Learning to Interact and Interacting to Learn: Active Statistical Learning in Human-Robot Interaction

Chen Yu, Tian Xu, Yiwen Zhong, Seth Foster and Hui Zhang

Abstract— **Learning and interaction are viewed as two related but distinct topics in developmental robotics. Many studies focus solely on either building a robot that can acquire new knowledge and learn to perform new tasks, or designing smooth human-robot interactions with pre-acquired knowledge and skills. The present paper focuses on linking language learning with human-robot interaction, showing how better human-robot interaction can lead to better language learning by robot. Toward this goal, we developed a real-time human-robot interaction paradigm in which a robot learner acquired lexical knowledge from a human teacher through free-flowing interaction. With the same statistical learning mechanism in the robot's system, we systematically manipulated the degree of activity in human-robot interaction in three experimental conditions: the robot learner was either highly active with lots of speaking and looking acts, or moderately active with a few acts, or passive without actions. Our results show that more talking and looking acts from the robot, including those immature behaviors such as saying non-sense words or looking at random targets, motivated human teachers to be more engaged in the interaction. In addition, more activities from the robot revealed its robot's internal learning states in real time, which allowed human teachers to provide more useful and "on-demand" teaching signals to facilitate learning. Thus, compared with passive and batch-mode training, an active robot learner can create more and better training data through smooth and effective social interactions that consequentially lead to more successful language learning.**

I. Introduction

Language is a central component of human intelligence which is fundamental and essential for human-human everyday communication. A basic function of language is to provide linguistic labels of objects and activities which people to refer to them in speech and share experiences in everyday communication[1]. Therefore, learning, understanding and using human languages by humanoid robot is critical for seamless human-robot interaction (HRI) [2-4].

Language learning and human-robot interaction are viewed as two related but distinct topics in developmental robotics. Researchers in artificial intelligence and machine learning are interested in how to build computational algorithms to acquire human languages. Meanwhile, researchers in HRI most often focus on how to design and implement smooth human-robot interfaces built upon speech recognition and natural language processing algorithms [5]. Thus, one obvious direction to connect the two topics is that better linguistic skills in robots lead to better interaction between humans and robots through more efficient verbal communication. The present paper investigates the link between learning and interaction in the other direction -- that is, better interaction can also lead to better language learning. More specifically, one effective way for robots to acquire human-like linguistic skills is to learn through social communication between a human language teacher and a robot language learner.

This idea is inspired by comparing language learning in machines and in humans. In machine learning, a typical paradigm is to first collect training data, and then to focus on developing advanced algorithms to extract and infer knowledge from data. The application of such algorithms is most often done in batch mode [6]. Thus, a machine learner/algorithm passively receives information from a training dataset in a one-way flow, without interaction nor feedbacks from human users. This scenario is quite different with how a young child learns the native language from his caregivers in everyday social contexts. In such contexts, caregivers as language teachers dynamically adjust their behaviors based on their understanding of the learner's mental state. Moreover, young language learners most often actively elicit information from teachers based on their own learning status. Good teachers then respond by providing "on-demand" information to young learners in real-time learning [7]. With responsive teachers, the learner plays an active role in leading learning-oriented interactions -- actively generating actions to interact with the physical environment and the teachers to seek data for successful learning. From this view, language learning should not be treated as the learner's task or the teacher's task. Instead, learning is a collaborative task in which both teachers and learners work together to achieve a shared goal [8].

Recently, there has been an increasing attention on linking learning with interaction[9-13]. For example, in [14], humans alter their behavior towards a robotic social partner by decreasing their hand movement velocity in action demonstration. In the study, robots seem to be treated by humans as infants with limited cognitive capabilities, as people modify their tutoring behavior in a similar way as what they do in adult-child interaction. In another study [15], a robot's real-time feedbacks shape the human tutor's demonstration, based on which tutors adjust their movement parameters, such as pauses, speech and height of motion. In [16], a reinforcement process is developed and used to utilize contingent interaction between a human teacher and a robot learner to extract word forms through a few minutes of dialogue. In [17], a socially-guided robot can not only learn by itself but also flexibly take advantage of the guidance of a human teacher who produces scaffolding acts to facilitate learning. One shared focus in those studies has been to

All of the authors are with Psychological and Brain Sciences, Cognitive Science, School of Informatics, at Indiana University, 1101 East 10th Street, Bloomington, IN, 47405, USA (corresponding author Chen Yu, phone: 812-856-0838; e-mail: chenyu@indianaa.edu).

incorporate user's feedbacks into machine learning algorithms. The results convincingly demonstrated that such feedbacks lead to effective learning.

In the present work, we argue that one compelling demonstration of the advantages of learning from social interaction is to show that given the *same* internal learning algorithm – *without* adding any new components to process additional feedbacks from the interaction – better social interaction lead to better learning. More specifically, a robot learner in our study generates different kinds of social behaviors when interacting with a human teacher, which change the teachers' responsive behaviors in the ways that teachers provide better teaching signals. Hence, even with the same learning machinery, better training data generated through social interaction lead to better language learning.

