DO YOU FEEL ME?: LEARNING LANGUAGE FROM HUMANS WITH ROBOT EMOTIONAL DISPLAYS

by

David McNeill

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Computer Science

Boise State University

May 2020

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

David McNeill

Thesis Title: Do You Feel Me?: Learning Language from Humans with Robot Emotional Displays

Date of Final Oral Examination: 1st May 2020

The following individuals read and discussed the thesis submitted by student David McNeill, and they evaluated the presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Casey Kennington, Ph.D.	Chair, Supervisory Committee
Jerry Fails, Ph.D.	Member, Supervisory Committee
Hoda Mehrpouyan, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Casey Kennington, Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

ACKNOWLEDGMENTS

I must acknowledge my adviser, Dr. Casey Kennington, for his support and guidance, which included the graduate assistantship that supported my research and allowed me to pursue a Master's degree in Computer Science at Boise State University. I must also acknowledge Dr. Casey Kennington for the gift of his time, which he always made me feel was available in abundance in our meetings, in spite of the many far more significant demands for his attention that I knew existed. I would also like to thank the members of my advisory committee, Dr. Jerry Fails and Dr. Hoda Mehrpouyan, for providing me with the attention and feedback that were necessary to fill the gaps in my knowledge and make my thesis complete.

There are many other professors at Boise State University I must acknowledge for the patience and support they offered in their classes, as well as the mentorship they have freely given outside the classroom. I must begin with Dr. Sole Pera, who allowed me to join her research group while I was still close to the beginning of my time in the program, and whose generous collaboration with her colleagues (both inside and outside the department), as well as with the students in her lab, I would hope to emulate. Dr. Edoardo Spezzano, Dr. Francesca Spezzano, Dr. Amit Jain, Dr. Tim Anderson, Dr. Jyh-haw Yeh, and Dr. Michael Ekstrand were all essential in making me a better programmer and researcher. I would also like to thank Luke Hindman for his instruction and collaboration by choosing me to help tutor CS 121.

I would also like to thank the members of the SLIM research group, both past and present, whom I had the pleasure of learning from, being critiqued by, and sharing a

jug of potato-chocolate milk with. Your names are included in order of milk consumed: Andrew Rafla, Daniele Moro, Alex Mussell, Gerardo Caracas, Stacy Black, Aprajita Shukla, Jake Carns, and Sam Schrader.

Finally, this thesis is dedicated to my parents, Kevin McNeill and Sheila McNeill, who find themselves in the unfortunate position of having given more love and support than seems fair when their contribution is compared against the object being dedicated to them. I love you both.

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. The second requirement poses a problem: If a robot is capable of asking a human teacher well-formed questions, it will lead the teacher to provide responses that are too advanced for a robot, which requires simple inputs and feedback to build word-level comprehension.

In a live interactive study, we tested the hypothesis that emotional displays are a viable solution to this problem of how to communicate without relying on language the robot doesn't–indeed, cannot–actually know. Emotional displays can relate the robot's state of understanding to its human teacher, and are developmentally appropriate for the most common language acquisition setting: an adult interacting with a child. For our study, we programmed a robot to independently explore the world and elicit relevant word references and feedback from the participants who are confronted with two robot settings: a setting in which the robot displays emotions, and a second setting where the robot focuses on the task without displaying emotions, which also tests if emotional displays lead a participant to make incorrect assumptions regarding the robot's understanding. Analyzing the results from the surveys and the Grounded Semantics classifiers, we discovered that the use of emotional displays increases the number of inputs provided to the robot, an effect that's modulated by the ratio of positive to negative emotions that were displayed.

TABLE OF CONTENTS

ABSTRACT vi

LIST OF TABLES x

LIST OF FIGURES xi

LIST OF ABBREVIATIONS xv

1 Introduction 1

1.1 A Child's First Words, a Robot's First Words 1

1.2 The Case of the Missing Co-located, Interactive Dialogue Setting 2

1.3 Thesis Statement 4

2 Related Work 6

2.1 Background 7

2.1.1 Why Use Emotional Displays to Address the Cold Start in Language Learning? 7

2.1.2 How to Represent Language on a Robot? 8

2.1.3 How Can Robots Use Emotion to Acquire Language? 9

2.2 Work Relevant to Robotic Emotional Displays in Human Interaction Studies 10

3 Preliminary Work 16

3.1 Capturing Cozmo's Emotional Displays 16
3.1.1 Sounds 18
3.1.2 Movement via Internal State 18
3.1.3 Face Animations 19
3.2 Amazon Mechanical Turk Worker Ratings 20
3.3 Emotional Interpretation of Robot Displays 23
3.4 Consistent Interpretation of Displays 25
3.5 Connections to Present Work 26

