Learning Word Groundings from Humans Facilitated by Robot Emotional Displays

David McNeill
Boise State University
1910 W. University Dr.,
Boise, ID 83725
davidmcneill@
u.boisestate.edu

Casey Kennington
Boise State University
1910 W. University Dr.,
Boise, ID 83725
caseykennington@
boisestate.edu

Abstract

In working towards accomplishing a human-level acquisition and understanding of language, a robot must meet two requirements: the ability to learn words from interactions with its physical environment, and the ability to learn language from people in settings for language use, such as spoken dialogue. In a live interactive study, we test the hypothesis that emotional displays are a viable solution to the cold-start problem of how to communicate without relying on language the robot does not–indeed, cannot–yet know. We explain our modular system that can autonomously learn word groundings through interaction and show through a user study with 21 participants that emotional displays improve the quantity and quality of the inputs provided to the robot.

1 Introduction

In any first language acquisition task, three questions must be resolved:

1. *What kinds of words to be learned?*
2. *How to model those words' semantics?*
3. *How to overcome the cold-start problem*?

To answer the first question, we note that co-located spoken dialogue interaction is the fundamental setting of first language acquisition for humans (Fillmore, 1981; McCune, 2008) and that children generally tend to focus on physical objects first, as evidenced by age-of-acquisition datasets. For this reason, concrete words that denote physical objects are learned earlier than abstract words (Kuperman et al., 2013). This informs the answer to the second question: the model of semantics should be able to connect language with the physical world, which is part of the goal of *grounded semantics* (e.g., grounding a color word like *green* with visual information).

This still leaves the third question: how can a system learn word groundings in a physical, co-located setting without using words it has yet to learn? In answering this, there is evidence that having a physical body is a requirement for bootstrapping semantic learning of concrete word denotations (Smith and Gasser, 2005; Johnson, 2008). Therefore, a system that can use extra-linguistic cues through physical signals can potentially overcome the cold-start problem and learn words without uttering words it has never heard.

In this paper, we test the hypothesis that *emotional displays*, specifically *confusion* and *understanding* displays performed by an embodied robot, are a viable solution to the cold-start problem. Our reasons are two-fold: emotional displays can relate the robot's state to its human teacher, and emotional displays are developmentally appropriate for the most common language acquisition setting (i.e., an adult teaching a child) (Adolphs, 2002), and would therefore not lead a human user to make incorrect assumptions regarding the robot's level of comprehension.

In an interactive study with 21 participants, our robot independently and autonomously explored a physical setting and elicited relevant word references and feedback from the participants, who were tested both with a robot that displayed emotions and a robot that did not. For grounded semantics, we opted for a model that is incremental (i.e., operates at the word level), that can map individual words to physical features, and that can learn a mapping between a word and physical features using only a few examples–the *words-as-classifiers model* (WAC) (Kennington and Schlangen, 2015). In the WAC model, each word is represented by its own classifier trained on "not / is" examples of real-world referents. The WAC model has been used in interactive dialogue scenarios with robots before (Hough and Schlangen, 2017). Importantly,

Proceedings of the SIGdial 2020 Conference, pages 97–106
1st virtual meeting, 01-03 July 2020.

our system not only learned word groundings as it interacted with participants, it also incorporated a reinforcement learning model to learn from positive or negative participant feedback which emotional valence (either understanding or confusion) to display. Analyzing the results from the surveys and the learned WAC classifiers, we discovered that the use of emotional displays improved the quantity and quality of the inputs provided to the robot, with the effect modulated by the valence and frequency of the emotional displays.

2 Background & Related Work

It has been shown that people assign anthropomorphic characteristics, social roles and models when interacting with robots (Kiesler and Goetz, 2002), which has implications for the kinds of settings and tasks that robots can carry out with human collaborators. One dimension that people anthropomorphically assign to robots is emotion. We cannot prevent users from making emotional judgements of a robot's behavior (Novikova et al., 2015). Instead, if a robot's behavior were designed to take these emotional judgements into account, the robot could be made more predictable and more interpretable by humans in a complex environment (Breazeal, 2005). Indeed, emotional features can make a robot appear more lifelike and believable to humans, thereby making humans more prone to accept and engage with them (Cañamero, 2005). Of course, the choice of emotions must be taken with care; Claret et al. (2017) showed that happiness and sadness emotional displays during primary tasks (e.g., such as transporting an object) could confuse human interlocutors as robot actions (e.g., jerkiness, activity, gaze), and robot movement are also emotionally interpreted.

