Investigating Transparency Methods in a Robot Word–Learning System and Their Effects on Human Teaching Behaviors

Matthias Hirschmanner[1], Stephanie Gross[2], Setareh Zafari[3], Brigitte Krenn[2], Friedrich Neubarth[2] and Markus Vincze[1]

***Abstract*— Robots need to understand words for references in social spaces (e.g., objects, locations, actions). Grounded language learning systems aim to learn these words from observing a human tutor. Teaching a robot is difficult for naive users due to the discrepancy between the users' mental model and the actual state of the robot. We introduce a grounded word-learning system with the Pepper robot which learns object and action labels and investigate two extensions geared towards increasing the system's transparency. The first extension utilizes deictic gestures (pointing and gaze) to communicate knowledge about object names, and to further request new labels. The second extension shows the current state of the lexicon on the robot's tablet. We performed a user study (n=32) to investigate the effects of the transparency methods on learning performance and teaching behavior. In a quantitative analysis, we did not see a significant performance increase for the two extensions. However, users reported higher perception of control and perceived learning success, the better they knew the current state of the learning system. In a qualitative analysis, we investigated the participants' teaching behaviors and identified factors that inhibited the learning process. Among other things, we found increased interactive behavior of users when the robot displayed deictic gestures. We saw that human tutors simplified their utterances over time to adapt to the perceived capabilities of the robot. The tablet was most helpful for users to understand what the robot had already learned. However, learning was impaired in all conditions, when the human input substantially deviated from the form required by the learning system.**

I. INTRODUCTION

Robots are moving into environments where they need to adapt to new situations and learn from people who are not robotics experts. Thus, the robot needs to be able to deal with language input, and to link agents, objects, actions, locations etc. with specific words used by its human collaborator in certain situations. This is necessary for understanding verbal instructions, such as *Put the ketchup into the fridge*. In other words, language learning needs to be grounded, including sensory-motor experience [1] in shared social spaces [2]. This also relates to work in developmental psychology emphasizing the importance of multi-sensory experience for early word learning in infants [3] and related approaches in robot word learning from multi-sensory channels [4].

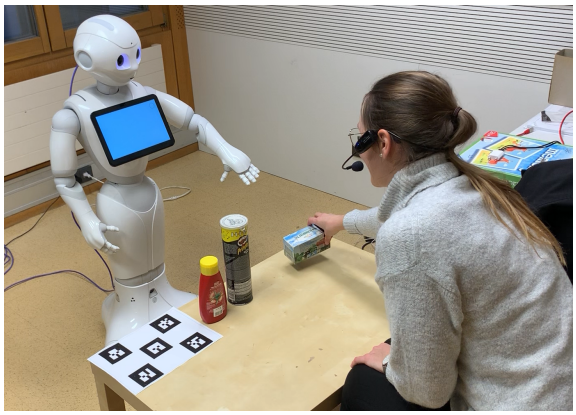

Fig. 1. User experiment setup: A user performs an object manipulation action and describes what they are doing. The Pepper robot observes the scene and learns object and action labels from the observations.

In the present paper, we investigate word learning in robots as a multi-modal and cross-situational learning task. Particular emphasis is put on how to make the robot's learning process more transparent to the human tutor. To do so, we make use of two strategies: (i) We implement an extension using deictic gestures (pointing and gaze) to an existing word learning model where the robot points at objects to request further information regarding action and object labels from its human tutor. (ii) We use a visualization on the robot's built-in tablet to provide the human with information on the current state of the robot's lexicon.

In an experimental setting, we let humans teach the robot three objects and actions. We use the humanoid robot Pepper in a table setting as shown in Fig. 1. Applying visual perception, the current action of a human tutor is detected by tracking the objects they are manipulating. In addition, the human tutor is describing their current actions. Our incremental word learning system uses statistical co-occurrences (npmi - normalized pointwise mutual information) of references (objects, actions) and words to learn how each individual tutor is referring to the objects and actions. Our approach does not require large corpora of language data, on the contrary a handful of utterances is sufficient to learn object and action names. It does not require a specific grammar and can therefore be seen to some extent as language-agnostic. The presented experiment was conducted

This research is partially supported by the Vienna Science and Technology Fund (WWTF) project RALLI (ICT15-045) and project "Human tutoring of robots in industry" (NXT19-005), the Austrian Science Foundation (FWF) project InDex (I3969-N30), and the Austrian Research Promotion Agency (FFG) Ideen Lab 4.0 project CoBot Studio (872590).

[1]Matthias Hirschmanner and Markus Vincze are with the Automation and Control Institute, TU Wien, 1040 Vienna, Austria {hirschmanner, vincze}@acin.tuwien.ac.at

[2]Stephanie Gross, Brigitte Krenn and Friedrich Neubarth are with the Austrian Research Institute for Artificial Intelligence (OFAI), 1010 Vienna, Austria {stephanie.gross, brigitte.krenn, friedrich.neubarth}@ofai.at

[3]Setareh Zafari is with the Institute of Management Science, TU Wien, 1040 Vienna, Austria setareh.zafari@tuwien.ac.at

in German.

