Children Use Non-Verbal Cues to Learn New Words From Robots as well as People

Jacqueline M. Kory Westlund[a]*, Leah Dickens[b], Sooyeon Jeong[a],

Paul L. Harris[c], David DeSteno[b], & Cynthia L. Breazeal[a]

[a]MIT Media Lab, 20 Ames St., Cambridge, MA 02139, USA

[b]Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA

[c]Harvard University, 13 Appian Way, Cambridge, MA 02138, USA

*Corresponding author:

jakory@media.mit.edu

Abstract

Social robots are innovative new technologies that have considerable potential to support children's education as tutors and learning companions. Given this potential, it behooves us to study the mechanisms by which children learn from social robots, as well as the similarities and differences between children's learning from robots as compared to human partners. In the present study, we examined whether young children will attend to the same nonverbal social cues from a robot as from a human partner during a word learning task, specifically gaze and bodily orientation to an unfamiliar referent. Thirty-six children viewed images of unfamiliar animals with a human and with a robot. The interlocutor (human or robot) oriented toward, and provided names for, some of the animals, and children were given a posttest to assess their recall of the names. We found that children performed equally well on the recall test whether they had been provided with names by the robot or by the human. Moreover, in each case, their performance was constrained by the spatial distinctiveness of nonverbal orientation cues available to determine which animal was being referred to during naming.

Keywords

Preschool children; non-verbal cues; social robots; word learning; reference

1 Introduction

Social robots are innovative new technologies that have considerable potential to support children's education as tutors and learning companions. Social robots share physical spaces with us and leverage our means of communication—e.g., speech, gestures, gaze, and facial expressions—in order to interact with us in more natural, intuitive ways. They have the potential to combine the general benefits of technology—such as scalability, customization and the easy addition of new content, and student-paced, adaptive software—with the embodied, social world. Prior research has shown that young children will not only treat social robots as companions and guides (Belpaeme et al., 2012; Kahn et al., 2012; Shiomi et al., 2015), but will also readily learn new information from them (Breazeal et al., 2016; Kennedy et al., 2016; Movellan et al., 2009; Tanaka and Matsuzoe, 2012; Vouloutsi et al., 2016).

Given this potential, it behooves us to study the mechanisms by which children learn from social robots, as well as the similarities and differences between children's learning from robots as compared to human partners. Some existing work investigates these differences. Kennedy et al. (2016) examined the effects of a human tutor versus a humanoid robot tutor on learning prime number categorization with children aged 8-9 years. With both tutors, children's scores on the math task improved from pretest to posttest, but the human led to a greater effect size than the robot. Serholt et al. (2014) compared the attitudes, success rate, and help-asking behaviors of children aged 11-15 years during a LEGO construction with either a humanoid robotic tutor or a human tutor. They found that while children with either tutor successfully completed the task, children were more likely to ask the human tutor for help, but were more eager to perform well with the robot tutor. These studies suggest that there may be important differences in how children treat human and robot tutors. However, both these studies were performed with older children. There is growing interest in developing social robots as tutors and learning companions for younger children aged 3-6 years (e.g., Breazeal et al. 2016; Kory, 2014; Kory & Breazeal, 2015; Movellan et al., 2009; Tanaka & Matsuzoe, 2012). How might these children respond? Furthermore, Kennedy et al. (2016) points out that they did not constrain the human's social behavior. For some kinds of learning tasks such as language learning, social cues may be very important (Kuhl, 2009; Kuhl, 2011). How do social cues impact children's learning from robots versus from humans? Do children respond to social cues from humans and robots in the same way?

In the present study, we examined whether young children will attend to the same social cues from a robot as from a human partner during a word learning task, specifically gaze and bodily orientation toward a novel referent.

Infants and young children are adroit at following another person's gaze and that capacity makes an important contribution to early social cognition. For example, gaze following helps infants and young children to determine what object or event has triggered another's emotion (Harris & Lane, 2014; Moses et al., 2001). Gaze following can also provide information about the goal of an agent's ongoing action (Phillips, Wellman & Spelke, 2002). In addition, following a speaker's gaze can provide information about his or her intended referent, facilitating the task of word learning. Baldwin (1991, 1993) demonstrated the key role of gaze following for language learning in a series of experiments with infants of 19-20 months. When infants heard a novel label, they did not immediately associate it with the object that they were concurrently looking at or exploring. Instead, by following the speaker's line of regard, they were able to determine what object the speaker was attending to and to link the novel name provided by the speaker with that referent. Recent findings have also shown that infants more readily associate names with novel objects if the speaker's gaze is directed to an object that is presented in a distinctive and consistent spatial locus (Samuelson, Smith, Perry & Spencer, 2011). By implication, infants treat a speaker's gaze direction as a major index of the particular target that is being named by the speaker within a shared space.

Granted the early importance of gaze following in human social interaction, investigators have begun to examine whether, and under what conditions, young children will follow a robot's gaze. Meltzoff, Brooks, Shon and Rao (2010) presented 18-month-old infants with a humanoid robot (HOAP-2, manufactured by Fujitsu Laboratories, Japan) that behaved in one of four ways. Infants in the *social interaction* group observed the robot as it interacted with an adult experimenter. In the course of the interaction, the robot answered the adult's questions and the two parties engaged in mutual imitation. By contrast, infants in the three other groups observed an interaction in which the impression of contingent, two-way communication between robot and adult was eliminated, because the adult remained stationary (*passive adult* group) or because the robot remained stationary (*passive robot* group) or because the gestures and utterances of the two parties were not aligned with each other (*robot-adult mismatch* group).

Following this observation period, infants' tendency to follow the gaze and bodily orientation of the robot was assessed. As they faced one another, the robot turned through 45° to look at an object located on either the left or right side of the infant. Infants who had observed the robot engage in social interaction with the adult were likely to shift their gaze to match the target that the robot was looking at, whereas the other three groups responded unsystematically. By implication, having observed the robot's capacity for contingent, social interaction, infants construed the robot as a partner or informant whose gaze signaled targets that were worth looking at.