Similar to *conversational grounding*, Jung (2017) explained how *affective grounding*–the coordination on content and process of affect–occurs between robots and human users. We handle this particular phenomenon by only considering a positive and negative valence of a single affective type (i.e., confusion vs. understanding), and by establishing through an evaluation that they are indeed interpreted the way we expect before we use them in a language learning task.

Robots have been used in many language grounding tasks; Matuszek (2018) gives an overview of the recent literature. In some cases the cold-start problem is handled by Wizard-of-Oz paradigm studies where a robot that knows no word denotations interacts with human participants, but the robot is in fact being controlled by a confederate. In this paper, our robot is fully autonomous and has no pre-programmed language production capabilities; that is, the robot will never utter words it hasn't encountered within an interaction.

Beyond word learning, our approach attempts to ground language and learn which emotions to display. This work builds on Ferreira and Lefèvre (2015) which outlined the approach we take for a reinforcement-learning based on "polarized user appraisals gathered throughout the course of a vocal interaction between a machine and a human". Their work outlined the design of a hypothetical experiment; we have taken this a step further by actually implementing this design in a live interactive study. We take user feedback to be the explicit reward signal (those user inputs that match the explicit positive or negative feedback). Like their work, our approach does require a lengthy explore phase at the outset.

3 System

In this section we explain our choice of robot, and how we modeled the dialogue for language learning with integrated robot modules.

Choice of Robot: Anki Cozmo Cozmo is small, has track wheels for locomotion, a lift, and a head with and OLED display which displays its eyes. The head has a small camera and a speaker with a built-in speech synthesizer (with a "young"-sounding voice). With a Python SDK, we can easily access Cozmo's sensors and control it. Importantly for our study, we will make use of Cozmo's camera for object detection, human face recognition, and locomotion functionality for navigation between objects. Cozmo does not have an internal microphone–we make use of an external one.

The choice of robot affects how humans will treat it, and it is important for our study that users perceive the robot as a young language learning child. We opted for the Anki Cozmo robot because Plane et al. (2018) showed that participants in their study perceived Cozmo as young, but with potential to learn. Cozmo's affordances are likewise consistent with this perceived age and knowledge-level. Cozmo is also a good option for this work because it has been recently demonstrated that humans perceive the same emotions and positive or negative valences from Cozmo's over 940 pre-scripted be-

haviors (McNeill and Kennington, 2019). Taken together, these studies show that (1) we can safely assume that human participants will treat Cozmo at an appropriate age level, and (2) we can assume that human participants will properly interpret Cozmo's behaviors as displays of emotion.

Indicating Objects If Cozmo is to learn denotations for physical objects, then objects must be present in the environment that Cozmo and a person share. Also, the person needs to be able to identify the object that Cozmo is attending to. Once these requirements are met, then Cozmo can learn the correct denotations for objects. Noting that Matuszek et al. (2014) has been able to successfully use deictic gestures to isolate objects, we assume participants will denote objects that the robots are already attending to, which is what adults do for children learning their first language (Hollich et al., 2000) (that is, the perspective Cozmo takes is egocentric). More practically, Cozmo is small, which places its camera very low to the surface of the shared environment. Therefore, Cozmo must be very close to objects to "see" them through its camera, which effectively isolates objects without the need for deictic gestures from the robot. When Cozmo does need to indicate an object, Cozmo moves its lift up and down while directly in front of the object of intended reference.

Social Conventions Motivated by Michaelis and Mutlu (2019), Cozmo needs to exhibit minimal "socially adept" behaviors if language learning is going to take place. We identify two behaviors that we incorporate into Cozmo: (1) *eye contact*; that is, in certain states (e.g., Cozmo is looking for feedback from the user) Cozmo looks up and turns in place until it finds a face, and (2) *motion*; that is, Cozmo must nearly always be moving–for several reasons, first to signal to an interlocutor that Cozmo is still functional and second, children who are learning language rarely sit still. These random motions occur outside of the task actions (explained below) and give priority to those task actions when they occur.

Learning To answer the question *can emotions serve as scaffolding to solve the cold-start language learning problem?*, we take a reinforcement learning (RL) approach. Given a dialogue state and a robot state, the RL regime learns which emotional valence to display: confusion or understanding. This learning takes place at the same time that the robot is learning grounded word meanings using WAC as it interacts with a person and its environment.