In robot learning, the difference between the user's mental model of the robot and its actual capabilities might impair learning performance. Chao et al. [5] define transparency in the context of robot learning as "revealing to the teacher what is known and what is unclear" which should improve the learning experience and reduce the workload for the teacher. Wallkötter et al. [6] define five groups of social cues for transparency: speech, movement, text, imagery, and other. In the presented experiment, we used robot pointing and head gaze (movements) to communicate known object names and/or request the object name by pointing at objects of interest, and we used Pepper's tablet to display information about the current state of the robot's lexicon (text and images). In a quantitative evaluation, we investigated whether these extensions improve the overall performance of the grounded word learning system and increase the self-efficacy of the users. In addition, we qualitatively analyzed the human teaching behaviors, compared teaching behaviors and robot conditions, and examined how robot behavior influenced the human teaching behavior and how this fitted with the requirements of the learning system.

In Section II we discuss related works from the domains of grounded language learning and transparency in HRI. The model and the two proposed extensions are introduced in Section III. We describe the setup of the user experiment in Section IV, and evaluate results in Section V. Section VI discusses the results and lists resulting challenges for language learning systems and Section VII concludes the paper.

II. RELATED WORK

In our npmi-based approach to learning word-object/action pairings from uncluttered visual scenes and temporally co-inciding utterances, we take up findings from developmental psychology. In particular, we take advantage of findings that egocentric views in both infants and adults are highly selective [7] combined with evidence that parents often name objects temporally coinciding with these moments when objects are prominent in the infants' view, and toddlers are highly likely to learn object names from those pairings of visual input and linguistic reference [8] using co-occurrence statistics [9]. Multi-modal, cross-situational word learning is a widely pursued approach in word learning for robots. See for instance Taniguchi et al. [4] for a comprehensive overview of different approaches, and Krenn et al. [10] for background and description of the npmi-model realized in the Base System described in Section III of the current paper.

The topic of transparency and explainability in HRI has received increased attention over recent years with the goal of increasing trust, robustness and/or efficiency [6]. Many different social cues are used to increase transparency. Baraka and Veloso [11] use programmable lights on a mobile robot to help humans understand its current state. Chao, Cakmak and Thomaz [5] utilize active learning with non-verbal gestures to increase transparency which increased accuracy and efficiency. De Greef and Belpaeme [12] used social cues (gaze and utterances) to signal learning preference in a language game setting. They found increased performance and better mental models of the human tutors. Deictic gestures (e.g., pointing, gaze) have shown great potential to direct human attention [13] [14]. In our work, we use pointing and gaze to communicate knowledge of object labels and actively request information from the human tutor.

Visualization on screens is a powerful tool to convey information to the user. Ramaraj et al. [15] use visual representation of the scene and allow the user to ask the robot about its perception as transparency mechanisms to improve the user's mental model so they can identify causes of interaction failures. Wortham, Theodorou and Bryson [16] use visualizations to display a low-cost mobile robot's plans which improved the users' mental models. Perlmutter et al. [17] compare a screen based and a virtual reality based visualization in a situated language understanding context. They observed increased efficiency and accuracy of the given commands by participants using the transparency measures. In our work, we utilize the built-in tablet to show the current state of a word-learning system as a transparency measure.

Teaching behavior of human tutors in HRI settings is influenced by the robot and has been investigated by multiple researchers [18]. Fisher, Lohan and Foth [19] found more interactivity of human tutors when interacting with an active embodied agent (iCub) that moved its head compared to the same agent with only eye movements and a simulated robot. Aliasghari et al. [20] found that splitting the robot's gaze between the task and the teacher can increase perceived eagerness to learn in a video study with a simulated iCub robot. Vollmer et al. [21] saw that adult-robot teaching is similar to adult-infant teaching with slower and more exaggerated movements compared to adult-adult teaching. Lohse, Wrede and Schillingmann [22] conducted a study in which pairs of users taught object labels to a robot. They found that participants used longer utterances and more motion peaks if the learning performance was bad. Kim et al. [23] found similar results of human teachers using more verbal guidance and feedback if the robot learner has been struggling before in an interaction experiment with the Pleo dinosaur robot. Pelikan and Broth [24] found that participants simplified their utterances to adapt to the perceived limited capabilities of a Nao robot in an interaction scenario. In our study, we investigate how teaching behavior changes over time depending on the interactivity of the robot and the user's perceived performance of the word-learning system.

III. CROSS-MODAL WORD-OBJECT AND WORD-ACTION LEARNING ON PEPPER

For word-object and word-action learning, we use cross-modal input and an incremental information theoretic model. We use the humanoid Pepper robot of SoftBank Robotics as an interactive embodied agent. It has all the necessary components built-in such as cameras for object tracking, a tablet for displaying information and arms for non-verbal communication (i.e. pointing at objects). We use an additional computer that is connected to Pepper via Ethernet for

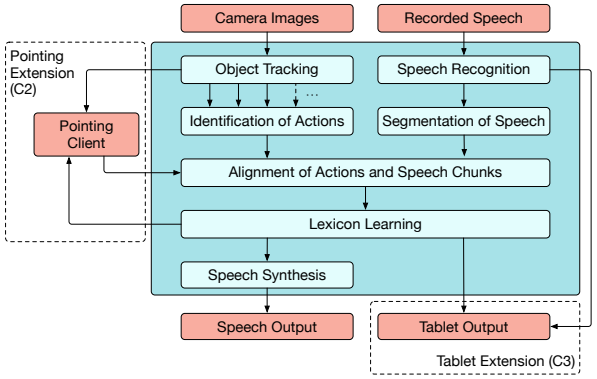

Fig. 2. Overview of the system architecture including the two proposed extensions.

processing the multi-modal inputs. The user wears a Bluetooth microphone for better input to the speech recognition module. An overview of the whole system including the two proposed extensions is shown in Fig. 2. Base model and extensions are detailed below.