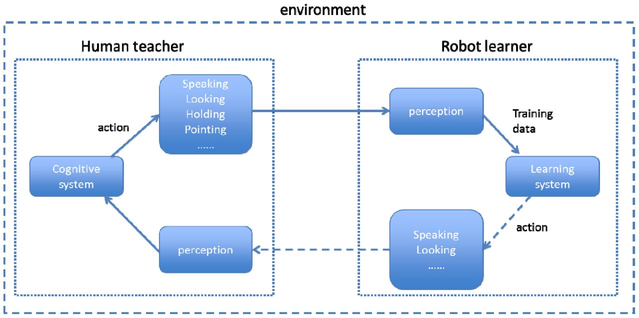

Figure 1. Overview of bi-directional human-robot interaction. We systematically manipulate the pathway indicated by dash lines to investigate how the dynamical coupling between human and robot can lead to successful statistical language learning.

Our study focuses on teaching robots the meanings of words, in particular, object names. This is a fundamental topic in human language acquisition and also such lexical knowledge is less likely to be pre-acquired by robots because of the evolution of vocabulary and flexibility of pragmatics [18]. Figure 1 shows two pathways of human-robot interaction in a word learning task: 1) From human to robot (solid lines in Figure 1) – a human teacher generates various actions and cues to elicit the robot's attention, and then names objects for the robot learner who receives visual and auditory information as training data for language learning; the task of the robot learner is to process information provided by the teacher to acquire language knowledge; 2) from robot to human (dash lines in Figure 1): the robot initiates actions based on its current learning state, producing spoken words and gazing at objects that it has already learned. Critically, the human teacher perceives the actions generated by the robot and adjusts his/her own teaching behavior accordingly. In this way, the behaviors from the teacher and the learner are closely coupled, forming a dynamic loop in the interaction. The present study systematically manipulates the pathway from robot to human (dash lines) and investigate how different responsive actions from the robot may change both the teacher's behavior and the dynamics of human-robot interaction, which lead to different learning outcomes.

II. REAL-TIME HUMAN-ROBOT INTERACTION AND LEARNING

The experiment was a language-learning task in which a human teacher was asked to teach a robot a set of object names in a shared environment. To do so, he needed to engage the robot and attract the robot's attention to the target object of his own interest, and then label the object for the robot learner. This joint task allowed participants and the robot to naturally interact with each other without any constraint on what they had to do or what they had to say. In order to teach the robot object names, human participants actively played the teacher's role and generated multimodal behaviors to attract the robot's attention, including eye contact, pointing to and manipulating objects in the shared environment, as well as speaking. Thus, the interaction itself was free-flowing, allowing participants to produce naturalistic behaviors. Each participant was asked to teach 16 objects with novel names to the robot. Learning those object-name mappings through a brief interaction was a challenging task.

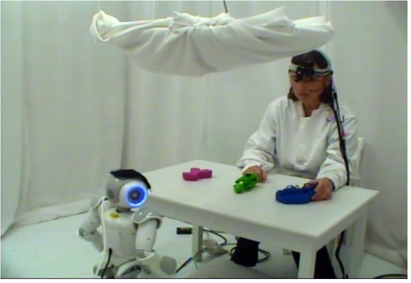

Figure 2. A human teacher was teaching a robot learner a set of object names in the interaction.

Figure 2 shows the experimental setup of our multimodal real-time human-robot interaction in which a human teacher attempted to teach a robot learner a set of object names. A Nao humanoid robot by Aldebaran Robotics was used in the experiment. The Nao robot has 35 DOFs as a whole. His eye unit is made of a CMOS camera with an image resolution of 640*480 at a sample rate of 30 frames per second. The camera's field of view is 58 degrees. In addition, stereo loudspeakers can be used to play back synthesized speech based on a text-to-speech module in the system. Since the Nao robot used here becomes a popular platform in various areas, the present study based on this particular robot platform can not only produce results to advance our general knowledge of building smooth human-robot interactions, but also have applied utilities for robot practitioners and educators using Nao to build real-world applications.

We've developed a learning component in the robot through which it was able to process visual and auditory signals in real time. Using this system, we then systematically manipulated the ways that the robot reacted to the human's behaviors based on the robot's own learning state. In the following, we will overview the learning system first and then present the experiment and results.

A. *Perception and Learning in Robot*

As shown in Figure 3, there are three components in the robot's learning system: visual processing, language processing and word-referent mapping. In the following, we will describe each of the three components respectively.