3.1 System Modules

For the balance of this section, we describe the modules that make up our word learning dialogue system and how they are integrated with the Cozmo robot. The modules include:

1. Visual Perception
2. Object Detection
3. Feature Extraction
4. Automatic Speech Recognition
5. Grounded Semantics
6. Action Management
   - Navigation
   - Emotional Displays
   - Word proposals
7. Emotion Management

Visual Perception The Visual Perception module handles the event of a new image being captured by Cozmo's camera. Cozmo's camera produces a color image at 30 frames per second (320x240 pixels). The output of this module is a single frame image.[1]

Object Detection This module uses the Mask RCNN graph (He et al., 2017) adapted taken from the tensorflow library. We used a model pre-trained on a dataset of sixty separately labeled grocery items from the MVTec D2S dataset (Follmann et al., 2018). We apply this configuration of the Mask RCNN model for drawing bounding boxes around objects in images received from the Visual Perception module. We discard the labels and only make use of the bounding box information. The output of this module is the bounding box information of all detected objects in view.

Feature Extraction The Feature Extraction module contains an image classification model built on the Keras implementation of VGG19 (Simonyan and Zisserman, 2014) which is trained using the ImageNet (Deng et al., 2009) corpus weights.[2] This module takes an image and bounding box information, extracts each sub-image containing each object, then passes those through the Keras model, thereby extracting features. We use the second-to-last (i.e., `fc2`) layer as the feature

[1] For our system, we only considered three frames per second and dropped the rest.

[2] We tested on more recent and principled models such as efficientnet (Tan and Le, 2019), but found the simpler Keras model to work better for our task.

representation of each object, which is a vector that represents the object. This model outputs a vector for each detected object.

We motivate this approach of using an existing object detector only for bounding box information and another model for object representation because pre-linguistic children can already detect isolated objects before they learn denotative words for those objects–our downstream *Grounded Semantic* module learns the mappings between words and objects. Moreover, this allows word learning to occur without relying on the limited vocabulary of any given object detector–those trained on imagenet only have a vocabulary of 1000 words, and those words are generally nouns, whereas attributes such as color and shape (i.e., adjectives) should be allowed.

Automatic Speech Recognition The *Automatic Speech Recognition* (ASR) module transcribes user speech. This module then categorizes user speech according to three exclusive dialogue acts:

- positive user feedback (e.g., *yes*)
- negative user feedback (e.g., *no*)
- object denotations (all other words)

The positive and negative feedback dialogue acts are used as environment signals to our reinforcement learning regime and are identified by simple word spotting. All other utterances are regarded as object denotations for the Grounded Semantic module.

Grounded Semantic Module The Grounded Semantic Module is tasked with learning word denotations as well as determining which word to utter in certain states. As noted above, for this we leverage WAC. This module takes in transcribed speech from the ASR module and the top (i.e., most confident) object feature representations from the Feature Extraction module (i.e., one set of object features per word use). In an *explore* state, the robot records the feature representations that it receives and assigns them as positive examples to words that are heard within a 10 second window. Negative examples for words are taken from the largest rectangular area of the image from outside of the top bounding box. Anytime a word has been heard three times, the WAC classifier for that word is trained. The classifiers themselves are scikit-learn logistic regression classifiers (with l2 normalization).[3] Trained classifiers can be improved each time a word is heard by re-training the classifier given the new training examples from the interaction.

Action Management For Action Management (which includes dialogue management), we use PyOpenDial (Jang et al., 2019). There are several navigational actions (the first three make up *explore* actions, the latter two *exploit* actions): `find-object`, `approach-object`, `indicate-object`, `propose-word`, `seek-face`. Several state variables are tracked to determine which of the above actions are taken, including the most recent navigation action, if the robot has found an object, and if the robot has approached an object. The robot begins in a `find-object` state where it does not yet see an object. This triggers random left and right turning, forward and backward driving until an object comes into view (determined by the Object Detection module). When an object is in view, the robot transitions to an `approach-object` state which alternates turning left and right to keep the object in the center of the robot's camera frame while driving short distances until the object takes up a specified percentage of the camera frame. At this point the robot transitions to `indicate-object` which it accomplishes by moving its lift quickly up and down multiple times. When the Action Management module enacts a `propose-word` action, the robot utters a word that it "thinks" it learned (i.e., the robot has a trained classifier for the word in question and it fits above a certain threshold for the object). After a proposal, the robot enters performs a `seek-face` action to ground with the interlocutor that it expects them to give it positive or negative feedback.