Granted that infants can and do follow a robot's gaze, it is plausible to ask whether young children will make use of a robot's line of regard when learning new words, as they do with human partners. More specifically, when a robot introduces a name, are children able to use line of regard to determine which particular object the robot is naming and thereby learn the name of the object? To begin to answer this question, O'Connell, Poulin-Dubois, Demke and Gray (2009) presented 18-month-old infants with two learning trials involving pairs of novel objects. Infants heard a robot offer a name for one of the paired objects. In the coordinated labeling condition, the robot uttered a novel label only when both the infant and the robot were focused on the same novel object whereas in the discrepant labeling condition, the robot uttered a novel label when focusing on a different object from the infant. Infants were subsequently tested to check if they had associated the name with the appropriate object, notably the object that the robot had focused on. They were shown the two novel objects and asked a comprehension question (e.g., "Where is the dax?").

Analysis of infants' attention during object naming indicated that they adjusted their gaze appropriately depending on the gaze direction of the robot. Thus, in the discrepant labeling condition, infants were prone to shift their gaze so as to focus on the same toy as the robot, a coordination that was present by default in the coordinated labeling condition. Nevertheless, infants performed at chance in the comprehension test following both conditions. By contrast, in a follow-up study, in which a human rather than a robot served as the speaker, infants not only adjusted their gaze, they also performed well in the comprehension test. Finally, in a third study, infants were re-tested with a robot but before proceeding to the word learning phase, they were given an opportunity to watch a 60-sec interaction in which the robot's utterances and movements were contingent on the immediately preceding behavior of an adult. Despite this opportunity, infants continued to perform at chance in the comprehension test. Accordingly, O'Connell et al. (2009) speculated that despite their tendency to follow the robot's gaze, infants did not think of the robot a reliable or conventional speaker from whom it is appropriate to learn new words.

Two aspects of the study by O'Connell et al. (2009) may have led infants to fail to learn new names from the robot. First, it is unclear whether the infants perceived the robot as an interlocutor with whom they could interact. During the familiarization phase, infants had only a brief opportunity to observe the robot communicate with an adult. It moved independently (turned its head) and vocalized (said "hello" and "ooh"). However, this may not have been sufficient for infants to regard the robot as a speaker from whom they could acquire language. Prior research suggests that a speaker's contingent responding to the learner appears to play a key role in early language acquisition. For example, Kuhl (2007) found that although infants will readily learn to differentiate new phonemes when they are presented by a live and contingent interlocutor, they fail to do so if they simply observe a video of the same interlocutor engaged in a conversation that is not directed at them. A second concern with the study conducted by O'Connell et al. (2009) is that they tested 18-month-olds. In a series of studies, Horst and Samuelson (2008) showed that, even at 24 months, infants can use a speaker's gaze to map a novel name onto the appropriate referent but display poor retention of that name on subsequent retention tests.

Accordingly, in the study to be reported, we made two changes aimed at giving the robot the best opportunity to serve as a teacher of language for young children. First, guided by previous research, we tested older children. Second, we sought to ensure that the robot would be perceived as a contingently responsive interlocutor for both the child and the experimenter in the context of an initial three-way conversation. We describe these two changes in more detail below.

In prior work, we investigated whether children aged 4-6 years displayed fast mapping when interacting with a robot or an adult (Kory Westlund et al., 2015). The study was modeled on a procedure adopted by Markson and Bloom (1997) in which preschoolers displayed stable retention. The procedure lends itself easily to either child-caregiver interaction or child-robot interaction. Children viewed a series of ten pictures of unusual animals, one picture at a time, with each interlocutor. For eight of the ten pictures, the interlocutor commented positively but uninformatively (e.g., "Cool animal!"); for the other two pictures, the interlocutor provided the name of the animal shown (e.g., "Look, a binturong! See the binturong?"). In this word-learning task, children learned equally well from the robot and the adult. Thus, they performed similarly in an immediate comprehension test and a retention test one week later, showing that they did remember the animal names.

In the present study, we created a similar but more challenging task that required children to attend to their interlocutor's gaze direction and bodily orientation in order to identify the referents of the new words, a task that more closely mirrors everyday language learning. Each child was shown multiple pairs of pictures depicting unfamiliar animals. For selected pairs, the child's interlocutor named one of the two animals. We asked if children would use the robot's gaze and bodily orientation, just as they do with humans, to identify which of the two animals was being named.

To increase the likelihood that children would regard the robot as an interlocutor from whom they could learn new names, the robot engaged in a brief conversation before name learning began. First, the experimenter spoke directly to the robot to introduce the child to the robot – thereby showing the child that the robot was an interaction partner. The robot then introduced itself to the child. The experimenter asked if the child and robot were ready to look at some animals, to which the robot expressed interest. The robot then invited the child to look at the pictures. The same procedure was followed when the adult female was the interlocutor to maintain consistency across conditions.

The study was also designed to find out how spatially distinctive the nonverbal cues had to be in order for children to identify which referent the interlocutor intended. Half the children were presented with pairs of animal pictures that were side-by-side whereas the other half were presented with pairs of animal pictures that were further apart. This enabled us to assess whether the ease with which children could differentiate the target of their interlocutor's naming would affect their learning from both a robot and a human. If children were to display a similar pattern of variation – depending on the spatial distinctiveness of the non-verbal cues – when learning from each interlocutor, this would strengthen the claim that children rely on such cues when learning new vocabulary whether from a robot or a human.

2 Methods

2.1 Participants

Thirty-six children aged 2-5 years (22 female, 14 male) from two Boston-area, English-language preschools serving predominantly middle-class neighborhoods participated in the study (17 children from one school; 19 from the other). One 4-year-old girl and one three-year-old boy were removed from analysis because they did not complete the study. The children in the final sample included 17 children from each school aged 2-5 years (21 female, 13 male), with a mean age of 3.69 years (*SD* = 0.908). Of these, 4 were 2-year-olds, 10 were 3-year-olds, 15 were 4-year-olds, and 4 were 5-year-olds. We do not have age data for one child.