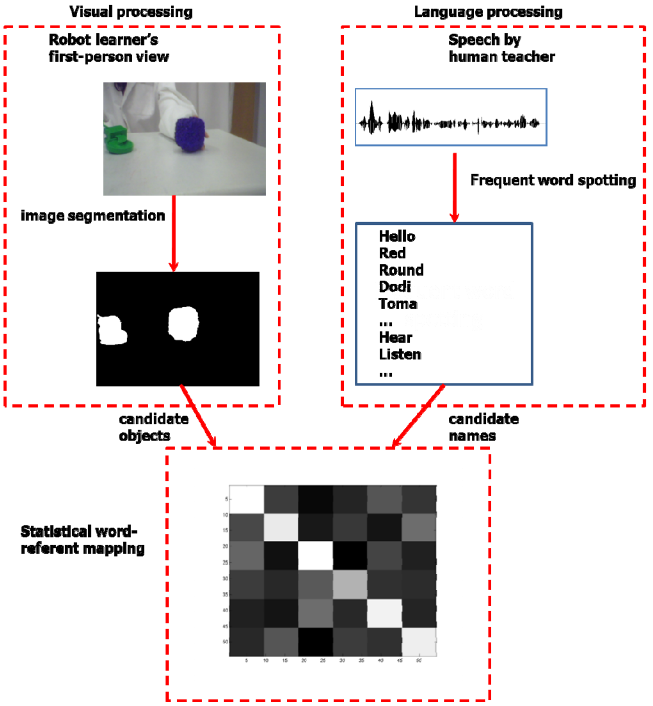

Figure 3. The system of statistical word learning consists of three subsystems: visual processing, language processing and visual-speech integration.

Visual Processing

We covered the interaction environment with white fabrics and chose objects with unique colors which facilitated visual processing. Given raw images collected from the robot's camera on the robot's head, the first step in image processing separates white background pixels from object pixels. In the second step, adjacent nonwhite pixels that have similar color values within a small threshold are grouped into several blobs. The method then attempts to create larger groups from the initial grouping by using a much tighter threshold. This follow-up step determines which areas of an image belong to the same object even if an object is visually segregated into multiple segments as for example when held in a teacher's hand. The third step assigns each blob into an object category. In this object detection task, we use Gaussian mixture models to pre-train a model for each object. By applying each object model to a segmented image, a probabilistic map is generated for each object indicating the likelihood of each pixel in an image as belonging to this specific object. Next, by putting probabilistic maps of all the possible objects together, and by considering the spatial coherence of an object, the detection algorithm assigns an object label for each blob in the segmented image. More technical details can be found in [19]. The whole object detection step on the robot's side took a running time of less than 100ms for each image to provide visual object candidates for language learning.

Language Processing

Figure 4 provides an overview of language processing. The robot perceived human speech in real time. A speech recognition software (Dragon naturally speaking from nuance, LLC) was first applied to convert speech into text. Next, a 10-second temporal window was used to define a local context. Spoken utterances within a context were then compared to spot frequent words that were further processed in two specific ways. First, frequent words were selected as candidate words for object names and would be linked with visual input to compute word-object associations. Second, frequent words were added to a word list that the robot maintained to keep track of those words that the robot heard before. In speech production, the robot would selectively produce those words. In a way, this mechanism made the robot like a copycat – repeating what it just heard most frequently in the recent past. For example, if a human teacher happened to say "hello" to the robot in multiple times, the robot would say "hello" back to the human teacher. Thus, the robot learning system was transparent and straightforward -- purely driven by statistical regularities in the data without complicated inference. The goal was to show how better data from interaction may lead to better statistical learning.

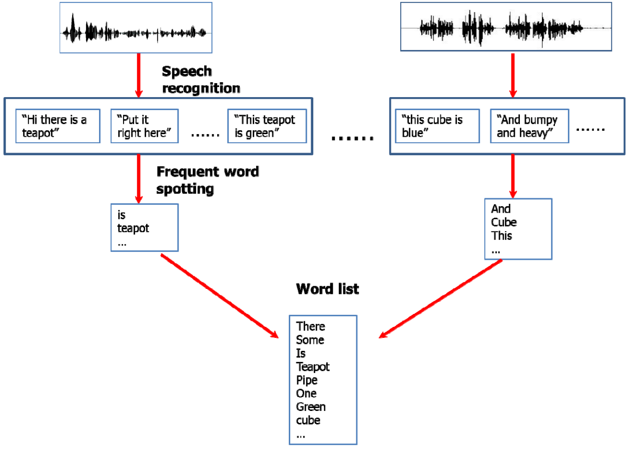

Figure 4. Overview of language processing. A word list is extracted and maintained in the robot's memory. Those words are both candidates for object names and also candidates for robot speech production.