Emotion Management This module is where the RL (i.e., reinforcement learning) takes place. The RL model (which leverages PyOpenDial Q-Learning functionality implemented as a dynamic Bayesian network with Dirichlet priors and a Gaussian posterior) tracks just a single variable: `robot-confidence` (RC), a number that represents the robot's internal confidence that it should move into a `propose-word` state. The following modules affect the RC:

- ASR: if a positive feedback occurs anytime, the RC increases by 2; RC decreases by 4 if

[3] We attempted to use other classifiers, such as multi-layer perception, as well as other feature representations, such as efficientnet, but found that this model is the most effective for fast language acquisition in this setting.

negative feedback is heard.

- Action Manager: if a `propose-word` state is reached (resulting in Cozmo uttering a word), *and* there is positive feedback from ASR, then the confidence increases by 5. If negative, the confidence decreases by 4.

The emotional displays take place before a `propose-word` action. This module uses RL to learn whether to display an *understanding* emotion or a *confusion* emotion. The above listed modules alter the RC dynamically over time (though the min/max values of RC are -10 and +10 respectively). The reward policy is as follows: if RC is positive, the policy is rewarded +5 for displaying understanding, and -5 if it displayed confusion; if RC is negative, the policy is rewarded -5 for displaying understanding and +5 for confusion. In this manner, the RL can determine, on its own, the RC threshold for producing understanding vs. confusion displays.[4] We chose *confusion* and *understanding* for two reasons: first, because prior work has shown that confusion and understanding are opposite valences of the same affect which are very interpretable, particularly when looking at Cozmo's movement and eyes (McNeill and Kennington, 2019); and second, because confusion and understanding are emotions that lend well to the language learning task–the robot can display confusion in states where it is unsure how to act, and understanding in states where it knows how to act. To determine which behaviors would be perceived by users as *confusion* or *understanding*, we collected Cozmo's behaviors that were labeled with high confidence as either of those emotions by the model in McNeill and Kennington (2019). We then asked 7 people to watch recorded videos of Cozmo performing those emotions and rate them on a 5-point Likert scale. This resulted in 11 highly-rated behaviors (i.e., lasting from 3-10 seconds) for *confusion* or *understanding*. The Emotion Management model randomly selects one of the 11 for each emotion when producing a display of that emotion.

The full learning pipeline is depicted in Figure 1. Object detection occurs while users say words that refer to the objects in Comzo's view. Object features are extracted and used for WAC to learn the fitness between words and objects. If the word fits above a threshold, then Cozmo proposes that word to the user.

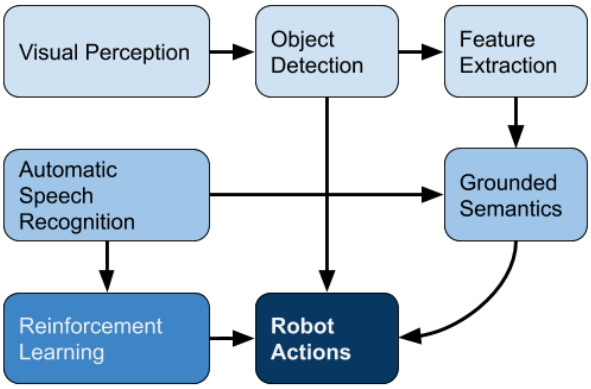

Figure 1: Schematic of our system: Visual perception passes camera frames to an Object Detector, which annotates objects with bounding boxes, then the Feature Extractor represents each of the detected objects as vectors that are passed to the Grounded Semantics module. The ASR transcribes speech, and passes those strings to the Grounded Semantics module and to the Reinforcement Learner (i.e., positive or negative feedback dialogue acts). Both Object Detection and Grounded Semantics pass their output to the Robot Actions (i.e., Action Management) which makes decisions about which actions to take, then actually performs those actions on the robot.

4 Evaluation

In this section, we explain how we evaluated our model with real human participants to determine if emotional displays increase engagement for language learning. We used two versions of our system: one which only performed the language learning task, and one which additionally included displays of emotion–the choice of which emotion was decided by a RL model. Our evaluation included objective measures logged by the system, as well as subjective measures collected using participant questionnaires.

4.1 Procedure

Study participants agreed to meet in a small room in the University's Computer Science building. The conference room is set up for the participant interaction as follows: a table is placed to one side of the room, with one chair positioned in the middle of the longer side for the study participant. The experimenter sits at the head of the table, with a laptop positioned between himself and the participant. This laptop is running the robot's interactive

[4] More principled models of deep reinforcement learning are available, but we opted for this approach because we wanted our RL module to learn from minimal real interactions–deep learning approaches are known to require large amounts of data.

script and the microphone that feeds the ASR module. A container of objects (specifically, pentomino blocks) is placed on the table; a handful of these have been randomly scattered on the table before the participant arrives in the room. The Cozmo robot is not introduced to the participant until the participant has signed an informed consent form and the task has been explained to them.

