor. To tackle this problem, expert interviews with preschool teachers have been conducted to identify children's affective states that are relevant during robot-child tutoring. The results suggest that different categories of engagement states seem to be most important, and that experts recognize and address those states in interaction. The findings from the interviews are currently used to inform the implementation of a computational model for tracing and managing the affective and cognitive state of a child learner with a robot tutor. To this end, we have laid out how to extend a previously developed knowledge-tracing and decision-making model based on a dynamic Bayesian Decision Network. The combined model will allow for finding an action policy that combines informative and affective actions of a robot tutor to manage the internal states (both, cognitive and affective) of a child learner more thoroughly, and to ensure an optimal course of learning.

Acknowledgment

We thank our colleagues from Tilburg University who provided the videos for the interviews. This work was supported by the L2TOR (www.l2tor.eu) project supported by the EU Horizon 2020 Program, grant number: 688014, and by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

References

[1] J. Kennedy, P. Baxter, and T. Belpaeme, "The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning," in *Proceedings of ACM/IEEE HRI 2017*. ACM, 2015, pp. 67–74.

[2] T. Schodde, K. Bergmann, and S. Kopp, "Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making," in *Proceedings of ACM/IEEE HRI 2017*. ACM Press, 2017, pp. 128–136.

[3] L. Vygotsky, *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press, 1978.

[4] S. Craig, A. Graesser, J. Sullins, and B. Gholson, "Affect and learning: an exploratory look into the role of affect in learning with autotutor," *Journal of educational media*, vol. 29, no. 3, pp. 241–250, 2004.

[5] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms," *Trends in cognitive sciences*, vol. 17, no. 11, pp. 585–593, 2013.

[6] N. Thompson and T. J. McGill, "Affective tutoring systems: enhancing e-learning with the emotional awareness of a human tutor," *International Journal of Information and Communication Technology Education*, vol. 8, no. 4, pp. 75–89, 2012.

[7] N. Schwarz, "Emotion, cognition, and decision making," *Cognition & Emotion*, vol. 14, no. 4, pp. 433–440, 2000.

[8] B. Lehman, S. DMello, and N. Person, "The intricate dance between cognition and emotion during expert tutoring," in *Intelligent Tutoring Systems*. Springer, 2010, pp. 1–10.

[9] S. Petrovica and M. Pudane, "Simulation of affective student-tutor interaction for affective tutoring systems: Design of knowledge structure," in *Proceedings of EET 2016*, vol. 7, 2016.

[10] S. DMello, N. Blanchard, R. Baker, J. Ocumpaugh, and K. Brawner, "I feel your pain: A selective review of affect-sensitive instructional strategies," *Design Recommendations for Intelligent Tutoring Systems*, vol. 2, pp. 35–48, 2014.

[11] R. A. Sottilare, J. A. DeFalco, and J. Connor, "A guide to instructional techniques, strategies and tactics to manage learner affect, engagement, and grit," *Design Recommendations for Intelligent Tutoring Systems*, vol. 2, pp. 7–33, 2014.

[12] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments," in *Proceedings of the CogSci 2007*, vol. 29, no. 29, 2007.

[13] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs." in *Proceedings of Interspeech 2006*, 2006.

[14] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child speech recognition in human-robot interaction: evaluations and recommendations," in *Proceedings of ACM/IEEE HRI 2017*. ACM, 2017, pp. 82–90.

[15] J. H. Kahn, R. M. Tobin, A. E. Massey, and J. A. Anderson, "Measuring emotional expression with the linguistic inquiry and word count," *The American journal of psychology*, pp. 263–286, 2007.

[16] S. D'Mello and A. Graesser, "Automatic detection of learner's affect from gross body language," *Applied Artificial Intelligence*, vol. 23, no. 2, pp. 123–150, 2009.

[17] D. McColl and G. Nejat, "Affect detection from body language during social hri," in *21st IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2012, pp. 1013–1018.

[18] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *Proceedings of ICME 2005*. IEEE, 2005, pp. 940–943.

[19] O. Villon and C. Lisetti, "A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors," in *15th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2006, pp. 269–276.

[20] M. H. Immordino-Yang and A. Damasio, "We feel, therefore we learn: The relevance of affective and social neuroscience to education," *Mind, brain, and education*, vol. 1, no. 1, pp. 3–10, 2007.

[21] A. Esposito, "Affect in multimodal information," in *Affective Information Processing*. Springer, 2009, pp. 203–226.

[22] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert)." *Emotion*, vol. 9, no. 5, p. 691, 2009.

[23] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of MM 2005*. ACM, 2005, pp. 677–682.

[24] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, "Emotion sensors go to school." in *Proceedings of AIED 2009*, vol. 200, 2009, pp. 17–24.

[25] L. Shen, M. Wang, and R. Shen, "Affective e-Learning: Using emotional data to improve learning in pervasive learning environment related work and the pervasive e-learning platform," *Educational Technology & Society*, vol. 12, pp. 176–189, 2009.

[26] S. Alexander, A. Sarrafzadeh, S. Hill *et al.*, "Easy with eve: A functional affective tutoring system," in *Workshop on Motivational and Affective Issues in ITS*. Citeseer, 2006, pp. 5–12.

[27] S. D'mello and A. Graesser, "Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back," *Interactive Intelligent Systems*, vol. 2, no. 4, p. 23, 2012.

[28] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Proceedings of 30th AAAI Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 3951–3957.

[29] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of CHI 2016*. ACM, 2016, pp. 3723–3726.

[30] S. Alexander, A. Sarrafzadeh, and S. Hill, "Foundation of an affective tutoring system: Learning how human tutors adapt to student emotion," *International journal of intelligent systems technologies and applications*, vol. 4, no. 3-4, pp. 355–367, 2008.

[31] J. de Wit, T. Schodde, B. Willemsen, K. Bergmann, M. de Haas, S. Kopp, E. Krahmer, and P. Vogt, "Exploring the Effect of Gestures and Adaptive Tutoring on Childrens Comprehension of L2 Vocabularies," in *Proceedings of the Workshop R4L at ACM/IEEE HRI 2017*, 2017.