4 Methods 28
4.1 Visual Perception 30
4.2 Object Detection 30
4.3 Feature Extraction 30
4.4 Automatic Speech Recognition 31
4.5 Grounded Semantics 32
4.6 Robot Actions 34
4.7 Reinforcement Learning 38

5 Experiment 42
5.1 Participant Recruitment 43
5.1.1 Study Setting 43
5.1.2 Study Task 43
5.1.3 Hypothesis 44
5.2 Evaluation 45
5.2.1 Language data 45
5.2.2 Participant Surveys 45

5.3 Results 46

6 Conclusion 59

6.1 Limitations and Future work 60

REFERENCES 62

A Parsing Feedback Lists 66

B Augmented Godspeed Questionnaire 67

C Cozmo Internal Features 71

LIST OF TABLES

3.1 Valence of 16 specific affects. 26

5.1 The effect of emotional displays on a language-acquisition task 46

5.2 The effect of valence on a language-acquisition task 47

LIST OF FIGURES

3.1 Taken from [27]: Three example frames of a video recording of Cozmo for a *bored* animation. 17

3.2 Taken from [27]: Example of face tracking and the corresponding extracted face frame. 19

3.3 Distribution of emotion labels as assigned by the workers. 22

3.4 Total counts of labels per task. 23

3.5 Taken from [27]: Common designer name tokens compared with annotator emotion labels; lower tokens map to counts on the left side of the bars. 24

4.1 The pipeline of data inputs and transformations for the Grounded Semantics module. **Step (1) Object Detection**: taking an image captured with the Cozmo robot's camera, this module's Mask RCNN model draws a bounding box around the most likely object in the image. **Step (2) ASR**: from the audio inputs provided by the laptop microphone, the Google ASR model creates a text transcription of the most likely words said by the user. **Step (3) Feature Extraction**: using the bounding box from Object Detection, crop the image from the Cozmo robot's camera and pass it to a pre-trained Image Classifier – the VGG19 – and take the second-to-last layer as a vector representation for the object. **Step (4) WAC**: use the vector representation from **(3)** as feature data to represent the word label obtained from the ASR; with this labeled data, build a classifier trained on this example (in addition to a negative example created in the same manner, but using the area *outside* the bounding box found in **(1)**. **Step (5) Propose word**: after training this classifier with two or more examples of the same word (and their negative examples), use this classifier to predict the best word label that describes the current object the Cozmo robot is looking at. 40

4.2 The four stages of an "Episode" for the Robot Actions module, moving clockwise and beginning from the top-left: 1) Task action: find object, go to object, or point to object; 2) Emotional display: emote understanding or confusion; 3) Word proposal: receive positive or negative feedback; 4) Locate human face. 41

5.1 The setting for the interaction. Top: Cozmo sees an object. Bottom: What the participant see upon entering the room. 50
5.2 X-axis: Participant ratings from 1: unintelligent to 5: intelligent. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials. 51
5.3 X-axis: Participant ratings from 1: foolish to 5: sensible. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials. 52
5.4 X-axis: Participant ratings from 1: ignorant to 5: knowledgeable. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials. 53
5.5 X-axis: Participant ratings from 1: unpleasant to 5: pleasant. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials.. 54
5.6 X-axis: Participant ratings from 1:not at all to 5:very much. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials. 55
5.7 X-axis: Participant ratings from 1: not at all to 5: very much. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials. 56
5.8 X-axis: Participant ratings from 1: dislike to 5: like. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials. 57

5.9 X-axis: Participant ratings from 1: not at all to 5: very much. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials. 58

LIST OF ABBREVIATIONS

WAC – Words-As-Classifiers Model of Lexical Semantics

ASR – Automatic Speech Recognition

Mask RCNN – Mask "Region Based Convolutional Neural Network"

Q-Learning – "Quality" learning

CHAPTER 1

INTRODUCTION

Wish I could see through, see deep into you
And know what you're thinking now
And if I were to need it, I need some kind of sign
Let me know 'cause I can't read your mind