The experimenter was present to monitor the state of the robot and the microphone, troubleshoot any problems that might arise, and answer any questions the participant might have over the course of the interaction. The experimenter was permitted to offer a constrained set of coaching tips to the participant during the interaction, if the participant needed a reminder of the task or the initial instructions. The study participant and the robot were observed with cameras, which recorded audio and video from the interaction. Following each interaction the user moved to the experimenter's laptop and completed a questionnaire. Following the completion of both interactions and subsequent surveys, the participant was paid eight U.S. dollars.

We recruited twenty-one study participants to interact with the Cozmo robot for two fifteen-minute periods over the course of a single session. Study participants were largely college students recruited from Boise State University's Computer Science department. Participants' ages range from their late teens to their forties. Eight of the participants were women; thirteen were men. Following each fifteen-minute interaction, the participant was asked to answer every question of the same questionnaire. The entire study took approximately one-hour.

We employed a within-group study design, meaning that each participant went through the same procedure twice, one time in which the independent variable (i.e., with emotional display) was present, and again when it was absent (i.e., without emotional display). To mitigate learning effects, the order in which the test condition was presented was alternated.

4.2 Task

First, the Cozmo robot was introduced to the participant, with an explanation of the following affordances and instructions: **(1)** Cozmo has a camera that can see them and the world; **(2)** Cozmo has a microphone that can hear them; **(3)** Cozmo doesn't know anything, but is "curious" to learn more about the world; **(4)** for the next 15 minutes, it is the participant's job to try to teach Cozmo as many words as they can, using the objects in the room, whatever they have on them, and their imagination; **(5)** if Cozmo gets off-track, they are allowed to pick Cozmo up and move it around; **(6)** when Cozmo is looking up, it is looking for their face; **(7)** when Cozmo "feels confident" enough, it will guess a word – if it gets it right, say "Yes." If not, say, "No." This feedback will help Cozmo learn. Figure 2 shows Cozmo in its task setting in two states: observing an object (left figure) and seeking a face (right figure).

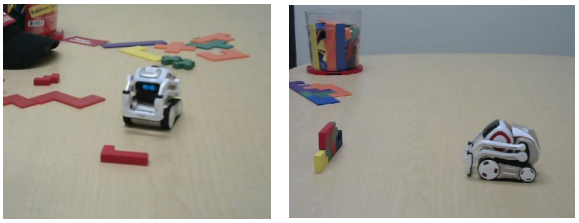

Figure 2: Cozmo looking down at an object (left) and looking up, seeking a face (right).

4.3 Metrics

System Logs We track the number of utterances (termed "Heard Words") made by the participants, including positive and negative feedbacks, and the number of proposals made by the robot which, taken together, form a proxy for engagement: higher numbers denote more engagement.

Participant Questionnaires We also evaluate the robot based on questionnaire responses written by the study participants following both sessions of the study. We used the Godspeed Questionnaire (Bartneck et al., 2009), a likert-scaled questionnaire with 24 questions ranging from negative to positive ratings of a robot's anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. In addition to the Godspeed questions, we also asked participants the following to ascertain their perceptions of our system and robot:

- How attached to the robot did the user feel?
- Were they engaged by the robot?
- What did they think the robot wanted?
- What did they think the robot was trying to do?
- Would they like to spend more time with the robot? Why or why not?

4.4 Results

Table 1 shows the results of the effect that emotional displays had on heard words, positive feedbacks, negative feedbacks, and proposals (note that

proposals represent trained WAC classifiers that reached the threshold for being uttered). Comparing the results of the experimental trials in which the robot displayed emotions to the control trials, it is apparent that the amount and quality of the user feedback to the robot improves in the presence of emotional displays. The sole caveat is negative feedback, which was offered the most on average by users interacting with a robot that wasn't making emotional displays.

Table 1: The effect of emotional displays on a language-acquisition task

(Mean / std. dev)	without emotions	with emotions
Heard Words	58.5 / **69.4**	**72.9** / 107.1
Positive Feedbacks	11.9 / **12.2**	**16.3** / 27.5
Negative Feedbacks	**7.4** / 7.0	6.6 / **6.5**
Proposals	7.8 / 7.8	**9.8** / **7.5**

Exploring the effect of participant learning on the language-acquisition task in Table 2 shows that users spoke more words and offered more positive feedback in the second trial than in the first, on average. Negative feedback was equivalent between the two trials, and the robot made more proposals in first trials, on average. This shows that learning effects had a minimal impact on user interaction with the robot.