A. Base System (Condition 1)

The goal of the word learning system is to acquire a lexicon of word-object and word-action mappings. A human tutor manipulates objects on a table while describing what they are doing. Utterance-situation pairs are detected and used as input to the word learning system. An episode with two utterance-situation pairs might be *<I take the box* - ACTION1 OBJECT1>, *<and put it next to the can* - ACTION2 OBJECT1 OBJECT2>. The situation is inferred from the vision system which tracks the objects. We use an 6D object pose tracker processing monoscopic RGB images using FAST features and recorded object models from [25]. This ensures high frame rates to follow the tutor's motions. However, problems of occlusions or objects leaving the field of view cannot be avoided with the single camera setup and can influence learning performance. The action the user is performing can be inferred from the change of the object position. If an object is moving, we can safely assume it is moved by the user deliberately.

In the current setup, we focus on the three basic object manipulation actions TAKE, PUT and PUSH. The TAKE observation is triggered when an object is lifted from the table plane and the PUT action when it is returned to the table. The PUSH action is registered if an object moves on the table plane. Actions are only registered if the movement is above a certain small threshold to mitigate registering wrong actions due to noise of the object tracker.

The speech of the human tutor is converted to text using Google Cloud Speech-to-Text engine. If it coincides with a registered action sequence it is processed as an utterance-situation pair. We employ the algorithm described in [26] to align these pairs. We use normalized pointwise-mutual information ($npmi$) to learn word-object/action mappings. It is a measures for the likelihood of an object/action-word co-occurrence. The $npmi$ value is updated after each detected situation-utterance pair. Therefore, it is an incremental learning system that does not depend on large corpora. Additionally, it can also be seen as a language-agnostic approach as it does not rely on a specific linguistic structure. If an object manipulation is detected for which the involved action and object labels have already been learned by the system, the robot utters the observation using the learned words (e.g., "take bottle", "push box"). For more details on the base system we refer to [10].

B. Pointing Extension (Condition 2)

We extend the base algorithm with an active component of the robot requesting information by means of non-verbal communication (i.e. pointing at objects). The passive learning algorithm is interrupted by a pointing sequence explained below. This approach is inspired by recent findings of child language acquisition which we summarize in [27]. The pointing sequences have two purposes, to request new information and to communicate if an object label has already been learned. Pointing is initiated by Pepper making a "Hmm?" sound to attract the attention of the human tutor. Subsequently, Pepper directs its gaze and arm at the object of interest. After the movements have finished, Pepper looks at the human tutor. If Pepper has not yet learned the word, it makes another "Hmm?" sound and waits for an utterance of the human. The utterance is connected to the specific object reference and its $npmi$ value is updated. If the object name is already in the lexicon, Pepper says the learned word associated with the object. The utterance of the tutor is still used to reinforce or attenuate the current believe. The pointing sequence is completed by the robot making an "Ah!" sound, retracting its arm and looking back at the table.

The pointing capabilities of Pepper are limited because its fingers cannot be actuated individually which prevents index finger pointing. Additionally, the tablet on Pepper's chest limits upper-arm movements. To circumvent these shortcomings, we use the direction of the forearm with all fingers extended to point at objects. The inverse kinematic problem is solved by an iterative method which aligns Pepper's forearm direction with the direction of the object. We use the Moore-Penrose pseudoinverse with a second term which ensures the joints staying within a reasonable range similar to [28] and apply gradient descent to calculate relative joint angles. To avoid the human tutor confusing the object being pointed at, a pointing action is only initiated if an object is spatially separated from all other objects on the table and all objects are currently detected by the object tracker.

C. Tablet Extension (Condition 3)

In the second extension to the base model, Pepper's tablet is utilized to visualize the current state of the lexicon. An example of the content shown on the tablet can be seen in Fig 3. The most recent speech recognition result is shown on top of the tablet screen. This is supposed to help the human tutor to adjust their voice (e.g., speed, dialect) for

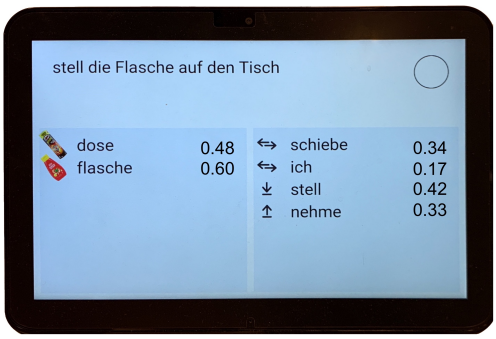

Fig. 3. Information shown on Pepper's tablet for C3 (tablet extension). The current speech recognition result is shown on top with the current state of the lexicon below.