2.2 Procedure

The experiment followed a 2 x 2 mixed design. Each child viewed pairs of pictures of unfamiliar animals with (1) a robot, and (2) an adult. The order of exposure to each interlocutor was counterbalanced such that half the children first viewed six pairs with the robot and then six pairs with the adult, whereas the other half first viewed pairs with adult and then with the robot.

Half the children viewed the pairs of animal pictures either side-by-side on the same tablet screen, making it hard to differentiate between them based on the interlocutor's direction of gaze and bodily orientation. The other half viewed the pairs of animals on two separate tablets, separated by approximately 15 inches, making it relatively easy to determine which was being referred to using direction of gaze and bodily orientation.

Each child participated in one 10-15 minute session, which took place in a quiet room at the child's preschool. As a warm-up, a female experimenter first showed each child pictures of the robot, a female adult, and a tablet device and asked questions about whether they thought the robot was more like the person or the tablet.

Next, the children were introduced to the first interlocutor (female adult or the robot), who began the interaction by looking at the child. Before beginning the task, the children's interlocutor (robot or person) engaged the children in a brief conversation in which the interlocutor introduced herself, expressed interest in animals, and suggested they look at some animal pictures together. Each interlocutor engaged in the same conversation with the child prior to looking at pictures (thus the child conversed with both interlocutors). The children then looked at two series of six pairs of pictures of unfamiliar animals, one pair at a time. Each pair of images was shown for ten seconds, either, as noted above, side-by-side or separated by about 15 inches on the table. Then the screen went blank, followed by the next pair of images appearing. The children viewed six pairs of pictures with a female adult and six pairs of pictures with the robot, for a total of twelve pairs (twenty-four individual pictures). The order in which the picture pairs were presented was held constant across participants. The experimenter was present in the room the entire time, seated behind the child out of sight. The experimenter stayed silent during the task.

During the display of the pictures, the interlocutor (robot or person) commented positively but uninformatively on one of the animals shown for 3 of the 6 pairs, e.g., “Look at that!” or “Cool animal!” but named one member of the other 3 pairs, e.g., “Ooh, a kinkajou! See the kinkajou?” These naming episodes presented children with a single, brief opportunity to learn the names of three novel animals. During the naming, the interlocutor looked down at the relevant picture, glanced up at the child, then resumed looking at the picture. Thus, for each pair of pictures with a named animal, the sequence was: 1) animal pictures are shown; 2) interlocutor looks at the named animal and says, “Ooh, a kinkajou”; 3) interlocutor glances at child and says “See the”; 4) interlocutor returns gaze to animal and finishes “kinkajou?” When neither animal in the pair was named, the sequence was: 1) animal pictures are shown; 2) interlocutor looks at a preselected animal in the pair and comments, e.g., “Cool animal!”; 3) interlocutor glances at child; 4) interlocutor returns gaze to animal. When looking at a picture, the interlocutor also turned the head slightly to orient toward the picture (rather than just using the eyes). Because the robot did not have a separate head joint that could turn, it turned its whole body slightly toward the picture. The robot also did not have arms or hands, so it could not point. To equate the human to the robot, the human interlocutor did not point at the pictures either. The animal pictures and the animal names are listed in the Appendix.

After observing the pictures with an interlocutor, the experimenter tested children's learning of the new names (one test after seeing pictures with the robot, another after seeing pictures with the human). With respect to each of the three named animals by the interlocutor, children were shown four animal pictures. One of these four animals was the named animal, one was the unnamed animal that had been shown in the same pair as the named animal, and two animals had been seen but unnamed. Children were asked to point to the named animal, for example, "Can you show me the kinkajou?"

Subsequently, as part of a separate study, children performed the Anomalous Picture Task with the person and with the robot, were asked more questions probing their conception of robots, and were invited to engage in an unscripted dialogue with the robot and with the adult. These data are not considered in the present report. Thus, details regarding these tasks are listed in the Appendix.