Statistical word-referent mapping

Given word candidates and visual objects in view, associating referents (objects, etc.) with words (object names, etc.) is viewed as the problem of identifying correspondences between the two. With multiple learning situations as shown in Figure 5, there are multiple possible pairs between words and objects wherein some are correct and others are not. The learning system computes association probabilities of all the possible pairs simultaneously and attempts to discover a set of reliable word-referent pairings across words, across objects, and across multiple situations. More specifically, the learning system estimates the association probability of every co-occurring word-referent pair in every learning moment. In this way, a word-referent association matrix shown in Figure 5 is built in which the rows represent all the referents in the training data, and the columns represent all the words in the word list. Each cell indicates the association probability of a specific word-referent pair. If a word-referent pair never co-occurs in any moment, the association probability is set to zero. Otherwise, each pair is considered to be a possible lexical item and its association probability is calculated. Using machine translation techniques, the learning method searches for an overall optimal solution of those individual association probabilities as a whole across all of the learning situations, but not just individual pairings. As a result, some word-referent associations (cells in the association matrix, etc.) with high probability can be viewed as learned pairs. Technical details can be found in [20].

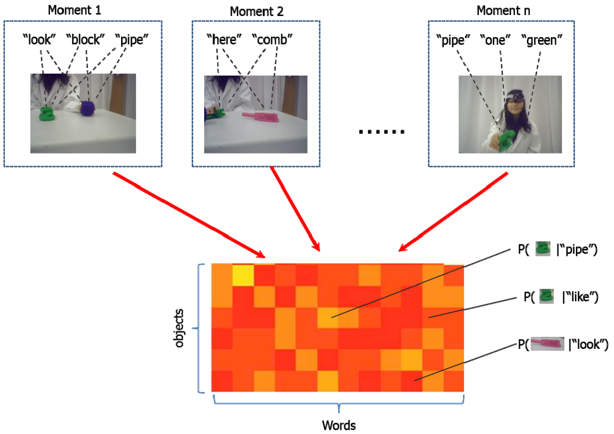

Figure 5. Overview of statistical word learning. The learning system takes multiple pairs of words and objects, one from each learning moment, and computes an association matrix containing all the possible associations between words and objects. Some cells in the matrix are assigned with high probability, which can be treated as learned word-object pairs.

B. *Looking and Talking Acts from Robot*

Through three components described above, the robot learner was able to process visual and audio data to spot frequent words and then build the associations between words and object names. This could be done in a completely passive way without any actions from the robot learner. The new contribution of the study was to link robot word learning with robot behavior. In particular, different from a passive learner, the active robot generated looking and speaking behaviors based on what it has learned.

As described earlier, the robot maintained a word list containing frequent words that the robot has heard multiple times in the recent interaction. When the robot heard one of those words from the human's current speech, the robot would recall that word as a familiar word heard previously and repeat it as a response to human speech. Moreover, those words in the candidate list were also considered to be associated with a particular object. When the robot decided to produce a word, it searched the association matrix shown in Figure 5 to check whether this familiar word was strongly associated to an object. If so, the robot would switch his attention toward that object while producing its name. This looking-while-talking behavior was naturally perceived and interpreted by human teachers as the robot understood the meaning of that spoken word.

The above control strategy created four possible responses when hearing a new utterance from the human teacher in the real-time interaction: 1) no talking or looking initialized by the robot: when the current spoken utterance by a human teacher didn't contain any words that the robot already knew, the robot would not produce any spoken word nor generate any gaze switches; 2) talking without looking: the robot spotted a familiar word from human speech, and therefore the robot repeated that word; however, that word was not strongly associated with any object, therefore no attention switch was made; 3) talking and looking at the target object: the robot produced a word and looked at the target object because the robot found a strong association between the two in its word-referent association matrix; and 4) talking and looking at a wrong target: the robot looked at a wrong target while producing a word; In this case, the target that the robot treated as the referent of the word was incorrect. Here is an example conversation between a human teacher and a robot learner:

Human: the blue object is a violin
Robot: ---
Human: violin (showing the violin)
Robot: ---
Human: can you say violin (showing the violin)
Robot: ---
Human: violin (slowly)
Robot: violin
Human: yes, violin
Robot: violin (looking at the violin)
Human: yes, good job
Robot: good job
Human: good job (with laugh)
……

This example shows that the overall interaction was smooth through which the robot gradually picked up lexical information based on statistical information provided by the human teacher.

To summarize, we equipped the robot learner with basic perception, learning and action skills. In the experiment described next, we systematically manipulated the robot's actions based on its learning states, and measured how this might change teaching behaviors from human teachers, and therefore learning outcomes in robot.

III. Experiment

There were three experimental conditions in which the robot attempted to learn new words and the meanings of the words from a human teacher. The differences between the three conditions lied in what trigged the robot's looking and talking actions in the interaction:

- **Passive**: the robot didn't generate any speech or attention switches among visual objects. Instead, it just passively perceived information while the teacher attempted to teach the robot object names. This condition served as a baseline to measure the results from passive statistical learning alone.
- **Moderately active**: the robot generated spoken words and also looked at the target object associated with spoken

words. However, it did so only after the robot has accumulated enough knowledge of words and referents. Therefore, the robot learner in this condition would not talk until after accumulating lots of statistical information about words and objects. And when it started talking, it was likely to say meaningful and relevant words, and also looked at the correct target objects. In this version, the robot didn't generate lots of activities at the beginning of the interaction (no talking nor looking) but started to be much more active at the end (talking while looking at the correct targets). This moderately active robot can be viewed as a rational and cautious learner – not saying much but always saying it right.