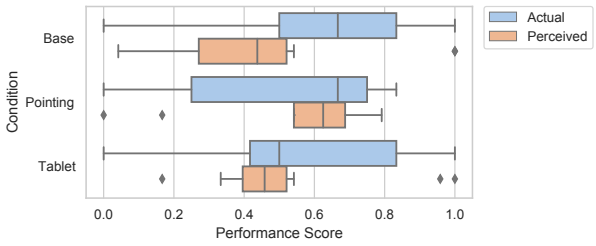

Fig. 4. Comparison between the actual and perceived performance scores averaged over all 6 references (i.e., objects, actions)

better speech recognition results. Additionally, it might help to understand why the robot learned incorrect words.

On the lower part of the display, we show the current entries in the lexicon. For each lexicon entry, a symbol is shown with the associated word and the current $npmi$ value next to it. We use images of the objects as their symbols. The actions are symbolized by arrows indicating the direction the object is moving (i.e. lift from the table, put on the table, move horizontally on the table). The $npmi$ value is shown to indicate which word has higher confidence if there are multiple word candidates for the same reference.

The tablet view is implemented as a local web page which is dynamically updated using JavaScript. It communicates with the rest of the system using ROS.

IV. EXPERIMENTAL SETUP

We conducted the experiment in the library of TU Wien. The participants sat in front of a low table with the Pepper robot across the table observing the scene as shown in Fig. 1. Three objects were positioned on the table. The participants were handed a consent form and the written explanation of their task. Additionally, a researcher explained the different parts of the system such as the speech recognition and the object tracker. They also gave an example of an action-utterance pair (e.g., *I take the can and put it over there.*). In the pointing extension (C2), the pointing of the robot was demonstrated and it was verified that the participant could identify the object pointed at. In the tablet extension (C3), a printout of a possible state of the tablet (Fig. 3) was shown to the participant and it was explained to them.

a) Participants: A total of 36 participants were recruited at TU Wien library. We excluded 2 participants due to technical failures during the experiment and 2 participants due to insufficient German language skills. Therefore, 32 participants between the ages of 18 and 62 ($M = 28.16$, $SD = 9.27$) remained in the evaluation. A total of 18 participants identified themselves as women and 14 as men.

b) Task: The task for the users consisted of teaching the robot 6 words (3 object labels, 3 actions labels) through object manipulation. Participants could choose the words they would usually use to refer to the objects, and to the actions (take, put, and push an object). The experiment was conducted in German.

c) Conditions: We conducted a between-subjects study with 10 participants in the base condition (C1), 11 in the pointing condition (C2) and the tablet condition (C3), respectively.

- C1 Robot utters Object+Action if it observes the user performing an action for which both labels (action and object) have been learned (Base System)
- C2 Base + Pointing at objects with "Hmm" if the object label is unknown, or uttering the object label
- C3 Base + Information on the state of lexicon shown on tablet

d) Procedure: After the explanation of the setup and the task, the learning process was started by announcement of the researcher and Pepper directing its gaze towards the table. Two researchers stayed in the room during the trial. For each participant, videos were recorded from two perspectives for the qualitative analysis. The learning progress is stored as log files. The researchers stopped the word-learning experiment after five minutes, if the participants did not stop on their own beforehand. The participants could stop at any point on their own. After the interaction, the participants filled in a questionnaire.

V. EVALUATION AND RESULTS

In this Section, we present the results of the user experiment. In a quantitative analysis we investigate the effects of the different transparency measures on learning performance and user experience. In a qualitative analysis we perform a detailed evaluation of the teaching behaviors of the individual participants and how they influence learning performance.

A. Quantitative Analysis

Hypotheses:

- H1 Pointing extension and tablet extension improve the learning performance of the robot
- H2 Pointing extension and tablet extension improve the perceived overall learning success, perception of control and self-efficacy in participants
- H3 Pointing extension and tablet extension increase knowledge of the system's state

a) Performance: As a performance measure, we consider the amount of correct word-reference pairs in the lexicon at the end of the experiment. A word-reference pair is considered to be correctly learned, if the intended word has the highest $npmi$ value of all words for a reference and $npmi > 0.1$. A Kruskal-Wallis test indicated no significant difference of this score between the conditions, $\chi^2(2) = 0.39$, $p = 0.82$. (C1: $M = 0.6$, $Mdn = 0.67$), the pointing extension (C2: $M = 0.52$, $Mdn = 0.67$) and the tablet extension (C3: $M = 0.56$, $Mdn = 0.50$). Thus, H1 is rejected.

b) Perceived success of learning process, perception of control, and self-efficacy: To evaluate the participants' confidence/accuracy in their performance, we paired the above mentioned score with subjective measures for each learned reference. We asked participants to rate on a 5-point Likert scale, how well they thought the robot has learned each word (1 = Not at all, 5 = Perfectly). The *perceived performance score* was then calculated by taking the average of the 6 scored items (Cronbach's alpha = 0.82) and rescaled to be in the range $[0, 1]$. The comparison between actual and perceived performance is shown in Fig. 4. A Wilcoxon Signed-Ranks test revealed no significant difference between the actual learning and perceived learning scores between the conditions, Z= -0.85, $p = 0.39$.