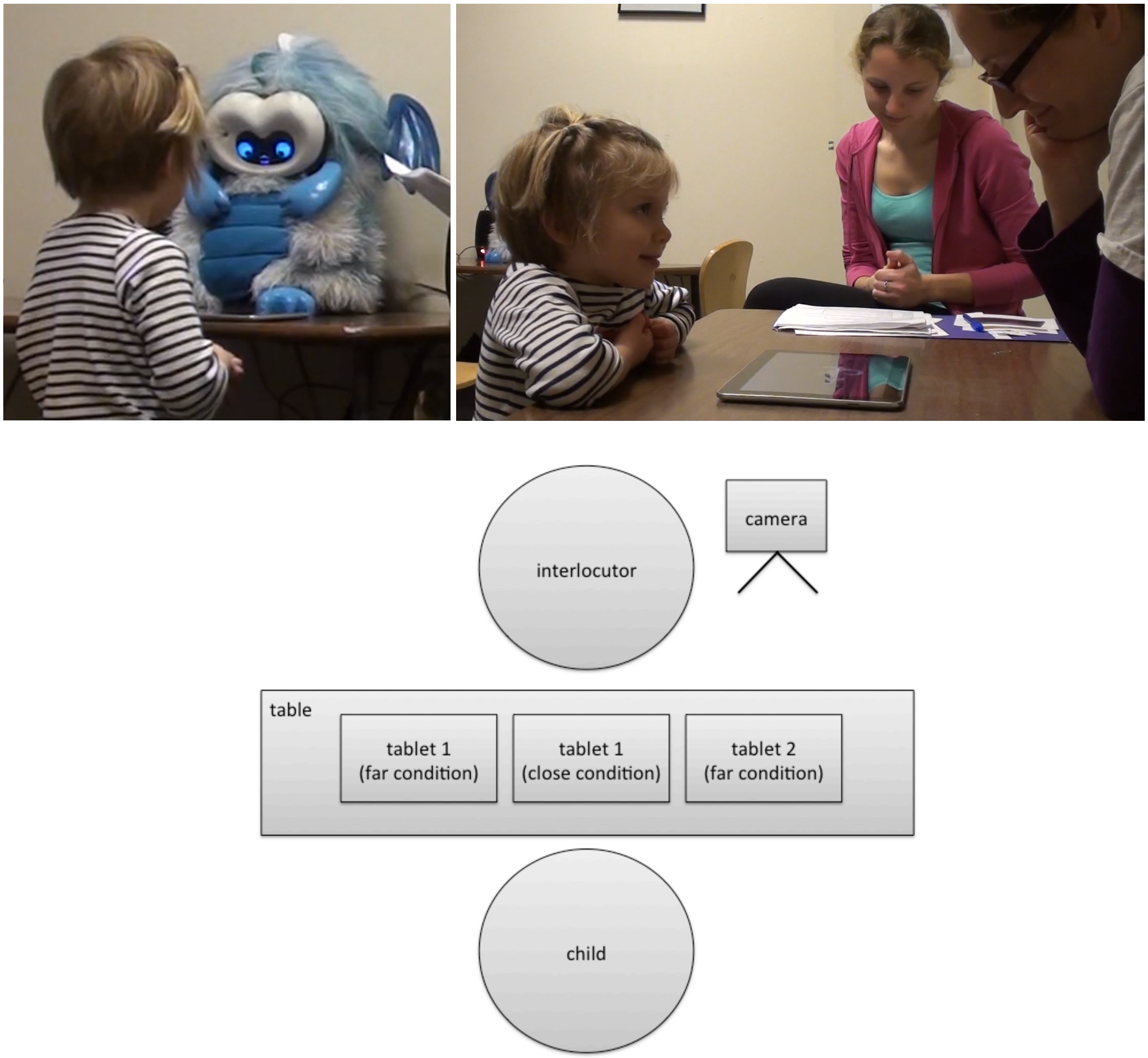

Figure 1: Children viewed pairs of pictures of unfamiliar animals on a tablet with either the DragonBot or with an adult. The bottom figure shows the study setup. On the top left, the robot gazes at one of the animals and provides the animal’s name. The child needs to follow the robot’s gaze to identify which animal in the pair is being named. On the top right, the child looks at the human partner as the animal is named.

2.3 Materials

Samsung Galaxy Tab 2 10.1” tablets were used to present the animal pictures. The robot was the DragonBot, a fluffy, “squash and stretch” robot designed to appeal to children (Freed, 2012; Kory et al., 2013; Setapen, 2012). The robot is shown in Figure 1. An android phone runs the robot’s control software while also providing a screen displaying the robot’s animated face. The face can display a variety of facial expressions with some expressions involving both the face and physical movements of the body. The robot’s gaze behavior involves movements of both the eyes on the screen and rotations of the head and body in space. The robot did not have a separate head joint that could rotate, so it moved both the head and body at once. The robot, which wore blue fur, was named “Blue” and was referred to in a non-gendered way by the experimenters throughout the study. Because the robot moves on its own, its appearance is unlike that of a puppet; it is not obviously controlled by or connected to a human.

The female adult recorded the robot’s speech so that the speech in the study would be the same in both conditions. These utterances were pitch-shifted to make them sound more child-like, which was consistent with the persona of the robot as a child-like peer, as well as making them reasonably distinct from the adult’s voice, so the child would not be confused. Note that although the robot was presented as a child-like peer, it was compared to an adult human for two primary reasons. First, practically speaking, it would more difficult to train a child’s human peer than an adult human to provide the target vocabulary words in accordance with the study procedure. Second, adult humans are frequent providers of novel vocabulary words for young children, arguably more often than their peers, and thus we wanted to compare the robot to this realistic scenario.

2.4 Teleoperation

We used a custom tele-operation interface to remote-operate the robot. This allowed the robot to appear autonomous to participants while removing several technical barriers, such as natural language processing and automation of the timing of speech. Furthermore, although the robot was remote-operated by a human**,** our goal was to understand what children's interactions with a robot might be like if an autonomous version of the robot were to exist. From this kind of study, we can gain some understanding of how children might interact with an "ideal" robot, which will indicate how we should design and build such robotic learning companions in the future.

The teleoperator triggered speech acts, movements, and facial expressions according to a script. The same speech and facial expressions were used in the adult condition. The robot's gaze was automated, such that the robot would glance down at the tablet and back up at the child at approximately the same times as the adult did in the adult condition. Given the simplicity of the study scenario, the teleoperator primarily needed to pay attention to timing in order to trigger the speech actions at the right times relative to when the primary experimenter spoke (e.g., when introducing the robot), and when images appeared on the tablet. The same script was used in the adult condition, although there were necessarily some differences in, for example, facial expressions.

The second experimenter was the teleoperator for all participants (as well as taking on the role of the adult in the adult condition). Thus, she was not blind to the experimental conditions or hypotheses. She was instructed to make the robot's interaction and her own actions as similar as possible. She had a great deal of prior experience in controlling the robot, having been trained to do so in multiple prior studies.

2.5 Data

Children’s responses to the animal recall tests, the pre-test questionnaires, and the post-test questionnaires were recorded on paper during the experiment and later transferred to a spreadsheet. Video and audio recordings of each session were made using a camera set on a tripod behind the robot and adult, facing the child. Three raters were trained to code the videos for children's gaze patterns—i.e., they judged whether children were looking at their interlocutor (person or robot), the pictures, or elsewhere. The reliability sample consisted of two of the videos. Each rater also independently coded one third of the remaining videos, for a total of 24 videos; video data were missing or incomplete for several children (e.g., the child’s eyes were not visible for the duration of the session). Cohen's kappa was used to determine the level of agreement between the raters from the reliability sample. There was high agreement between the raters' judgments regarding gaze direction, k = 0.974, k_max = 0.995.