Table 2: The effect of participant learning on the language-acquisition task

(Mean / std. dev)	first trial	second trial
Heard Words	60.6 / **70.3**	**64.1** / 103.3
Positive Feedbacks	9.8 / **11.4**	**16.8** / 26.5
Negative Feedbacks	**6.7** / 6.9	**6.7** / **6.7**
Proposals	**9.1** / 7.8	7.5 / **7.6**

Next, we analyze the participant surveys to see if the presence of emotional displays biased the participant toward higher estimations of robot intelligence. For both the control and experimental trials, the average estimated age of the robot is two years old, which follows prior work using Cozmo (Plane et al., 2018) and is an appropriate assigned age range for this study. Additionally, the participant surveys reinforce the ambiguous role of emotion in human estimations of robot intelligence, irrespective to trial order, as seen in Figure 3.

User engagement also appeared largely uninfluenced by the presence of robot emotional displays, or the trial order, as seen in Figure 4. This is reinforced by the high p-value between user responses to the Godspeed questionnaire and the total number of emotional displays produced by the robot. As

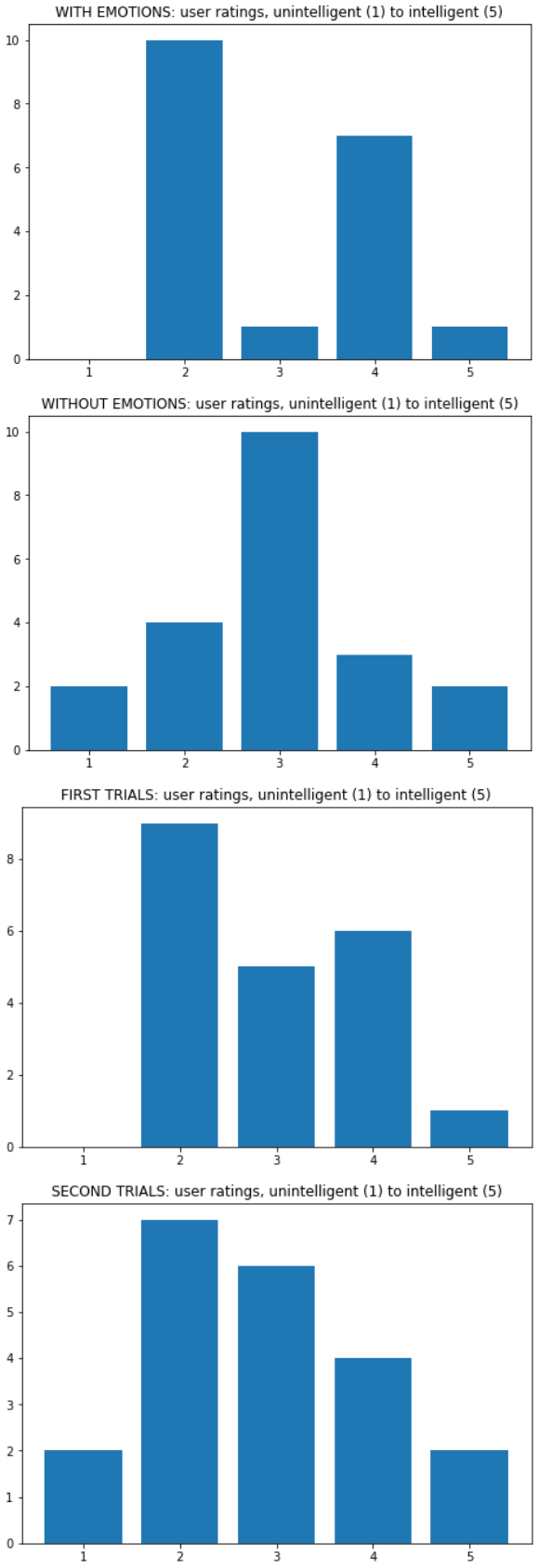

Figure 3: X-axis: Participants' ratings of robot intelligence from 1: unintelligent to 5: intelligent. Y-axis: the number of participants who selected that response.