Additional to the per reference performance ranking from above, we used 4 subjective items to determine the perceived overall success of the learning process (Cronbach's alpha = 0.79). A sample item is "The robot was able to learn the objects I taught it." The participant's perceived self-efficacy in interacting with a robot was measured by 6 items from SE-HRI scale [29] (Cronbach's alpha = 0.83). A sample item is "I could get a robot to perform a specific task". Perception of control was measured by 2 items from [30]. A sample item is "I felt that I had control over what the robot was learning". All items were rated on a 5-point Likert-scales (1 = strongly disagree, 5 = strongly agree).

Results from a Kruskal-Wallis test indicated no significant difference in the participants' perceived success of learning process, ($\chi^2(2) = 2.20$, $p = 0.33$) and perceived self-efficacy among the conditions, ($\chi^2(2) = 1.26$, $p = 0.53$). We also found no significant difference on participants' perceived control among the conditions ($\chi^2(2) = 2.31$, $p = 0.32$), rejecting H2. A box-plot of the self-efficacy and perception of control can be seen in Fig. 5.

c) Knowledge of the system's state: We used 1 subjective question to determine the knowledge of the current state of the lexicon on the same 5-point Likert scale (i.e., "While teaching the robot, it was clear to me which words the robot knew already and which ones it still had to learn."). A Kruskal-Wallis test indicated a significant difference of knowledge of the system's state between the conditions ($\chi^2(2) = 7.37$, $p < 0.05$), with a mean rank score of 12.45 for the base extension (C1), 14.45 the pointing extension (C2) and 22.23 for the tablet extension (C3). This suggests that transparency of the robot's learning process is highest when there is a visualization of robot's internal state. The difference is significant between the base and the tablet condition, and thus H3 holds for C1 compared to C3.

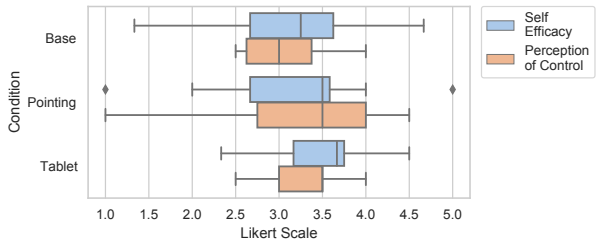

Fig. 5. Self-efficacy and perception of control per condition

Furthermore, we tested the correlation among our variables. We found positive correlation between knowledge of the system's state and perceived success of the overall learning process (Spearman's $\rho = 0.35$, $p < 0.05$) and perceived control (Spearman's $\rho = 0.54$, $p < 0.01$). That implies the more transparent the robot's learning process is, the higher is the perception of robot's learning success and the perception of control in the participants. Furthermore, the perception of control is positively correlated with the perceived success of learning process (Spearman's $\rho = 0.57$, $p < 0.01$), and with self efficacy (Spearman's $\rho = 0.45$, $p < 0.01$). That is the more participants feel in control, the higher their perception of self-efficacy is and the higher the perceived robot's learning success is.

B. Qualitative Analysis

Research Questions:

a) Do participants show different behaviors when they start to tutor the robot?
b) Do participants change their teaching behavior over time?
c) Is there a correlation between teaching behavior and condition, i.e., how the robot interacts with the human?

a) Teaching Behavior: In general, participants (n=32) showed different behaviors when teaching the robot. However, the majority of participants (n=27) started as instructed with utterances of the following pattern: "[agent="I"], [action x] [object x]" (e.g., "I take the box"). Some of these utterances also contained a location (e.g., "I put the box on the table"), or a spatial relation and another object ("I move the box next to the bottle"). For these participants 70.4% of all object labels and 50.6% of all action labels were learned correctly by the system. Only five participants showed divergent behaviors: (i) introducing objects first ("This is a ...") (n=1), (ii) introducing actions first ("I take, I move" etc.) (n=1), (iii) using one noun ("box") for referring to two objects (can and box) (n=1), (iv) introducing one of the three actions only in the second half of the interaction (n=2). For the five participants with a divergent teaching behavior, only 40% of object labels were learned correctly and 13.3% of action labels. This might be due to the lack of compatibility of the given input with the learning algorithm.

b) Change of Teaching Behavior: When participants received feedback from Pepper either in form of utterances (C1, C2, C3) or pointing gestures (C2), 14 did not change their teaching behavior when describing the conducted actions. This group includes all three conditions and interactions resulting in high and low learning scores.