"Do You Feel Me"-

Anthony Hamilton, Diane Warren

1.1 A Child's First Words, a Robot's First Words

Imagine a child's first words. Only, imagine that instead of saying, "Mom," or "Dad," with a burst of delighted giggles, this child held up a stuffed animal and asked you, *what is this?*, with a clarity that demonstrated not only an understanding of what each word meant but also of the underlying structures they contain (and also a disapproval of contractions): *what* indicating a question, *is* indicating existence, and *this* pointing to a sensory phenomenon. Thankfully for parents, their children do not possess such a preternatural level of language understanding right away. Yet this is precisely how automated systems, such as robots, are currently programmed to "learn" words: they ask explicit questions like *what is this?*

A system can't ask a person questions like, "what is this?" because that leads a person to assume the system not only understands the question it has just asked, but

also possesses the commensurate knowledge and experience that would accompany such language understanding in a human [32]. This would lead the human partner to treat the system as a peer rather than a completely naive entity which possesses a limited understanding of the world and requires their direct attention and support. As a result, users would be less likely to provide the robot the simple inputs and feedback and frustrate its attempt to learn language.

This is a type of *cold-start* problem. From recommender systems, the cold-start problem refers to the beginning of a recommender's history with a user, when the recommender lacks data with which to make a recommendation. In language acquisition, this cold-start problem can be understood as the beginning of the system's exposure to language, when it has the capacity to learn language but lacks the ability to actually use language in such a way that would provoke a person into teaching it correctly (e.g., by asking the person questions).

1.2 The Case of the Missing Co-located, Interactive Dialogue Setting

One attempt to circumvent this cold-start problem is through offline training on reams of labeled language data. This method is clearly artificial, which, although fitting for the artificial robot that is learning the language, does skew the robot's model of what language is and how it is used.

Humans learn their first language by face-to-face transmission from an experienced user within a shared (co-located) context [9]. Here, the learner is not only learning a language, they are learning how to make connections between the language and those things in the world that both they and the teacher are experiencing. This

ability to build associations between symbols and the things they signify (termed "grounded semantics") is leveraged by language-learners to build links between words and gestures, abstract words and emotions [27], and even to other words. But the first crucial steps are taken in dialogue, grounding concrete words to their easily imageable referents [29].

Training a language model strictly from text, or a static dataset, changes language learning from being an interactive method humans use to *engage* with others vis a vis the physical world, to an entirely *abstract* framework with no basis in *concrete* reality. This may be appropriate for higher-levels of human conversation, when a majority of words are in reference to entirely abstract concepts, but unless all knowledge is shared equally among the speakers, a shared understanding of concrete concepts is required if participants encounter a point in which one of them is unfamiliar. This core set of concrete concepts – and their physical referents – are built in childhood; developmentally, they should be the first concepts we strive to teach an artificial system, such as a robot, in order to lay a plausible foundation for a human-level understanding of language. Keeping language as an entirely abstract set of symbol-patterns through text-only models may be sufficient for limited language tasks, such as one-word prediction, but has been proven insufficient for novel language generation or more extensive extrapolations on a theme [14].

The cold-start problem then remains: if we must train a robot to use language in an interactive environment, and that robot cannot use words it doesn't already understand (otherwise violating the concept illustrated above by the gifted child), how then is that robot meant to *acquire* language if it can't *use* language?

1.3 Thesis Statement

In a live interactive study, we tested the hypothesis that *emotional displays* are a viable solution to an initial lack of language information. Our reasons are two-fold: emotional displays can relate the robot's state to its human teacher 2; also, emotional displays are developmentally appropriate for the most common language acquisition setting (an adult teaching a child), and would not lead a human user to make incorrect assumptions regarding the robot's comprehension.

We programmed a robot to independently and autonomously explore the world and elicit relevant word references and feedback from the participants, who were tested both with a robot that displayed emotions and a robot that did not. Analyzing the results from the surveys and the Grounded Semantics 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 (positive or negative) of the emotional display, and the total number of emotional displays in the trial.

