Where do people look when tutoring a robot?

Katrin Solveig Lohan
Instituto Italiano di Tecnologia
iCub Facility
Italy, Genova 16163
katrin.lohan@iit.it

Kerstin Fischer
University of Southern Denmark
IDK Alsion 2
DK-6400 Sonderborg
0045-6550-1220
kerstin@sdu.dk

Christian Dondrup
University of Bielefeld
Applied Informatics
Germany, Bielefeld 33615
cdondrup@techfak.uni-bielefeld.de

Chrystopher Nehaniv
University of Hertfordshire
Adaptive Systems
United Kingdom,
AL10 9AB, Hatfield
C.L.Nehaniv@herts.ac.uk

I. Introduction

In this paper, we investigate the relationship between tutors' gaze behavior and particular kinds of linguistic behaviors. In particular, we describe how word classes are distributed over different gazing classes. For this, we collected data from human-robot interactions and used a classification of our participants' gazing behavior to create subsets of their utterances. The participants' speech was transcribed for 3 sessions (2 min each) of interaction with the robot and classified based on the detected gazing classes (looking at the robot, looking at the object or looking somewhere else). The analysis shows that there are, for instance, more object related keywords when people are gazing at an object, and more personal pronouns when people are looking at the robot. Understanding the relationship between human tutors' linguistic and gazing behavior can facilitate bootstrapping the one capability from the other, such that gazing behavior is exploited for narrowing down the search space for object-relevant linguistic material.

II. Human Robot Interaction Study

In the study, human tutors (naive users) were asked to present three differently sized cubes and to explain the colors and shapes of the markers on these cubes to the iCub robot [1]. The participants were asked to come three times to describe to Deechee (the name of the iCub robot), the same cubes for two minutes in total. Between the sessions, there was a pause of at least one day. During this pause, the spoken utterances of the participants were transcribed. The study took place at the University of Hertfordshire. All participants were fluent English speakers. There were altogether eight participants, four in each group, each one participating three times for two minutes.

A. Design

Data were elicited in two conditions, representing two different robot feedback strategies. In one condition, the robot was reacting based on the tutoring spotter (see section III). In the other condition, the robot's movement was controlled by the non-tutoring spotter. In both conditions the robot was capable of looking at the face of the interaction partner, the objects presented to it or elsewhere in the room. Furthermore, the robot had the ability to point towards an object. A learning algorithm related the object-detection input with the spoken utterances and resulted in the robot expressing words which were salient and information-wise more relevant based on its sensorimotor perceptions (for a detailed description of this work see [2]). These words were uttered when this object reappeared in the following session. Thus, there was no verbal feedback in the first session, but only in the second and third.

B. Setup

For the technical realization of the tutoring spotter, we equipped the robot with a kinect sensor. The participants were asked to wear a headset, and a face and object tracking module used a separate webcam. In the pre-phase of the experiment, the participants were initializing the kinect and face tracking. Then they were seated across the robot in front of a table.

III. Implementation

A. Tutoring spotter

The tutoring spotter system was created in the ITALK project and is capable of giving contingent feedback in eye gaze and looming/pointing [3], [4]. To spot a tutor enables the system to pay attention to an ostensive action and the crucial parts or circumstances that are helpful in resolving the question of what and when to imitate [5]. Furthermore, mechanisms that detect (and produce) contingency can be a precursor for later dialogical competencies as described in the framework of grounding.

1) Gazing feedback: The robot detects three gazing classes (gaze at the robot, gaze away and gaze at an object) and responds with the same behavior. Thus, the robot would look at you when you are looking at it, the robot would look around when you are looking elsewhere and when you are looking at the object, the robot would follow your gaze.

2) Pointing feedback: Since child-directed action was shown to be important in a tutoring situation, Matatyaho and Gogate [6] further investigated the kind of action that is typically applied. They found that the looming action, which is an action that describes a movement of a tutor moving an object towards a learner's face, is used more frequently than upward or backward motions in temporal synchrony with the spoken words. This looming motion is likely to highlight a novel word-object relation [7].

In the tutoring spotter, we implemented a detector for this kind of movement by facilitating the distance. The iCub robot would respond to a looming behavior by pointing towards the object that the tutor is presenting.

B. Non-tutoring spotter

For the non-tutoring spotter setup we used an implementation based on tracking the objects, the face of the participant to control the iCub's behavior, as in the tutoring spotter condition the robot was capable to look at the participant's face, the object or somewhere else and it was using pointing or non pointing gestures.

1) Gazing feedback: The non-tutoring spotter saturates a 'boredom' filter if the same face or object is being seen for too long and the robot switches to random gaze. That means the robot is changing the gazing behavior based on timing.

2) Pointing feedback: In the non-tutoring spotter condition, the robot tracked the object and occasionally (on a random basis) pointed at the object. The robot made no use of the contingent feedback of the tutor in this condition.

C. Language acquisition system

The system uses prosodic salience and shared belief to bootstrap word learning via interactions with the iCub robot. We automatically analyzed intonational contours from the transcribed sessions in order to extract words which are both prosodically salient and which may have been chosen by the human to express particular pragmatic relationships. The association of these words with the sensorimotor experiences of the robot allows the robot to derive rudimentary meaning and to bootstrap the word learning process.

IV. Language analysis

As our language acquisition system needs phonemically transcribed input, we manually annotated the videos between the sessions. Furthermore, we used the automatic detection of the gazing classes to divide the spoken utterances into three subsets for each participant. In a next step, we counted the words for each class. This was followed by a search in each subset for specific keywords that describe the objects: 'blue', 'red' and 'green', 'arrow', 'cross', 'circle', 'heart', 'moon' and 'star', 'small', 'medium' and 'large' [8]. Finally, we analyzed how often and where concerning the gazing classes the participants used personal pronoun in the interaction. We searched for the following personal pronouns: 'we', 'I', 'you', 'us' and 'Deechee'.