Participants who changed their teaching behavior in the course of the interaction did it in the following way:

- omitting locations and thus shortening the utterances, e.g., "I put the bottle" (n=6 participants)
- increasing interactive behavior, such as giving verbal feedback on what Pepper uttered (e.g., "very good", "no"), uttering the object name and pointing at or lifting the object (n=5)
- omitting the subject ("I"), e.g., "push the bottle" (n=4)
- uttering object name + infinitive (n=3), e.g., "box take"
- simplifying action descriptions by omitting certain action labels (n=1)
- changing to a very repetitive behavior over time, repeating object and action labels on their own, e.g., "crisps [pause] crisps [pause] crisps" (n=1)
- using passive voice and descriptions of situations, such as "the can is moved forward", or "the bottle is next to the box" (n=1)

As reflected in these items, most of the participants for whom the learning process did not go well tried to simplify their utterances over time. However, the analysis shows that the end results in experiments where participants changed their teaching behavior (learned object labels: 66.6%, learned action labels: 46.3%) are very similar to those in experiments where participants did not change their teaching behavior (learned object labels: 64.3%, learned action labels: 45.2%). This reflects that for some participants the results improved after adapting the teaching process. However, others simplified their utterances in a way that did not contain all information necessary for our learning system and thus negatively affected word learning. Also, the simplifications bore the risk that the granularity of the described actions did not match the granularity in which the system was able to perceive the actions any more (e.g., "I put the box on the table" versus "I take the box and put it on the table" describing the same action). This resulted that on average the learning success was very similar to the group that did not adapt the teaching behavior.

Some participants, for whom learning went well, checked whether Pepper learned everything correctly by the end of the experiment. They manipulated objects without verbally describing their actions but listened if it was correct what Pepper uttered. One participant used more complex utterances by adding spatial relations and other objects, after Pepper had learned all the labels correctly.

c) Differences Between the 3 Conditions: With regards to the change of behavior between the different conditions, very few participants changed their behavior in C1 (base system) and the most in C2 (pointing extension). This might be related to the increase of interactive behavior of Pepper in C2. The replies by the 11 participants in C2 given to the 43 pointing actions of Pepper diverged from the other utterances. Participants (i) uttered a correct object label, such as "tea box", sometimes preceded by "this is a..." (n=11), (ii) uttered verbal feedback - only on correctly learned object labels - such as "yes", "very good" or "correct" (n=3), or (iii) manipulated that object accompanied by a task description (n=1), or (iv) did not reply at all (n=3). When Pepper started pointing in C2, participants reacted to Pepper's utterances also in diverging ways, even if it was not pointing: participants (i) uttered "yes", "no" or "bravo" etc., partially including nods (n=5), (ii) corrected Pepper's utterances (n=3), and (iii) pointed at objects or lifted them and uttered their name or "This is a ..." (n=2). On the other hand, participants in C3 (tablet extension) did not provide any feedback of that kind and only 2 participants in C1 nodded, when Pepper uttered words correctly. Thus, the type and amount of feedback participants gave to robot's utterances and actions was a conspicuous difference between the three conditions.

Participants in all three conditions mixed up words or used different words to refer to the same action or object, between and within participants. For example, 3 to 12 different nouns were used per participant to refer to the three objects (on average 4.2 nouns per participant; Mdn:3). There was a higher variation in object names in C2 (on average 5.1 nouns per participant; Mdn:3). However, it did not affect the learning, as especially for the participants who varied a lot learning worked well. This shows that our learning system was able to deal with this variation, which is a positive affect of cross-situational and cross-modal language learning systems in general. For learned action labels, the learning scores of the three conditions are comparable: C1: 46.7%, C2: 45.5%, C3: 45.5%. However, there is a difference in the percentage of correctly learned object labels: C1: 73.3%, C2: 57.6%, C3: 66.7%. The reason why the learning scores for C2 are worse might be due to the increase of interactive behavior. Participants provided feedback and utterances other than describing the conducted action, which is not compatible with the learning algorithm.

Although on average word learning did not work better in C3 than in the other two conditions, of the 4 participants who were able to teach all words, 3 were in C3 and 1 in C1. Therefore, we assume that if word learning worked well and the shown teaching behavior provided suitable input for the learning system, the tablet supported participants to focus on words which were not yet learned correctly. However, if the teaching behavior did not suit the learning algorithm, no information was provided on how the teaching behavior should be adapted and therefore the experiment was not successful despite the additional information given to the participants.

Robotic factors, such as object tracking performance (wrong or no actions detected) or performance of speech recognition (wrong or no words detected) also impaired learning performance on an individual level. However, there was no correlation between tracking/speech recognition performance and the performance of the word-learning system.

We hypothesize that the reason for this is the relatively low amount of data required to teach our system. In the four experiments where all labels were learned correctly, only 17-36 (Mdn: 20.5) utterances were needed. As long as sufficient examples are recognized from which the system is able to learn, it does not negatively influence learning success if actions are not detected (e.g., due to occlusions). Additionally, these robotic factors will always be prevalent. This emphasizes the importance of designing language learning systems being able to deal with wrong or missing input.

VI. DISCUSSION

In the quantitative analysis, we investigated the influence of transparency on user experience and performance. We found that transparency (knowledge about the system's state) correlates positively with the users' perception of control and perceived learning success. Additionally, perception of control and perceived learning success were positively correlated with self-efficacy. All of these factors are important to keep a user motivated and interested to interact with a system over a longer time horizon [29]. In our setup, displaying information on the tablet was most beneficial to increase transparency. Most effects were not significant due to the small sample size, but there was a tendency of increased self-efficacy for both extensions (Base $Mdn = 3.25$, Pointing $Mdn = 3.5$, Tablet $Mdn = 3.67$) and increased perception of control (Base $Mdn = 3$, Pointing $Mdn = 3.5$, Tablet $Mdn = 3.5$). However, we did not see an actual performance increase for the extensions.