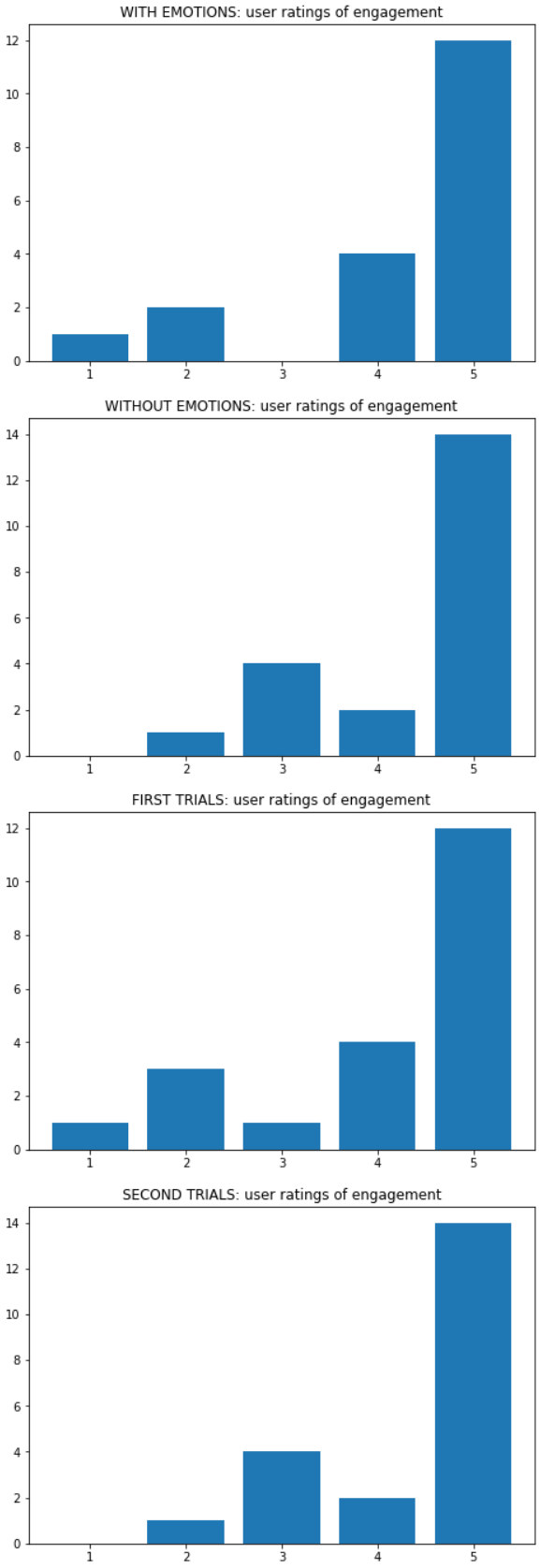

Figure 4: X-axis: Participants' responses to the question, "Would you like to spend more time with the robot?" from 1: not at all to 5: very much. Y-axis: the number of participants who selected that response.

see in 3, there was a weak correlation and weak evidence to support a relationship between user interest and engagement with the robot, and the total number of emotional displays produced by the robot.

Table 3: Correlations between the total number of emotional displays and the following user questionnaire responses

	correlation	p-value
moves elegantly	0.48	0.03
is nice	0.40	0.09
is interesting to interact with	0.34	0.15
would like to spend more time with	0.17	0.49

In our RL module, the Q-Learning algorithm learned to put all weight onto one emotional display to the exclusion of the other for each interaction. This may have been due to the training batch size and training time for the Q-Learning algorithm (10 max samples and a 5 ms sample rate, rate to keep the interaction from slowing down). This did not have a negative effect on the choice of emotional displays produced by the robot; to the contrary, the emotional displays chosen by the RL module facilitated engagement.

5 Conclusion

We conducted an experiment with twenty-one participants who had to rely on the robot's displays of *confusion* and *understanding* and their own performance in a language acquisition task as context. We analyzed our results by comparing the participants' survey responses and the robots' Grounded Semantics classifiers between the experimental and control trials. We found that a robot that displayed a combination of confused and understanding emotional displays – positive- and negatively-valenced emotion – gathered more inputs, and more useful inputs (positive feedback), than a robot that only engaged in task-specific actions (i.e., orienting to objects; seeking out the user's face). This in turn led to the robot making more word proposals, which did not lead to greater engagement. User estimations of the robot were generally more positive estimations, supporting our choice of the Anki Cozmo robot for this task. Emotional displays did not incline participants to over-estimate the robot's language understanding. We can conclude that emotion is an important aspect in handling the cold-start problem where a system can only use words it has heard.