- **Highly active**: the robot tended to generate speech and look at objects very often, even from the beginning wherein it might not have enough statistical data to extract meaningful words and objects. Therefore, the robot learner might produce words that were less meaningful and irrelevant to the learning task (e.g. "there", "look") by simply repeating what the human teacher just said without context. Hence, there were lots of talking without looking, and talking while looking at wrong objects. However, with more statistical evidence accumulated, the robot would gradually produce more appropriate words and looks. Through the whole interaction, the robot was talkative and also actively switched to look at candidate objects. A highly active robot could be viewed as a young child who is active with high energy but can be "annoying" sometimes as what it says may or may not make sense. Nonetheless, just as a young human learner, the highly active robot was not afraid to produce many talking and looking behaviors.

Two primary questions in the present study are: 1) how different behaviors from the robot may alter what human teachers behave in language teaching; and 2) how this may ultimately lead to different learning results in robot. Using the passive condition as a baseline, our main interest is to compare how different degrees of activity from the robot learner may influence interaction and learning. One hypothesis is that human teachers may like both moderately and highly active robots more than the passive one as activities and responses from the robot are critical components in social interaction to create coordinated behaviors between the two social partners. In addition, participants may like the moderately active robot most as it is calm and rational. In contrast, more inappropriate looking and talking from the highly active robot may be viewed as "annoying" and therefore disrupt interaction. Alternatively, more looks and more speech may better engage human teachers. Just like how young children interact with their parents to learn their native language, it is beneficial for a robot learner to generate more acts, even with immature behaviors such as producing non-sense words or looking at random targets.

21 students at Indiana University participated in the study. They were divided into three experimental conditions with 7 participants in each condition. Participants were given four sets of four novel objects, with a total of 16 objects. More specifically, each set contained one blue, one green, one red and one pink object. Each object was given a name that roughly matched with the overall shape of the object, i.e. pipe, comb, or violin. Participants were provided with those object names and asked to memorize them in advance before they entered the experiment. Each participant was asked to teach the robot in four trials with each trial lasting around 2 minutes. Trial orders were randomized across participants. At the end of each trial, an experimenter signaled participants to stop and asked participants to take a voluntary break before starting a new trial with a new set of four objects. The whole interaction was free flowing without any particular instructions to participants on what they should do and what they should say to the robot.

IV. RESULTS

With the learning system run in real time, including online image processing and online language processing, we collected multimodal data during the interaction, which consisted of: 1) human speech; 2) robot speech; 3) visual information from the robot's view and 4) word learning results. Our data analyses focused on two perspectives in the interaction: 1) teaching: how human teachers might behave differently; and 2) learning: how robot learners in the three conditions might vary in their learning performance.

A. Teaching Behavior from Human

In this section, we first report linguistic acts from human teachers followed by the results derived from non-linguistic acts.

1) Speech Act

A summary of a set of measures is presented in Table I. Overall, in terms of the proportion of talking time (in the 2nd row of the table), human teachers tended to talk to the robot learner more when the robot learner was more active. This result may seem to be not intuitive as we know turn-taking is a reliable pattern in speech conversation, including in human-robot interaction. If human teachers would not talk when the robot was talking, then more talking from the robot should lead to more listening and less talking from human teachers. However, we noticed that participants spent only 21% of time talking to the robot in the passive condition, with 18.75 spoken utterances per minute. There were long silences in which participants were not much engaged in teaching the robot. Without any responses from the robot, participants in the passive condition might just hesitate to take next actions. Instead, they probably anticipated some responsive behaviors from the robot before they proceeded. When the robot became more active with talking and looking acts, even with the turn-taking principle that both social partners intended to follow, those behaviors from the robot broke the ice as participants produced 34.63 spoken utterances per minute which is almost doubled compared with the number in the passive condition. As shown in Table I, we found no difference in speech length across the three conditions. The significant difference in talking time was mostly caused by the number of spoken utterances produced. Not only did participants generate more spoken utterances in the highly active condition, they also produced more naming utterances (14.25) compared with what they did in the moderately active (9.03) and passive (7.65) conditions. Meanwhile, mean length of naming utterance (M_{highly}=0.53; $M_{moderately}$=0.52; $M_{passive}$=0.50) was similar in the three conditions. In addition, we measured the number of unique words (tokens, etc.) and found a significant difference between the three conditions shown in the last row of Table I. Thus, participants not only produced more words, but also used different kinds of words in the highly active condition as the size of vocabulary increased from 103 to 142. Using different words and naming objects in different ways may facilitate learning as we will show next.