We will cover this experiment's procedures in greater detail in Chapter 5 – an overview of the relevant systems to this interaction task, and their dependencies, is given here: For our interaction we make use of the Anki Cozmo robot as a learning platform. The Cozmo contains a camera that we use to acquire visual information about the world and to locate the study participant by identifying their face. The researcher is located beside the interaction with a laptop; the researcher's laptop has a microphone that is used to record the words that are said by the participant. Using these parallel streams of audio and visual data, a classifier is trained on the researcher's laptop. When it is the appropriate time to propose a word for the object the Cozmo robot is currently seeing, the robot will refer to this classifier. The robot

attempts to elicit feedback from the user by locating their face; if this is insufficient, the researcher can prompt the participant to provide the Cozmo positive and negative feedback to its word proposals. Based on this feedback, in the experimental condition, pre-scripted emotional displays are triggered by a Reinforcement Learning module that is training in the background.

The purpose of this study is to solve the cold-start problem, and also to provide further inspiration for more engaging, flexible, responsive, and natural spoken dialogue systems. These systems would include 'language acquisition' strategies, which may use emotions for signaling confusion or understanding to alert the user to the system's state in the language-learning task, and also to reward them for their attention as the system adapts and responds to their input.

In the next chapter, we show how our proposed solution has the theoretical support to motivate further research in strategies for designing Spoken Dialogue Systems that can double as "natural language acquisition" devices.

CHAPTER 2

RELATED WORK

The work presented in this chapter attempts to resolve the cold-start problem and leverages emotional display as a method of bootstrapping communicative cues for word learning. Building off of Plane et al. (2018) [32], we are confident that the Cozmo platform is the right platform for this task because prior work has shown that study participants' perceptions of age and the knowledge-level of this robot are consistent with a human child who is still acquiring language (further motivating a human user to adopt the 'teacher' role in a language-acquisition task). Moreover, the robot's affordances are likewise consistent with this perceived age and knowledge-level (it can identify objects, recognize human faces, and navigate obstacles). The same study showed that Cozmo is an agreeable social partner to a variety of users, who find that it avoids the uncanny valley by its deliberately mechanical design combined with its demonstrative "personality". More recently, we demonstrated that humans perceive the same emotions and positive or negative valences from Cozmo's over 940 pre-scripted behaviors [27]. 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 emotional displays.

In the following section, we explore the theories that form the background for our

study with the Cozmo robot and its emotional interaction with a casual human user for the express purpose of learning word meanings.

2.1 Background

In this section we explore those theories – pertaining to spoken dialogue, human-robot interaction, language learning, and emotion – that form the background for our study with the Cozmo and its emotional interaction with a casual human user.

2.1.1 Why Use Emotional Displays to Address the Cold Start in Language Learning?

From Developmental Psychology, we learn that emotion is pre-linguistic, universal, interpretable and imitated by infants [2]. This is relevant for the reasons stated in the introduction: the user must interact with the robot as if it is learning its first language, otherwise the context for the robot's language will be skewed. From Neuroscience, we see that emotions are composed, meaning that the experience of emotion is conscious and not a reflex [1]. This is significant because if emotions were reflexive responses to the environment they would be an entirely internal, autonomic process, and therefore useless as communication, or as an intentional protocol between sender and receiver.

In Cognitive Science, *The Interactive Brain Hypothesis* is a framework for human cognition that incorporates the enactive approach. According to this view, artificial systems can only be thought of as "thinking" insofar as they can interact with other thinking systems; *affect* (as demonstrated by emotional displays) is an important piece of this interactive cognition between humans [11]. "Situated dialogue between 2 embodied entities" is the original setting for language; all else is derived, according

to Clark (1996) [9] and Fillmore (1981) [13]. According to the enactive approach as explained in [36], cognition is not an act of passively processing information as it encounters our brains, cognition is an active, embodied, and situated process of engaging and relating to an external environment via sense and perception.

According to our background taken from a variety of different disciplines: emotion is appropriate for the setting of language acquisition; emotions are within the scope of a conscious interaction; understanding is not possible without interacting with another thinking system.

If we are to use emotions as part of an extra-linguisitic communicative protocol between a robot and a human user, we must observe the guiding principles from [16], whose review of emotional displays by artificial agents found that in order for such displays to effective they must be reliably recognized by people. According to [34], humans can interpret humanlike affective nonverbal behavior in robots. We explore these human interpretations of robot affect in greater detail – and their implications for this thesis – in Chapter 3.