3 Results

3.1 Recall

Figure 2 shows the mean number (maximum = 6) of correct animal responses, near miss responses (i.e., the animal that had appeared together with the named animal on learning trials), and incorrect animal responses as a function of condition. Overall, children produced a mean of 2.58 correct animal responses (43.0% correct, *SD* = 1.58). They chose the near miss animal for a mean of 1.70 (28.3% of the time, *SD* = 1.10), and an incorrect animal for a mean of 1.61 (26.8%, *SD* = 1.20).

A two-way ANOVA with interlocutor (person vs. robot) and condition (close vs. far) as factors and age as a covariate revealed that children in the far condition chose correct animals more often than children in the close condition, $F(1,57) = 6.37$, $p = .014$ (see Figure 2). There was no significant effect of interlocutor, $F(1,57) = .073$, $p =.789$; nor was there a significant interaction of condition with interlocutor, $F(1,57) = .282$, $p =.598$. Age was not significant as a covariate. The eta-squared value ($\eta_p^2 = .101$) indicates a medium-small effect. Given the main effect of condition, we examined children's pattern of responding in each condition more closely.

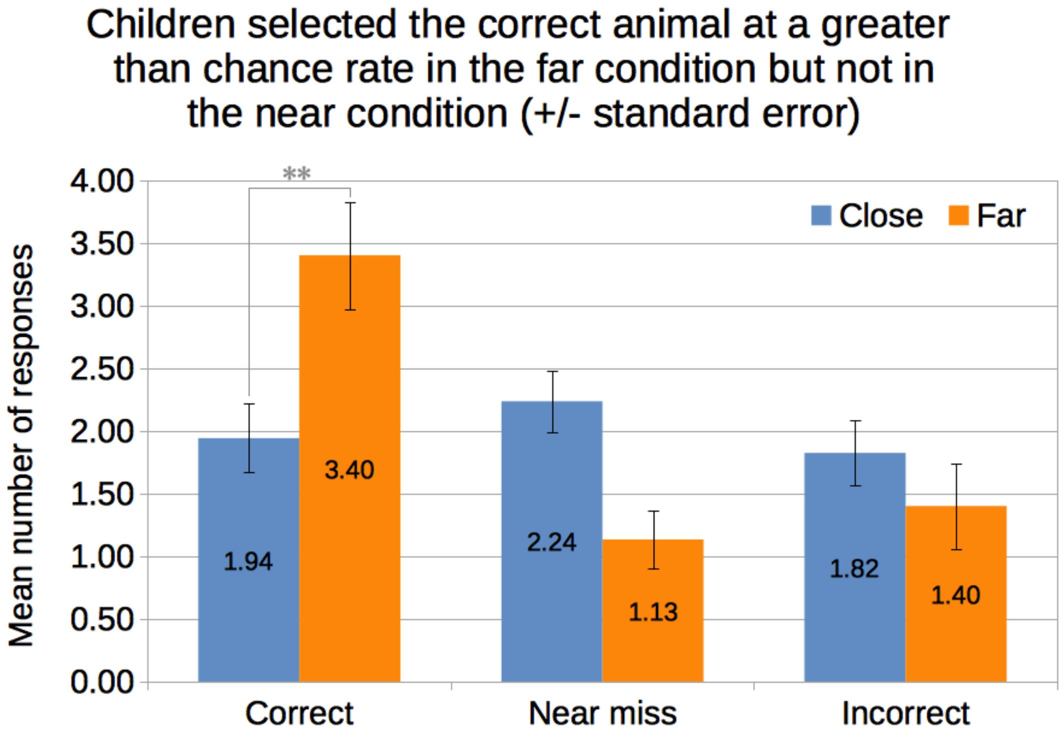

Figure 2. Mean number of correct animal responses, near miss responses, and incorrect responses as a function of condition (close versus far).

Inspection of the findings shown in Figure 2 reveals that children's responding displayed little departure from chance (1.5) with one notable exception. In the far condition, children selected the correct animal a mean 3.4 of 6 times (56.7% of the time, *SD* = 1.65). This was significantly greater than chance (1.50) and had a large effect size, $t(14)$ =4.38, $p < .001$, d = 1.13. Moreover, further analysis confirmed that children produced a similar rate of correct animal responses for both interlocutors. With the robot, children selected the correct animal a mean of 1.76 of 3 times (58.7% of the time, SD = 1.11). This was significantly greater than chance (.75) and had a large effect size, $t(14) = 3.52$, $p = .0034$, $d = .909$. Similarly, with the person, children selected the correct animal a mean of 1.73 of 3 times (57.6% of the time, SD = 1.03). This was also significantly greater than chance (.75), $t(14) = 3.69$, $p = .0024$, $d = .952$ and had a large effect size.

In examining the close condition, we found that children selected one of the animals present during naming (i.e., either the correct animal or the animal shown in the same pair) at a greater than chance rate (.50), $t(16) = 2.37$, $p = .031$, $d = .575$, even if, as noted above, selection of either the correct or near miss animal was no better than chance when considered separately.

3.2 Gaze

A two-way ANOVA with interlocutor (person vs. robot) and condition (close vs. far) as factors and age as a covariate revealed that children spent more time gazing at the robot (M = 46.9% of the time, SD = 13.2%) than at the person (M = 13.9% of the time, SD = 7.92%), $F(1,43) = 106.5$, $p < .001$, $\eta_p^2 = 703$ (see Figure 3). A second two-way ANOVA of interlocutor X condition confirmed that children also spent more time gazing at the pictures shown on the tablets when their interlocutor was a person (M = 73.6% of the time, SD = 18.3%) than when their interlocutor was the robot (M = 46.9% of the time, SD = 15.1%), $F(1,43) = 34.2$, $p < .001$, $\eta_p^2 = .773$. In the case of both ANOVAS, there was no significant effect of condition, no significant interaction, and age was not a significant covariate. However, age significantly affected how much time children spent looking elsewhere (e.g., at the experimenter), $F(1,43)$ =

9.78, $p = .003$, $\eta_p^{\ 2} = .170$. Younger children looked away more often than older children.