Table I: A summary of human speech. Means, standard deviations (in parentheses) and statistical results are reported. "n.s" stands for not (statistically) significant

	Highly active	Moderately active	passive	statistics
Prop of talking time	36.12% (8.28%)	0.26.82% (7.82%)	21.36% (6.24%)	F(2,18) =8.22; P<0.001
Freq. of spoken utterances	34.63 (7.28)	25.57 (6.25)	18.75 (5.78)	F(2,18) = 9.74; P<0.001
length of spoken utterance	0.63 (0.23)	0.61 (0.18)	0.64 (0.35)	n.s.
Freq. of naming	14.25 (2.52)	9.03 (2.01)	7.65 (1.56)	F(2,18) = 7.48; p<0.005
length of naming (sec)	0.53 (0.28)	0.52 (0.22)	0.50 (0.25)	n.s.
vocabulary size	142 (24)	110 (32)	103 (25)	F(2,18) = 3.65 P<0.05

2) Non-Linguistic Behavior

Now that we know more naming utterances were created by participants in the highly active condition compared with the other two conditions. We next zoomed into those naming moments to further evaluate the quality of teaching signals from human teachers. We measured the size of a named object at naming moments when a human teacher produced a spoken utterance containing an object name. As shown in Figure 6, a comparison between the three conditions suggests that the target object was larger and more dominant in the highly active condition compared with the other two conditions ($t_{moderately}$(13)=3.24,p<0.01; $t_{passive}$(13) =4.11 , p<0.005). On average, across all the moments in the interaction, a visual object occupied 2.12% of the robot's view. Therefore, in both active conditions, the target object was larger than other objects in view (t_{highly}(13) =5.80 , p<0.001; $t_{moderately}$(13)=2.64,p<0.01) while the target object was not significantly different with other objects in the passive condition($t_{passive}$(13) =0.59, p=0.28). Human teachers in the highly active condition not only produced more naming events; they were also selective for when to name an object. The naming moments chosen were those that the target object was dominant the robot learner's view. A central challenge in word learning is reference uncertainty – given a learning moment with several words and several objects, there are many possible associations between candidate words and candidate objects. However, if the human teacher labels an object when that object is visually salient, this teaching behavior can significantly reduce the uncertainty problem in word-to-object mapping [21]. In the present experiment, we show that human teachers in the highly active condition tended to do so, which provided an external solution of the uncertainty problem through teacher-learner interactions.

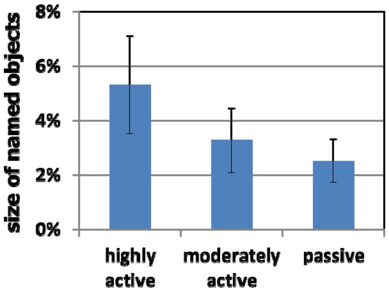

Figure 6.The mean size of target objects at naming moments across three experimental conditions.

B. Learning outcome in robot

Different behaviors from the robot led to different teaching behaviors from human teachers. Our conjecture is that human teachers interacting with more active robots provided better teaching signals. To confirm this hypothesis, the next step is to measure whether and how the learning system may take advantage of better training data provided by the teacher.

In the experiment, the robot has discovered and maintained a set of familiar words and stored them in a word list. Only a word produced in human speech frequently and repeatedly was chosen to be a familiar word. Then the robot linked those familiar words with objects through building associations between words and objects. Since our learning system was run in real-time interaction, we can directly access the learning outcome at the end of the interaction.

To measure word learning results, we found 16 object names in the association matrix accumulated in the interaction (as shown in Figure 5) and computed the mean association strength of those target words. As shown in Figure 7, target word-referent pairs have much higher association probabilities in the highly active condition compared with the other two conditions ($t_{moderately}$(13)=3.92,p<0.005; $t_{passive}$(13) =4.28, p<0.001), showing that the highly active robot has successfully learned those word-object pairs.

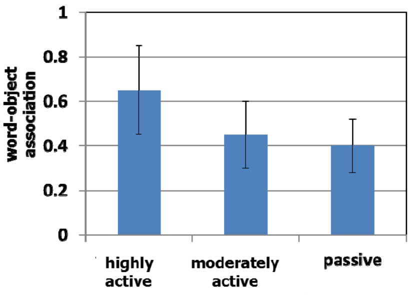

Figure 7. Association probabilities between spoken object names and target objects in three experimental conditions.

In addition to access the association matrix as an internal representation of word learning, learning results can also be evaluated through the robot's external behavior. Producing an object name while looking at the target object was a clear demonstration that the robot learner knew the referent of the word, because the robot learner did so only after it built a strong association between an object and its linguistic label. In the two active conditions, the robot produced spoken words from its word list as a way to demonstrate its knowledge to those words. Given that the highly active robot in general produced more words than the moderately active one as a part of our experimental design, we measured instead the relative proportion of object names produced, normalized by the overall number of spoken words. The highly active robot produced not only more words overall but a larger proportion of spoken words were object names (M=42.83%), compared with the moderately active robot (M=27.21%, t(13)=2.87,p<0.01), suggesting more successful learning from the highly active robot.