2.1.2 How to Represent Language on a Robot?

Grounded Semantics: Grounded semantics is the study of how the meaning of words ("semantics") is connected ("grounded") into our human senses and experiences. For example, when a child is learning their first words, those words often denote physical objects, such as *ball*. Grounded semantics takes into account things like how the ball looks, how it feels, etc. – features that are significant to a robot engaging a human in dialogue in a real-world setting, to take the task from this thesis as an example. This task is at the center of what this thesis is trying to accomplish: to program a robot that can naturally engage and learn language from a human, "language" here

represented by these grounded semantic mappings between words and the physical objects detected by a robot an interactive, spoken dialogue setting.

In the words-as-classifiers model (WAC) [21], research conducted by the thesis writer's advisor, Dr. Casey Kennington, each word in a language is built into a classifier. Each word's classifier is trained on "not / is" examples of that word – these examples are real-world referents that either exemplify the word or do not (e.g., in this paper, the classifier for "red" would be trained on pictures of things that are red and pictures of things that are *not* red). In this paper, the referents are image data taken from Cozmo's camera.

The reasoning behind choosing WAC are itemized as follows:

- WAC is grounded – any semantic model that claims to learn words that children learn needs to link the physical world with words spoken by an adult
- WAC can learn from minimal amounts of training data – children can "fast map;" i.e., they can learn how words map to objects with only a few examples [35]

2.1.3 How Can Robots Use Emotion to Acquire Language?

Reinforcement Learning: In Q-learning – a type of Reinforcement Learning (RL) algorithm, and the method used in our experiment – the agent moves between states by taking actions which are determined by reinforcement (positive or negative) from the environment. We use reinforcement learning because it is adaptive; using a pre-defined policy based on the environment's feedback, an agent (i.e., our robot) can learn the best actions to take given the environment's feedback to prior actions given whatever state the agent was in when it made the action. In this case, the environment

is the human user – this means that the robot's display of emotions is adaptive to the user's feedback to it. Within Human-robot interaction, it is important for the user to have a "shared" sense of space with the robot, one in which the robot is aware of and responsive to the user's presence and actions [20]. For example, in [17], researchers construct a mapping from basic "emotions" to RL primitives and demonstrate an artificial agent behaves according to its emotions in a way that is consistent with psychological and behavioral literature.

2.2 Work Relevant to Robotic Emotional Displays in Human Interaction Studies

In this section we explore other researchers' work that is comparable to our own: an interactive study between a robot and a casual human participant, in which the robot incorporates live grounding of semantics by gathering inputs from a dynamically changing environment, which it then uses to inform separate decision-making processes: a reinforcement learning regime (in the experimental condition) and a deterministically defined action model (in both the experimental and control condition).

Jekaterina Novikova found that emotionally expressive robot behavior improved human-robot collaboration in [31]: "[W]hen the robot is acting in an emotionally expressive way, the human puts more effort into the activities that add to the success of the collaborative task so they last longer compared to the situation when the robot is acting neutrally, without showing emotional expressions." For their research, the task was one in which a human and a robot had to work cooperatively to move objects from one location to another, across a virtual environment. In this task, emotional displays caused the average distance between the robot and the human to

increase – this may have lead to an increase in self-reported enjoyment for the human participant, but it also lead to a decrease in the human-robot team's efficiency in completing the task – they took longer. For a task like language acquisition, however, this decrease in efficiency could lead to an increase in effectiveness; that is, when the goal of the task is not externally focused, but is in fact interactional (give the robot more words), we demonstrate this increase in human user enjoyment also leads to improved task effectiveness.

The work of Mason Bretan, Guy Hoffman, and Gil Weinberg demonstrated it was easier for users to recognize a robot's intended emotion when that emotion was presented as a dynamic embodied display, rather than a static pose [6]. Also, they showed that "automatically generated affect responses cause participants to show signs of increased engagement and enjoyment compared with arbitrarily chosen comparable motion parameters." This research motivates our decision to select emotional displays at random from a pre-selected list of understanding and confusion displays, that have already been demonstrated to be interpreted with the appropriate emotion by a group of people (our research group, in the Preliminary Work chapter). This way we hope to increase user's engagement and enjoyment with emotional displays that feel spontaneous and appropriate.