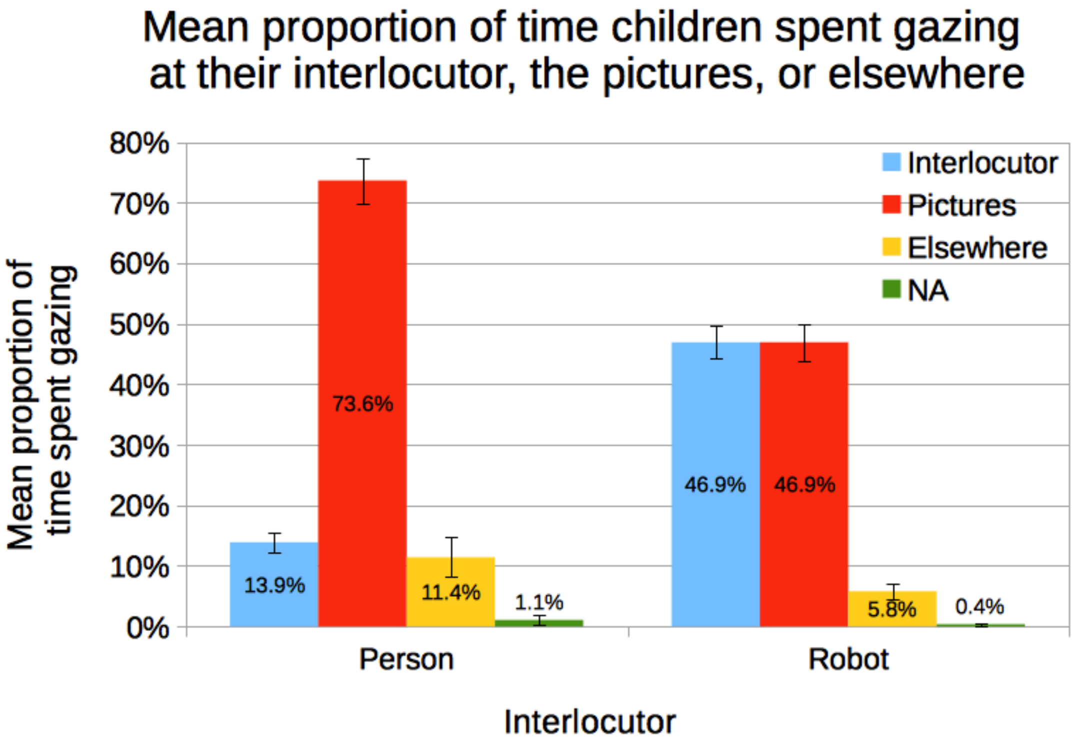

Figure 3. Mean proportion of the total time that children spent gazing at their interlocutor, the pictures, or elsewhere (+/- standard error). NA indicates the portion of time that the child's gaze could not be coded.

4 Discussion

We asked whether preschool children ranging from 2-5 years would treat the robot like a human interlocutor by attending to its social cues, specifically its gaze direction and bodily orientation, when learning the referent of a new word. The pattern of results that was obtained underlines important parallels between children's learning from a robot as compared to a human interlocutor. Moreover, the pattern of results was stable across the age range tested. Two findings are especially noteworthy.

First, when the paired animal pictures were presented close together on a single tablet, children’s subsequent identification of the named animal was no better than chance. The most plausible interpretation of this failure is that when the two pictures were presented close together, the non-verbal cues of the interlocutor were not sufficiently distinct for children to determine which animal was being referred to – whether by the robot or by the adult. This result is supported by recent research by Yu and Smith (2013). Whilst children can discriminate where someone is looking when given strong spatial cues—such as when gaze is supplemented by head and body orientation or by pointing gestures—gaze direction alone is not spatially precise enough for children to be able to discriminate between objects that are very close together.

The plausibility of this interpretation is strengthened by the second notable finding: When the paired animals were presented further apart on separate tablets, children’s subsequent identification of the named animal was considerably better than chance for both interlocutors. Children’s successful learning in the far condition renders it extremely unlikely that their failure to learn in the close condition can be attributed to incidental factors such as distractibility, or difficulty in retaining the names from the naming phase to the subsequent test phase. Such demands were also present in the far condition but children succeeded in learning the new names. Furthermore, results of the gaze analysis showed no differences in how long children looked at their interlocutor or at the animal pictures in the close versus far conditions, suggesting that children paid similar amounts of attention to the task in both conditions. Reciprocally, it is also plausible to conclude that the opportunity to determine the intended referent on the basis of non-verbal cues provided by the interlocutor was a key component of children’s learning. Had children not needed to decode and use such cues in order to learn the new animal names, it would have been reasonable to expect them to learn the names equally well in both conditions –

but this was not the case.

Further support for this interpretation can be obtained by looking more closely at the pattern of responding in the close condition. Although children's recall score for the correct animal was no greater than chance, they did choose either the correct or near miss animal at a greater than chance rate. This implies that, in the close condition, children paid attention to the name and to both of the animals but could not determine which was being named given the subtlety of their interlocutor's non-verbal cues.

Indeed, past work has shown that children will learn new words from various media (e.g., Kory, 2014; Movellan et al., 2009; Naigles & Mayeux, 2001; Tanaka & Matsuzoe, 2012; Willoughby et al., 2015), but these prior studies did not directly study how children learn new words from a person as compared to a robot. Nor did they investigate the extent to which children can and do use the gaze direction and orientation of a robot to determine what is being named. O'Connell et al. (2009) studied the extent to which 18-month-old infants would learn new words from a mechanical robot, finding little evidence for word learning. Our study differed in two key ways and either or both may have led to children's success. First, our robot was presented as a social actor through a brief three-way conversation, as described earlier. By contrast, the robot in O'Connell et al.'s (2009) study was merely observed by the infants; they did not interact with it directly. This factor of social interaction may be important for learning. Second, the children in O'Connell et al.'s (2009) study were significantly younger than the children in our study. Indeed, past research has show that word-learning that is guided by speaker orientation cues is relatively fragile among infants (Horst & Samuelson, 2008).