In future work, we will test different policies for the reinforcement learning regime including measures for novelty rewards (i.e., hearing new words) as well as repeated words. Another aspect that demands further investigation would be the *timing* of emotional displays in the language learning interaction. Importantly, we will go beyond the two basic emotions explored here and incorporate additional emotions (e.g., the 8 valence pairs used in McNeill and Kennington (2019)) as the basis for additional engagement and perhaps use emotional states as features for the grounded classifiers.

Acknowledgements We thank the anonymous reviewers for their insights and feedback. This work was approved under the Boise State University IRB #126-SB20-012.

References

Ralph Adolphs. 2002. Recognizing Emotion from Facial Expressions: Psychological and Neurological Mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, 1(1):21–62.

Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81.

Cynthia Breazeal. 2005. *Designing Socially Intelligent Robots*. National Academies Press, Washington, D.C.

Lola Cañamero. 2005. Emotion understanding from the perspective of autonomous robots research. *Neural Networks*, 18(4):445–455.

Josep Arnau Claret, Gentiane Venture, and Luis Basañez. 2017. Exploiting the Robot Kinematic Redundancy for Emotion Conveyance to Humans as a Lower Priority Task. *International Journal of Social Robotics*, 9(2):277–292.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Emmanuel Ferreira and Fabrice Lefèvre. 2015. Reinforcement-learning based dialogue system for human-robot interactions with socially-inspired rewards. *Computer Speech and Language*, 34(1).

Charles J. Fillmore. 1981. Pragmatics and the description of discourse. *Radical pragmatics*, pages 143–166.

Patrick Follmann, Tobias Böttger, Philipp Härtinger, Rebecca König, and Markus Ulrich. 2018. MVTec D2S: Densely segmented supermarket dataset. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings. *arXiv*.

G J Hollich, K Hirsh-Pasek, and R M Golinkoff. 2000. Breaking the language barrier: An emergentist coalition model for the origins of language learning. *Monographs of the Society for Research in Child Development*, 65(3).

Julian Hough and David Schlangen. 2017. It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot. In *Proceedings of the 2017 Conference on Human-Robot Interaction (HRI2017)*.

Youngsoo Jang, Jongmin Lee, Jaeyoung Park, Kyeng-Hun Lee, Pierre Lison, and Kee-Eung Kim. 2019. PyOpenDial: A Python-based Domain-Independent Toolkit for Developing Spoken Dialogue Systems with Probabilistic Rules. In *Proceedings of EMNLP*.

Mark Johnson. 2008. *The meaning of the body: Aesthetics of human understanding*. University of Chicago Press.

Malte F Jung. 2017. Affective Grounding in Human-Robot Interaction. In *Proceedings of HRI'17*.

Casey Kennington and David Schlangen. 2015. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China. Association for Computational Linguistics.

Sara Kiesler and Jennifer Goetz. 2002. Mental Models and Cooperation with Robotic Assistants. *CHI'02 extended abstracts on Human factors in computing systems*, pages 576–577.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2013. Age-of-acquisition ratings for 30,000 English words.

Cynthia Matuszek. 2018. Grounded Language Learning: Where Robotics and NLP Meet. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*.

Cynthia Matuszek, Liefeng Bo, Luke S Zettlemoyer, and Dieter Fox. 2014. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In *Proceedings of AAAI 2014*. AAAI Press.

Lorraine McCune. 2008. *How Children Learn to Learn Language*. Oxford University Press.

David McNeill and Casey Kennington. 2019. Predicting Human Interpretations of Affect and Valence in a Social Robot. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany.

Joseph E. Michaelis and Bilge Mutlu. 2019. Supporting Interest in Science Learning with a Social Robot. In *Proceedings of the Interaction Design and Children on ZZZ - IDC '19*, pages 71–82, New York, New York, USA. ACM Press.

Jekaterina Novikova, Gang Ren, and Leon Watts. 2015. It's Not the Way You Look, It's How You Move: Validating a General Scheme for Robot Affective Behaviour. In *Human-Computer Interaction – INTERACT 2015*, pages 239–258, Cham. Springer International Publishing.

Sarah Plane, Ariel Marvasti, Tyler Egan, and Casey Kennington. 2018. Predicting Perceived Age: Both Language Ability and Appearance are Important. In *Proceedings of SigDial*.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition.

Linda Smith and Michael Gasser. 2005. The Development of Embodied Cognition: Six Lessons from Babies. *Artificial Life*, (11):13–29.

Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv*.