In the qualitative analysis we looked at the different factors that influence performance and need to be addressed by future word-learning systems. We investigated which kind of teaching behavior participants show, in order to develop mechanisms which enable the learning system to consider relevant information. In addition, we investigated whether different behaviors of the robot, such as providing information on its current learning status via (i) utterances, (ii) combined with deictic gestures, or (iii) combined with a tablet, influence the teaching behavior of the human.

Based on the observed behaviors, we identified the following *challenges for language learning systems*:

User try to facilitate learning of the robot. The approach used by the majority of participants was to simplify their utterances, e.g., by omitting locations or subjects ("I"), by uttering object name + infinitive. While Lohse, Wrede and Schillingmann [22] and Kim et al. [23] found that participants used longer utterances if learning did not work well for the robot, our observations are in line with Pelikan and Broth [24] where participants simplified their utterances to adapt to the perceived limited capabilities of the robot. Some users also changed the granularity of their verbal descriptions by uttering only action or object labels, when learning did not go well. However, the actions perceived by the system (take, put, push) need to be of the same granularity as the actions described by the utterances of the human tutor, i.e., there needs to be one word for one action. In our user study, the granularity of descriptions varied between and within participants. Therefore, the system needs to be able to combine the perceived actions flexibly (e.g., "put" with and without a preceding "take" action). Also, new actions were added spontaneously by participants. All this requires a robust language learning system.

Labels for a specific object or action varied between and within participants, participants sometimes mixed up words or used pronouns instead of object names. Therefore, mechanisms for coreference resolution are required, and the learning algorithm needs to be robust enough to deal with lexical variety and to "forget", if words are confused by the human or wrongly recognized by the system.

The type of interactivity influences the behaviors of users. Depending on the communicative cues given by the robot, the interactive behavior of users might increase, e.g., in addition to replies to deictic gestures from the robot. However, the system then needs to be able to deal with these different types of input, such as feedback on the system's learning behavior (e.g., "very good" or "correct"). To deal with this type of input, inspirations can be derived from studies on developmental language learning, where children receive verbal feedback from their care-takers and individual objects are visually prominent [3], [7].

Even if incorporating more interactive robot behavior bears the risk that the input given by humans can not be fully dealt with by the learning algorithm, it still makes sense to include this type of interaction. We saw a tendency that participants in the interactive condition C2 perceived higher learning success, self-efficacy and perception of control (see Fig. 4 and Fig. 5) but it would require a larger sample size to get a definitive result. The observed tendency is in line with similar research that showed people ascribing more competence to a more active robot [19].

Utilizing a visualization (e.g., Pepper's tablet) as a transparency mechanism to visualize the current state of the lexicon to the tutor has potential to support human tutors in their teaching process, only if their teaching behavior provides suitable input for the learning system. It does not automatically support them in adapting their teaching behavior. Future research is needed to investigate in how far the visualization can be used to provide feedback to humans on how to adapt their teaching behavior in order to increase learning success.

VII. CONCLUSION

In this paper, we investigated two extensions for a grounded object and action word-learning system with the aim of increasing transparency. For the first extension we implemented deictic gestures (pointing and gaze) to communicate known/unknown object names and actively request the object name from the user. The second extension we added a visualization of the current state of the learned lexicon (object and action names) on the robot's display. In a user-study (n=32) with the Pepper robot we found benefits of both extensions and identified challenges for future word-learning systems. The tablet was perceived as most helpful to communicate the current state of the word-learning system.

The deictic gestures increased interactivity of the users. We saw that human tutors in all conditions simplified their utterances over time to adapt to the perceived capabilities of the robot.

ACKNOWLEDGMENT

The authors would like to thank: Clara Haider for providing the motion algorithms and tablet interface for Pepper, and for assistance during the user experiments; Helena Frijns for help displaying information on Pepper's tablet; Christiana Tsiourti and Astrid Weiss for help and guidance for the user experiment; Matthias Samonig and the "TU Wien Bibliothek" for providing the rooms for the experiments.

REFERENCES

[1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[2] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, N. Pinto, and J. Turian, "Experience grounds language," in *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 8718–8735.

[3] S. H. Suanda, L. B. Smith, and C. Yu, "The multisensory nature of verbal discourse in parent–toddler interactions," *Developmental Neuropsychology*, vol. 41, no. 5-8, pp. 324–341, 2016.

[4] A. Taniguchi, T. Taniguchi, and A. Cangelosi, "Cross-situational learning with bayesian generative models for multimodal category and word learning in robots," *Frontiers in neurorobotics*, vol. 11, p. 66, 2017.

[5] C. Chao, M. Cakmak, and A. L. Thomaz, "Transparent active learning for robots," in *5th ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2010, pp. 317–324.

[6] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani, "Explainable Agents Through Social Cues: A Review," *arXiv:2003.05251 [cs]*, 2021, arXiv: 2003.05251.