V. GENERAL DISCUSSION

In the present study, we argue that one of the strongest demonstrations of linking interaction and learning is to show that even given the same learning algorithm – without any changes to process additional feedbacks from human users, the learning system can improve its performance through social interaction. That is, social behaviors from a robot learner make a human teacher provide better and more teaching signals, and by doing so, better and more training data through social interaction lead to better statistical learning. The results can be explained from both cognitive and social perspectives. At the cognitive level, on-demnad information provided by human teachers can directly enter the learning system to improve learning. At the social level, more activities from the robot engage human teachers to more actively interact with the robot. In the following, we will discuss those two aspects respectively.

First, how exactly did better data lead to better learning with the same learning mechanism across the three conditions? A close examination can categorize the robot's looking and talking behaviors into two common cases. First, at the beginning of interaction, the highly active robot started talking without accumulating enough statistical information yet. Therefore, the robot was likely to produce seemingly non-sense words and also generate more-or-less random looks on objects. This kind of behavior provided clear signals for human teachers on what the robot did or didn't know –its internal learning states. If the robot looked at object A while naming it as object B, or looked at object A while producing an irrelevant word (e.g. function words), then the human teacher was likely to correct that by showing the right object to the robot and naming it. The learning system can then use such clear information to update its knowledge. This can be viewed as a case of learning from making mistakes with three consequential steps: 1) the robot named a wrong object; 2) the human teacher noticed that and corrected it; and 3) the robot's learning system took correct statistical data and updated its lexical knowledge. Different from self-correction, in the context of human-robot interaction, it relied on human teachers to correct those mistakes by providing correct signals through interaction. In the second case, when the robot happened to name an object correctly while switching attention on the target object, the human teacher was likely to echo that and encourage the robot learner to do so. Thus, both correct and incorrect behaviors from the robot can elicit useful feedbacks from human teachers. This is accomplished without adding any components in the learning system to explicitly use feedback signals. Instead, this solution worked well because it counted on better statistics provided by the teacher; and moreover, such statistical information was provided as needed and as soon as possible in real-time learning. As shown in Figure 1, we can view this as active learning through interaction [22], in which the robot learner generated behaviors to reveal its current learning states, and those behaviors gave human teachers first-hand information about what the robot learner needed next for successful learning. In summary, real-time learning using on-demand information elicited through human-robot interaction can lead to successful statistical learning without complicated internal algorithms.

Second, why were human teachers willing to provide what the robot needed in the interaction? More generally, how can a robot learner successfully elicit better teaching signals from human teachers? In the case of human language learning, when communicating with young children, parents tend to speak slowly, with high pitch and hyperarticulation. They also tend to use repeated and simple words [23]. As a result, child-directed speech certainly provides better signals for language learning compared with overheard speech from adult conversation or TV programs. In the context of training robots to learn a language or other cognitive skills, if a robot learner, through its appearance and its behavior, makes human teachers treat the robot as an immature learner, human teachers would be willing to adjust their behaviors to interact with and teach the robot. In the present study, it is unexpected that human teachers were not distracted by non-sense words and random looks generated by the highly active robot. Instead, those immature behaviors somehow engaged and motivated human teachers to provide better teaching signals, suggesting that in the context of learning, immature behaviors demonstrated by the learner may facilitate learning through teacher-learner interactions. More interaction means more data for learning; and better interaction means better data. Of course, this idea may not work in general human-robot interaction in which we want to design robots to be co-workers and assistants. However, there are many applications in which robots cannot be equipped with all the knowledge and skills in advance. Instead, it needs to learn from interacting with human users. Hence, it is an important topic to understand how to make such learning possible and effective. Toward this goal, instead of focusing on how to develop complicated machine learning algorithms that are able to infer knowledge from noisy data, we should also think of how to design better human-robot interactions and interfaces through which robot learners can elicit and gather better teaching signals from human teachers. It is not surprising that better data lead to better learning. But the important lesson here is to emphasize on how to elicit better teaching signals from human users, which can be an important direction of robot learning in social contexts.

We note that the present findings are derived from one word-learning task with a specific experimental setup. The goal is to make a convincing and clear case to demonstrate important connections between interaction and learning. For that purpose, we intentionally designed and implemented the same learning machinery across multiple experimental conditions, and showed that better learning can be achieved through better interaction. Of course, learning algorithms that can better capture social signals in the interaction will lead to even better learning [24]. In the present work, we are less interested in obtaining the best learning results using the best algorithms available. Instead, we are more interested in proposing and understanding fundamental principles and ideas of linking interaction with learning, such as active real-time learning, and elicitation of humans' teaching signals for learning. Those principles have the potential to apply to different applications and learning tasks. More generally, we suggest that a deep theoretical understanding of how a robot can learn to interact, and interact to learn, what principles such learning system should have, and how incorporate such principles in different contexts and different tasks, will provide useful guides for future HRI design.