The following works by Lola Canamero are essential to this thesis, in which she directly asks – and answers – the question of why robots should have emotional features. It is Canamero's view that emotional features can make a robot appear more life-like and believable to humans, and "therefore, humans [will be] more prone to accept them and engage in interactions with them." Canamero also lays out the rationale for this thesis' means of testing and evaluating emotional displays: "[W]e must be able to show ... that emotions improved the performance or the interaction

capabilities of our robot and how... [a]n obvious way of doing this is by running control experiments in which the robot performs the same task 'with' and 'without' emotions and comparing the results" [7]. The latter two papers, however, are where our thesis diverges. The goal of [25] is to imitate a neural "emotion circuit" with their Reinforcement Learning module – according to Canamero's view, this would teach the robot which action to pursue according to a "hedonic reward." In [26], they use an "emergent neural network" that increases its number of nodes according to the number of different stimuli the robot encounters . Our own experiment is in contrast to these approaches – rather than modeling human emotions in a robot as a means of better understanding the ways that humans process and respond to their environments, our RL regime is designed to learn the emotional displays purely for the purposes of engaging a user. It is assumed by our research that these processes are understood well enough to be used in an interactive task such as language learning – language learning being the true focus of this research.

Ada Kim et al. identified important design principles of an interaction between a teenage user and a robot in [23]. These principles included collaboration and characterization. Characterization is addressed by the Cozmo robot's cartoonish animated face design, and the expressiveness of its emotional displays. Collaboration is addressed by our task, in which the user works with the robot toward it learning which words it should use to describe different objects in the environment.

The work of Cynthia Breazeal must also be addressed here, for her groundbreaking work in defining social robotics, and continuing on to more recent user studies measuring user perceptions and acceptance of a social robot. Breazeal noted that "emotion-inspired mechanisms" could be used to make a robot function better in a complex environment, by allowing it to interact more appropriately with others

"than it could with its cognitive system alone" [5, p. 5]. Predicting the design of the Cozmo robot, Breazeal noted that such an emotion system could "implement the style and personality of the robot" [5, p. 5]. Cozmo, with its 940 pre-scripted emotional displays, each programmed by the same team of designers at Anki, bears such a unique personality. This design was important to the success of our study in that it would inform a user's mental model for how the robot operated. "Social and emotional factors also greatly affect the individual's willingness to adopt the technology," Breazeal notes, citing the study of Kiesler and Goetz (2002) that suggests that people apply a social model when observing and interacting with autonomous robots [22]. Breazeal writes this may be because, paraphrasing the work of Don Norman (1990), "in order for people to interact with another entity, they must have a good conceptual model of how that entity operates" [5, p. 4] [30]. If they have such a model, people can explain and predict what a robot may do, understand its reasons for doing it, and know how to elicit the desired behavior from the robot. For our thesis, by adhering to natural signals and mappings (e.g., emotional displays), the state of our robot's Grounded Semantics classifiers could "become intuitively understandable to people" [5].

With J. M. Kory-Westlund, Cynthia Breazeal measured children's views of a robot over time, and found that children were slightly more accepting of the robot at the posttest. Likewise, they noted that if a child's acceptance of the robot is connected to their view of it being animate, or human-like, than that view was less likely to change. Although we conducted our study with adults rather than children, these results have a direct bearing on our survey results of participants between the first trial and the second – if in the first trial the robot displays emotions, it affects the user's responses more significantly in the second trial, when they will be less likely to

back off their initial measurements of the robot as being more animate, likeable, and human [24].

J.E. Michaelis and B. Mutlu (2019) demonstrated that a robot utilizing "socially adept" behaviors, which included expressive speech and eye contact, found the robot to be "friendlier and more attractive, reported a higher level of closeness and mutual-liking for the robot, had higher situational interest," and performed better on a task-related evaluation. For our Cozmo robot, we incorporate eye contact in both the control and experimental condition, considering this "socially adept" feature to be a basic requirement of the language learning task, insofar as it demonstrates to a user that their presence is noticed and required by the robot to continue at the end of each episode. Additionally, researchers programmed their robot to produce small semi-randomized head motions to demonstrate to the child that the robot was working. For our thesis, we incorporate this same approach of producing small semi-randomized movements with the Cozmo robot's track wheels and lift, in the event the robot identifies an object, or when a detected object fails to pass a check that would warrant the robot navigating to it. This is done for the same reason as this paper – to avoid the impression the robot has stopped working simply because it is still [28].