[7] L. B. Smith, C. Yu, H. Yoshida, and C. M. Fausey, "Contributions of head-mounted cameras to studying the visual environments of infants and young children," *Journal of Cognition and Development*, vol. 16, no. 3, pp. 407–419, 2015.

[8] C. Yu and L. B. Smith, "Embodied attention and word learning by toddlers," *Cognition*, vol. 125, no. 2, pp. 244–262, 2012.

[9] A. R. Romberg and J. R. Saffran, "Statistical learning and language acquisition," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 1, no. 6, pp. 906–914, 2010.

[10] B. Krenn, S. Sadeghi, F. Neubarth, S. Gross, M. Trapp, and M. Scheutz, "Models of cross-situational and crossmodal word learning in task-oriented scenarios," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 3, pp. 658–668, 2020.

[11] K. Baraka and M. M. Veloso, "Mobile Service Robot State Revealing Through Expressive Lights: Formalism, Design, and Evaluation," *Int. Journal of Social Robotics*, vol. 10, no. 1, pp. 65–92, 2018.

[12] J. de Greeff and T. Belpaeme, "Why Robots Should Be Social: Enhancing Machine Learning through Social Human-Robot Interaction," *PLOS ONE*, vol. 10, no. 9, p. e0138061, Sept. 2015.

[13] H. Admoni, T. Weng, B. Hayes, and B. Scassellati, "Robot nonverbal behavior improves task performance in difficult collaborations," in *11th ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2016, pp. 51–58.

[14] R. M. Holladay, A. D. Dragan, and S. S. Srinivasa, "Legible robot pointing," in *23rd IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, 2014, pp. 217–223.

[15] P. Ramaraj, S. Sahay, S. H. Kumar, W. S. Lasecki, and J. E. Laird, "Towards using transparency mechanisms to build better mental models," in *Advances in Cognitive Systems: 7th Goal Reasoning Workshop*, vol. 7, 2019, pp. 1–6.

[16] R. H. Wortham, A. Theodorou, and J. J. Bryson, "Improving robot transparency: Real-time visualisation of robot ai substantially improves understanding in naive observers," in *26th IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, 2017, pp. 1424–1431.

[17] L. Perlmutter, E. Kernfeld, and M. Cakmak, "Situated Language Understanding with Human-like and Visualization-Based Transparency," in *Robotics: Science and Systems XII (RSS)*, 2016, pp. 40–50.

[18] A.-L. Vollmer and L. Schillingmann, "On Studying Human Teaching Behavior with Robots: a Review," *Review of Philosophy and Psychology*, vol. 9, no. 4, pp. 863–903, Dec. 2018.

[19] K. Fischer, K. Lohan, and K. Foth, "Levels of embodiment: Linguistic analyses of factors influencing HRI," in *7th ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2012, pp. 463–470.

[20] P. Aliasghari, M. Ghafurian, C. L. Nehaniv, and K. Dautenhahn, "Effects of Gaze and Arm Motion Kinesics on a Humanoid's Perceived Confidence, Eagerness to Learn, and Attention to the Task in a Teaching Scenario," in *ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI)*, 2021, pp. 197–206.

[21] A.-L. Vollmer, K. S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. J. Rohlfing, and B. Wredek, "People modify their tutoring behavior in robot-directed interaction for action learning," in *8th Int. Conf. on Development and Learning*. IEEE, 2009, pp. 1–6.

[22] M. Lohse, B. Wrede, and L. Schillingmann, "Enabling robots to make use of the structure of human actions - A user study employing Acoustic Packaging," in *IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*, 2013, pp. 490–495.

[23] E. S. Kim, D. Leyzberg, K. M. Tsui, and B. Scassellati, "How People Talk When Teaching a Robot," in *4th ACM/IEEE International Conference on Human Robot Interaction (HRI)*, 2009, pp. 23–30.

[24] H. R. Pelikan and M. Broth, "Why That Nao?: How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk," in *2016 CHI Conf. on Human Factors in Computing Systems*. ACM, 2016, pp. 4921–4932.

[25] J. Prankl, A. Aldoma, A. Svejda, and M. Vincze, "RGB-D object modelling for object recognition and tracking," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 96–103.

[26] S. Gross, M. Hirschmanner, B. Krenn, F. Neubarth, and M. Zillich, "Action Verb Corpus," in *2018 Language Recognition and Evaluation Conference*, 2018.

[27] B. Krenn, C. Tsiourti, F. Neubarth, S. Gross, and M. Hirschmanner, "Active Language Learning Inspired from Early Childhood Information Seeking Strategies," in *Workshop on Cognitive Architectures for Human-Robot Interaction at AAMAS2019*, 2019.

[28] A. Liégeois, "Automatic supervisory control of the configuration and behavior of multibody mechanisms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 7, no. 12, pp. 868–871, 1977.

[29] A. R.-V. D. Pütten and N. Bock, "Development and Validation of the Self-Efficacy in Human-Robot-Interaction Scale (SE-HRI)," *ACM Trans Hum.-Robot Interact*, vol. 7, no. 3, pp. 1–30, Dec 2018.

[30] P. J. Hinds, *User control and its many facets: A study of perceived control in human-computer interaction*. Hewlett Packard Laboratories, 1998.