VI. Acknowledgment

We thank Hong-Wei Shen and Amanda Favata for help with running experiments. This research was supported by NIH R01HD074601and NSF BCS 0924248.

References

[1] H. H. Clark and M. A. Krych, "Speaking while monitoring addressees for understanding," *Journal of Memory and Language,* vol. 50, pp. 62-81, 2004.

[2] M. Scheutz, P. Schermerhorn, and J. Kramer, "The utility of affect expression in natural language interactions in joint human-robot tasks," in *Proceedings of the ACM SIGCHI conference on Human-robot interaction*, 2006, pp. 226-233.

[3] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Transaction on Systems, Man and Cybernetics, Part A: Systems and Humans,* vol. 35, pp. 460-470, 2005.

[4] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: how robots might shape participant roles using gaze cues," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009, pp. 61-68.

[5] M. Heerink, B. Kröse, B. Wielinga, and V. Evers, "Enjoyment intention to use and actual use of a conversational robot by elderly people," in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 2008, pp. 113-120.

[6] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing* vol. 999: MIT Press, 1999.

[7] C. S. Tamis-LeMonda, Y. Kuchirko, and L. Tafuro, "From Action to Interaction: Infant Object Exploration and Mothers' Contingent Responsiveness (June 2013)," *IEEE Tranactions on Autonomous Mental Development,* vol. 5, pp. 202-209, 2013.

[8] K. L. Marsh, M. J. Richardson, and R. Schmidt, "Social connection through joint action and interpersonal coordination," *Topics in Cognitive Science,* vol. 1, pp. 320-339, 2009.

[9] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems,* vol. 57, pp. 469-483, 2009.

[10] Y. Nagai, C. Muhl, and K. J. Rohlfing, "Toward designing a robot that learns actions from parental demonstrations," in *IEEE International Conference on Robotics and Automation (ICRA 2008)*, 2008, pp. 3545-3550.

[11] Y. Nagai, Y. Kawai, and M. Asada, "Emergence of mirror neuron system: Immature vision leads to self-other correspondence," in *IEEE International Conference on Development and Learning (ICDL)*, 2011, pp. 1-6.

[12] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence,* vol. 172, pp. 716-737, 2008.

[13] J. de Greeff, F. Delaunay, and T. Belpaeme, "Active robot learning with human tutelage," in *IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 2012, pp. 1-6.

[14] A.-L. Vollmer, K. S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, *et al.*, "People modify their tutoring behavior in robot-directed interaction for action learning," in *IEEE International Conference on Development and Learning*, 2009, pp. 1-6.

[15] K. Pitsch, A.-L. Vollmer, and M. Muhlig, "Robot feedback shapes the tutors presentation: How a robots online gaze strategies lead to micro-adaptation of the humans conduct," *Interaction Studies,* vol. 14, pp. 268-296, 2013.

[16] C. Lyon, C. L. Nehaniv, and J. Saunders, "Interactive language learning by robots: The transition from babbling to word forms," *PloS one,* vol. 7, p. e38236, 2012.

[17] A. L. Thomaz and C. Breazeal, "Experiments in socially guided exploration: Lessons learned in building robots that learn with and without human teachers," *Connection Science,* vol. 20, pp. 91-110, 2008.

[18] H. H. Clark and S. E. Brennan, "Grounding in communication," *Perspectives on socially shared cognition,* vol. 13, pp. 127-149, 1991.

[19] C. Yu, L. B. Smith, H. Shen, A. Pereira, and T. Smith, "Active Information Selection: Visual Attention Through the Hands," *IEEE Transactions on Autonomous Mental Development,* vol. 2, pp. 141–151, 2009.

[20] C. Yu and D. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions," *ACM Transactions on Applied Perception (TAP),* vol. 1, pp. 57-80, 2004.

[21] C. Yu and L. B. Smith, "Embodied Attention and Word Learning by Toddlers," *Cognition,* vol. 125, pp. 244-262, 2012.

[22] T. M. Gureckis and D. B. Markant, "Self-Directed Learning A Cognitive and Computational Perspective," *Perspectives on Psychological Science,* vol. 7, pp. 464-481, 2012.

[23] Y. Nagai and K. J. Rohlfing, "Can motionese tell infants and robots" what to imitate"," in *Proceedings of the 4th International Symposium on Imitation in Animals and Artifacts*, 2007, pp. 299-306.

[24] C. Yu and D. H. Ballard, "A unified model of early word learning: Integrating statistical and social cues," *Neurocomputing,* vol. 70, pp. 2149-2165, 2007.