V. Results

We used a standard t-test to validate the results and we found significantly higher numbers of keywords over both conditions between the last two sessions when participants were looking at the face of the robot (session2: M= 0.114, session3: M= 1.313, SD = 1.21, p = 0.0263). We found significant more use of keywords and uttered words in total towards the robot in the tutoring spotter condition when participants were looking at an object (keywords: tutoring spotter condition: M= 23.375, non-tutoring spotter condition: M= 35.375, SD = 13.69, p = 0.0250), (all words uttered: tutoring spotter condition: M= 171.75, non-tutoring spotter condition: M= 264.25, SD = 111.32, p = 0.0283). In addition, we observed an increase from session 1 to session 3 in the used words while looking at the object in the tutoring spotter condition (tutoring spotter condition: session1: M= 211, session3: M= 132.5, SD = 28.7692, p = 0.0121).

Also over all sessions, we found significantly more keywords while participants were looking at an object than while they were looking at the robot or elsewhere for the tutoring spotter condition (keywords: tutoring spotter condition: object: M= 22.8334, elsewhere: M= 1.9167, SD= 7.9825,p=0.000), (keywords: tutoring spotter condition: object: M= 22.8334, robot: M=1.0833, SD= 8.8374,p=0.000). We found the same effect over all sessions for the non-tutoring spotter condition; in particular, we found significantly more keywords while tutors were looking at the object then while looking at the robot or elsewhere (keywords: non-tutoring spotter condition: object: M= 33.0833, elsewhere: M= 4.3334, SD= 7.9825, p=0.000), (keywords: non-tutoring spotter condition: object: M= 33.0833, robot: M=1.25, SD= 8.8374, p=0.000). Finally we found in the non-tutoring spotter condition significantly more use of personal pronouns while tutors were looking at the object than when they were looking at the robot (personal pronouns: non-tutoring spotter condition: object: M= 0.0924, robot: M=0, SD= 0.021, p=0.0162). In the tutoring spotter condition, we found no significant difference between all sessions for the use of personal pronouns.

VI. Conclusion

To summarize, we found that the participants used more object keywords while looking at the object. We found that the participants in the tutoring spotter condition used more keywords in the 3rd session than in the first one. In all participants we found more use of the keywords while they were looking at the robot during the 3rd than during the 2nd session. This is in line with previous finding that suggest that tutors adapt closely to the robot's linguistic development [9]. These results can be exploited to extract object-relevant linguistic forms for automatic language learning. Concerning the use of personal pronouns, we found that in the non-tutoring spotter condition the participants used them more often while looking at the object than while looking at the robot, but there were no clear results concerning the use of personal pronouns in the tutoring spotter condition. We suggest that this result is due to the fact that mutual gaze provides participants with enough common ground to render further, more explicit elaboration of the interpersonal relationship superfluous.

Acknowledgment

Kerstin Fischer, Chrystopher Nehaniv and Katrin Solveig Lohan gratefully acknowledge the financial support from the FP7 European Project ITALK which supported the data collection. Katrin Solveig Lohan gratefully acknowledge the financial support from the EU FP7 Marie Curie ITN RobotDoc. Contract No. 235065. Finally we would like to thank Hagen Lehmann, Joe Saunders, Britta Wrede and Katharina Rohlfing.

References

[1] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. Von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor *et al.*, "The icub humanoid robot: An open-systems platform for research in cognitive development," *Neural Networks*, vol. 23, no. 8, pp. 1125–1134, 2010.

[2] J. Saunders, H. Lehmann, Y. Sato, and C. L. Nehaniv, "Towards using prosody to scaffold lexical meaning in robots," in *Proceedings of ICDL-EpiRob 2011*. IEEE, 2011.

[3] K. Lohan, K. Rohlfing, K. Pitsch, J. Saunders, H. Lehmann, C. Nehaniv, K. Fischer, and B. Wrede, "Tutor spotter: Proposing a feature set and evaluating system," *International Journal of Social Robotics*, 2012.

[4] K. Lohan, K. Pitsch, K. Rohlfing, K. Fischer, J. Saunders, H. Lehmann, C. Nehaniv, and B. Wrede, "Contingency allows the robot to spot the tutor and to learn from interaction," in *Development and Learning (ICDL), 2011 IEEE International Conference on*, vol. 2. IEEE, 2011, pp. 1–8.

[5] C. Nehaniv and K. Dautenhahn, "Like me?-measures of correspondence and imitation," *Cybernetics and Systems*, vol. 32, no. 1, pp. 11–51, 2001.

[6] D. Matatyaho and L. Gogate, "Type of maternal object motion during synchronous naming predicts preverbal infants' learning of word–object relations," *Infancy*, vol. 13, no. 2, pp. 172–184, 2008.

[7] L. Gogate, L. Bolzani, and E. Betancourt, "Attention to maternal multimodal naming by 6-to 8-month-old infants and learning of word–object relations," *Infancy*, vol. 9, no. 3, pp. 259–288, 2006.

[8] C. Dondrup, K. Lohan, J. Saunders, H. Lehmann, C. Nehaniv, and B. Wrede, "Keyword detection in human-robot tutoring scenarios."

[9] K. Fischer and J. Saunders, "Getting acquainted with a developing robot," in *Proceedings of 'Human User Behavior', IROS, Portugal*. Springer, 2012.