Several limitations to the current findings should be noted. First, the analysis of children's gaze revealed that children spent more time looking at the robot than at the person.

One hypothesis is that this is due to the novelty of the robot. However, in one three-session study where children aged 10-13 years were tutored by a social robot in a map reading task, there was no decrease in the amount of time children spent looking at the robot over the three sessions (Serholt et al., 2016). This may indicate that over time, children's gaze behavior with the robot may not change, or alternatively, that three sessions was not long enough for the novelty of the robot to wear off. Additional work by Serholt et al. (2014) found that during a LEGO construction task, children aged 11-13 years spent more time looking at the task when with a human tutor than when with a robot tutor. Our results reflect the same pattern: children spent more time looking at the animal pictures when with the human. Several other studies have found that children will gaze more at a robot that displays greater social contingency than at an asocial robot (Breazeal et al., 2016; Kennedy et al., 2015). In this regard, despite our attempts to control for social behavior in our study procedure, the human still likely displayed more social contingency than the robot, which could have led to our similar result (i.e., that children looked more at the more social agent). Nevertheless, despite their tendency to look at the robot, children still spent nearly half their time looking at the animal pictures when with the robot, and they subsequently performed just as well on the recall test, whether they had learned the names from a person or a robot.

A second limitation is that because the robot did not have a joint enabling the head to turn independent of the body, it turned its whole body slightly toward the picture**.** It is possible that this bodily cue increased children's ability to detect which picture the robot was naming. As robot technology becomes more fine-grained in its simulation of human movements, it will be possible to make a more detailed analysis of the role of particular orientation cues.

Third, there were other differences between the robot and human that we did not control that could have affected the results. For example, it is difficult to control for all the human's nonverbal behaviors. This, in combination with the robot's limited repertoire of facial expressions and gaze behavior, likely led to some differences in the nonverbal expressions used by the human and by the robot, such as the human being perceived as more social or responsive. We also specifically did not assign the robot a particular gender. However, prior work has found that most preschool children will assign gender to a robot, often matching their own, even if is referred to in a non-gendered way by experimenters (Kory, 2014). This may be because preschool children often show affiliation toward same-sex peers (see, e.g., La Freniere et al., 1984). Such affiliation could have affected their behavior, perhaps especially if the gender they assigned to the robot was different from that of the human (female) interlocutor. In addition, the voice we used for the robot was child-like rather than an adult voice. Although our goals in shifting the pitch of the robot's voice were to keep the voice in line with the robot's persona and keep it distinct from the female interlocutor's while keeping all other auditory features constant, it could be that the difference in voices led children to view the two interlocutors differently, e.g., perhaps the adult was seen as more of an expert or teacher than the robot. Again, however, despite these differences, children did perform just as well on the recall test, whether they had learned the names from a person or a robot.

However, despite the parallels in children's recall, it is important to note that our posttest of children's learning was quite restricted. We only assessed children's ability to re-identify the referent that had been paired with the name when it had been presented earlier. We did not assess whether children could extend the name to other referents falling into the same category or could produce the name themselves when shown an appropriate referent. It is conceivable that differences between learning from a robot and human would emerge on these, potentially more demanding tests, of word learning.

5 Conclusion

Given the potential of social robots as tutors and learning companions for children, it is important to explore the mechanisms by which children learn from robots, and how learning with this kind of social technology compares to learning with a human partner. In this study, we examined whether children will attend to the same social cues—specifically gaze direction and bodily orientation—from a robot as from a human partner during a word-learning task. Our results confirm that children are able to acquire new vocabulary in a spontaneous and natural fashion from a robot using gaze direction and bodily orientation as a cue to linguistic reference. In this respect, their learning from a human and from a robot was comparable.

Acknowledgments

This research was supported by the National Science Foundation (NSF) under Grant 122886 and Graduate Research Fellowship Grant No. 1122374. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not represent the views of the NSF.

References

Baldwin, D.A., 1993. Infants' ability to consult the speaker for clues to word reference. J. Child Lang. 20, 395–418. doi:10.1017/S0305000900008345

Baldwin, D.A., 1991. Infants' Contribution to the Achievement of Joint Reference. Child Dev. 62, 874–890. doi:10.1111/j.1467-8624.1991.tb01577.x

Belpaeme, T., Baxter, P.E., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., Racioppa, S., Kruijff-Korbayová, I., Athanasopoulos, G., Enescu, V., Looije, R., Neerincx, M., Demiris, Y., Ros-Espinoza, R., Beck, A., Cañamero, L., Hiolle, A., Lewis, M., Baroni, I., Nalin, M., Cosi, P., Paci, G., Tesser, F., Sommavilla, G., Humbert, R., 2012. Multimodal Child-Robot Interaction: Building Social Bonds. J. Hum.-Robot Interact. 1, 33–53.

Breazeal, C., Harris, P.L., DeSteno, D., Kory Westlund, J.M., Dickens, L., Jeong, S., 2016. Young Children Treat Robots as Informants. Top. Cogn. Sci. 1–11. doi:10.1111/tops.12192

La Freniere, P., Strayer, F.F., Gauthier, R., 1984. The Emergence of Same-Sex Affiliative Preferences among Preschool Peers: A Developmental/Ethological Perspective. Child Development 55, 1958–1965. doi:10.2307/1129942.

Freed, N.A., 2012. "This is the fluffy robot that only speaks French": language use between preschoolers, their families, and a social robot while sharing virtual toys (Master's Thesis). Massachusetts Institute of Technology, Cambridge, MA.

Harris, P.L., Lane, J.D., 2014. Infants Understand How Testimony Works. Topoi 33, 443–458. doi:10.1007/s11245-013-9180-0

Horst, J.S., Samuelson, L.K., 2008. Fast Mapping but Poor Retention by 24-Month-Old Infants. Infancy 13, 128–157. doi:10.1080/15250000701795598

Kahn, P.H., Kanda, T., Ishiguro, H., Freier, N.G., Severson, R.L., Gill, B.T., Ruckert, J.H., Shen, S., 2012. “Robovie, you’ll have to go into the closet now”: Children’s social and moral relationships with a humanoid robot. Dev. Psychol. 48, 303.