A confounding factor that is not accounted for in this thesis but could be explored in future research is the interpretation of emotions from robot behavior that is purely task task-oriented. This concern is addressed by [8], in which a Pepper robot was programmed to convey emotions simultaneously as it pursues a primary task, such as completing a wave gesture or transporting an object. However, researchers discovered that using this approach the emotions that were well conveyed to users were limited to happiness and sadness, the motion features that mediated these emotions

being jerkiness, activity, and gaze. Because these motion features were identical for our experimental and control conditions, we take this to indicate that whatever incidental perceptions of robot happiness or sadness were communicated to the study participants, were equivalent across the trials, and did not effect our results.

The research of Ferreira et al. outlines the approach of our thesis to reinforcement-learning based on "polarized user appraisals gathered throughout the course of a vocal interaction between a machine and a human" [12]. As with the above-mentioned research, this paper was outlining 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 delineated in two separate lists). Our research, however, does suffer the shortcoming addressed in this research: a lengthy explore phase at the outset, during which the robot produces a jumble of confused and understanding emotional displays that bear very little intelligible correspondence to the state (i.e., termed below as *Robot Confidence*).The researchers of this paper cite Williams (2008) as an example of how expert domain knowledge can be applied to circumventing this explore phase [19].

The work of E. J. Jacobs et al. demonstrates the mapping of Reinforcement Learning onto the corresponding emotional states of joy, distress, hope, and fear: "[F]rom a human-robot interaction point of view the emotional signal can be expressed to a human observer." The authors of this paper, however, do not extend their simulation beyond validating emotion dynamics established by the "psychological and behavioral literature" [17], rather than measuring their effect on the human participant of a live interactive study with a co-located robot.

CHAPTER 3

PRELIMINARY WORK

Before we can answer the question of whether emotion can help a robot acquire language, we first need to address the assumption that humans who interact with our robot will:

- Interpret the Cozmo robot's displays as emotional
- Assign predictable emotion labels to those displays

This chapter is based on the work of the author, David McNeill, and his advisor, Casey Kennington, from: [27]. The work of that paper is presented in this thesis to provide context for how we choose the emotional displays for the Cozmo robot to display in our experiment, as well as informing our understanding of how humans form their perception of the Cozmo's overall affect and valence based on the particular features of its emotional displays.

First, we explain the data we collected, and then provide analyses of that data to answer the questions listed above.

3.1 Capturing Cozmo's Emotional Displays

In this section, we explain the data we collected and offer some analysis of that data. Our goal in this data collection is to better understand how people interpret the

affective display of Cozmo as it performs its pre-scripted animations, and how their perceptions of those animations differ from what we interpreted to be the affective display that the animation designer intended the robot to portray.

For each of Cozmo's 940 available, pre-scripted animations, we recorded video and audio of the robot's behavior. For each recording, we position Cozmo in a starting position where it faced the camera, then initiated the animation. We kept the camera as close to Cozmo as possible while still recording the animations from within a single camera position (i.e., for some animations, Cozmo moved around, requiring wider camera coverage).

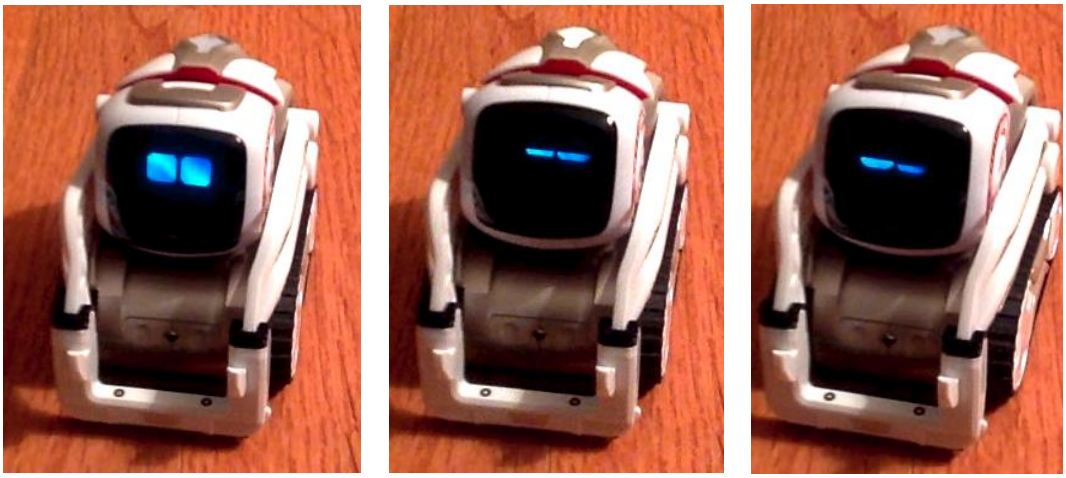

Figure 3.1: Taken from [27]: Three example frames of a video recording of Cozmo for a *bored* animation.