Kennedy, J., Baxter, P., Senft, E., Belpaeme, T., 2016. Heart vs hard drive: Children learn more from a human tutor than a social robot, in: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Presented at the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 451–452. doi:10.1109/HRI.2016.7451801

Kennedy, J., Baxter, P., Belpaeme, T., 2015. The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning, in: Proc. HRI.

Kory, J., 2014. Storytelling with robots: Effects of robot language level on children’s language learning (Master’s Thesis). Massachusetts Institute of Technology, Cambridge, MA.

Kory Westlund, J., Breazeal, C., 2015. The Interplay of Robot Language Level with Children’s Language Learning During Storytelling, in: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts, HRI’15 Extended Abstracts. ACM, New York, NY, USA, pp. 65–66. doi:10.1145/2701973.2701989

Kory, J.M., Jeong, S., Breazeal, C.L., 2013. Robotic learning companions for early language development, in: In J. Epps, F. Chen, S. Oviatt, & K. Mase (Eds.), Proceedings of the 15th ACM on International Conference on Multimodal Interaction. ACM, New York, NY: ACM, pp. 71–72.

Kory Westlund, J., Dickens, L., Sooyeon Jeong, Paul Harris, David DeSteno, Cynthia Breazeal, 2015. A comparison of children learning new words from robots, tablets, and people, in:

New Friends: The 1st International Conference on Social Robots in Therapy and Education. Almere, The Netherlands.

Kuhl, P.K., 2007. Is speech learning "gated" by the social brain? Dev. Sci. 10, 110–120.

Kuhl, P.K., 2011. Social Mechanisms in Early Language Acquisition: Understanding Integrated Brain Systems Supporting Language, in: Jean Decety, John T. Cacioppo (Eds.), The Oxford Handbook of Social Neuroscience.

Markson, L., Bloom, P., 1997. Evidence against a dedicated system for word learning in children. Nature 385, 813–815. doi:10.1038/385813a0

Meltzoff, A.N., Brooks, R., Shon, A.P., Rao, R.P., 2010. Social robots are psychological agents for infants: A test of gaze following. Neural Netw. 23, 966–972.

Moses, L.J., Baldwin, D.A., Rosicky, J.G., Tidball, G., 2001. Evidence for Referential Understanding in the Emotions Domain at Twelve and Eighteen Months. Child Dev. 72, 718–735. doi:10.1111/1467-8624.00311

Movellan, J., Eckhardt, M., Virnes, M., Rodriguez, A., 2009. Sociable robot improves toddler vocabulary skills, in: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction. ACM, pp. 307–308.

Naigles, L.R., Mayeux, L., 2001. Television as incidental language teacher. Handb. Child. Media 135–152.

O'Connell, L., Poulin-Dubois, D., Demke, T., Guay, A., 2009. Can Infants Use a Nonhuman Agent's Gaze Direction to Establish Word–Object Relations? Infancy 14, 414–438. doi:10.1080/15250000902994073

Phillips, A.T., Wellman, H.M., Spelke, E.S., 2002. Infants' ability to connect gaze and emotional expression to intentional action. Cognition 85, 53–78. doi:10.1016/S0010-0277(02)00073-2

Samuelson, L.K., Smith, L.B., Perry, L.K., Spencer, J.P., 2011. Grounding Word Learning in Space. PLOS ONE 6, e28095. doi:10.1371/journal.pone.0028095

Serholt, S., Basedow, C.A., Barendregt, W., Obaid, M., 2014. Comparing a humanoid tutor to a human tutor delivering an instructional task to children, in: 2014 IEEE-RAS International Conference on Humanoid Robots. Presented at the 2014 IEEE-RAS International Conference on Humanoid Robots, pp. 1134–1141. doi:10.1109/HUMANOIDS.2014.7041511

Serholt, S., Barendregt, W., 2016. Robots Tutoring Children: Longitudinal Evaluation of Social Engagement in Child-Robot Interaction, in: Proceedings of the 9th Nordic Conference on Human-Computer Interaction, NordiCHI '16. ACM, New York, NY, USA, p. 64:1–64:10. doi:10.1145/2971485.2971536

Setapen, A.M., 2012. Creating robotic characters for long-term interaction (Master's Thesis). MIT, Cambridge, MA.

Shiomi, M., Kanda, T., Howley, I., Hayashi, K., Hagita, N., 2015. Can a Social Robot Stimulate Science Curiosity in Classrooms? Int. J. Soc. Robot. 1–12. doi:10.1007/s12369-015-0303-1

Tanaka, F., Matsuzoe, S., 2012. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. J. Hum.-Robot Interact. 1, 78–95.

Vouloutsi, V., Blancas, M., Zucca, R., Omedas, P., Reidsma, D., Davison, D., Charisi, V., Wijnen, F., Meij, J. van der, Evers, V., Cameron, D., Fernando, S., Moore, R., Prescott, T., Mazzei, D., Pieroni, M., Cominelli, L., Garofalo, R., Rossi, D.D., Verschure, P.F.M.J., 2016. Towards a Synthetic Tutor Assistant: The EASEL Project and its Architecture, in: Biomimetic and Biohybrid Systems. Presented at the Conference on Biomimetic and Biohybrid Systems, Springer, Cham, pp. 353–364. doi:10.1007/978-3-319-42417-0_32

Willoughby, D., Evans, M.A., Nowak, S., 2015. Do ABC eBooks boost engagement and learning in preschoolers? An experimental study comparing eBooks with paper ABC and storybook controls. Comput. Educ. 82, 107–117. doi:10.1016/j.compedu.2014.11.008

Yu, C., Smith, L.B., 2013. Joint Attention without Gaze Following: Human Infants and Their Parents Coordinate Visual Attention to Objects through Eye-Hand Coordination. PLOS ONE 8, e79659. doi:10.1371/journal.pone.0079659