An example of three frames derived from one of these video recordings is in Figure 3.1: though Cozmo does not appear to move, its eyes have the appearance of looking around and portraying boredom.

In [27], we explored three sources of information that are available to people when observing and interpreting the robot animations: (1) Cozmo's produced sounds,

(2) facial animations, (3) and movements. For the experiments below, we obtained representations of each of these modalities.

3.1.1 Sounds

Obtaining Cozmo's produced sounds was straightforward: we extracted the audio from the recordings for each animation. The other two modalities required additional steps which we explain in the following subsections.

3.1.2 Movement via Internal State

The Cozmo SDK allows developers to obtain the internal state of the robot at any state change update event. Some examples are itemized below; the entire set of 47 state variables is listed in the Appendix:[1]

- `left_wheel_speed`
- `lift_position_height`
- `accelerometer_x`
- `gyro_x`

On average, animations had 73 state change updates with sorrow-labeled animations being the longest (92 on average), and surprise-labeled animations being the shortest (58 on average). For each change in the state of the robot, we recorded the entire state of the robot resulting in a sequence of state changes for each animation, which we used to represent movement over time.

[1] We only considered variables that did not remain constant across all animations.

3.1.3 Face Animations

The internal state updates do not include information about the state of the face. Cozmo's face display is an OLED (organic light-emitting diode) where the facial animations are pre-defined and inaccessible through the SDK. To obtain facial information for each animation, we passed the video recordings through a computer vision processing script that located the eyes by color (which was unique to the scenes in the recordings) and created a bounding box around them. Each frame of each video recording for each animation was passed through the script, resulting in an extracted face for each frame. An example of what this looked liked for a single frame is depicted in Figure 3.2; Cozmo is facing the camera, the script located the face (i.e., the eyes) and formed a bounding box, then extracted the contents of that bounding box into an individual face image. Processing each animation recording in this manner resulted in a sequence of face images, one for each frame where the face was found in the frame (i.e., there were some frames where Cozmo was not facing the camera, and therefore no face images were extracted).

Figure 3.2: Taken from [27]: Example of face tracking and the corresponding extracted face frame.

3.2 Amazon Mechanical Turk Worker Ratings

We then posted these recordings (i.e., containing the audio and video) on Amazon Mechanical Turk with the following instructions for the workers:

"You will be shown a video of a small robot. Please describe what the robot is doing in the video, and provide a selection of the emotions that you think the robot is displaying (in this paper we only focus on the resulting emotion labels).

Following [33], we used the following 16 emotions:

- interest
- alarm
- confusion
- understanding
- frustration
- relief
- sorrow
- joy
- anger
- gratitude
- fear
- hope

- boredom
- surprise
- disgust
- desire

Taking note from [3] that there is no mutual exclusivity between emotions, we allowed workers to be able to select any number of these emotions using check boxes, thereby not constraining the number of emotions they could assign, however we did not give them a free-form input so as to keep the task within reasonable constraints. Each worker was paid $1.00 to describe and label 10 randomly assigned videos and could repeat the process for another set, if they desired. The emotion check boxes were arranged randomly. Each animation recording was labeled by two workers.

This resulted in 1,870 labeled recordings (to ensure that each worker received the same number of labels, some were labeled 3 times). Figure 3.3 shows the distribution over the labels. The most common label is interest at 12.2%, the least common is disgust at 2.82%, with a fairly uniform distribution over each of the 16 labels. We take this to mean that no single label was either over- or under-favored by the workers.

Figure 3.4 shows a count of the number of labels for each animation. For example, if an animation has a recording that the worker labeled as surprise and disgust, then that animation received a count of 2. Of the 1,870 labeling tasks, 1008 only received one label, 486 received 2, 165 received 3, and we found smaller counts for higher numbers of labels. From this we infer something important: while more than half of the workers assigned a single label to recordings, nearly half received more than one label. This is the first evidence of ambiguity in interpreting the affective display of a particular behavior.

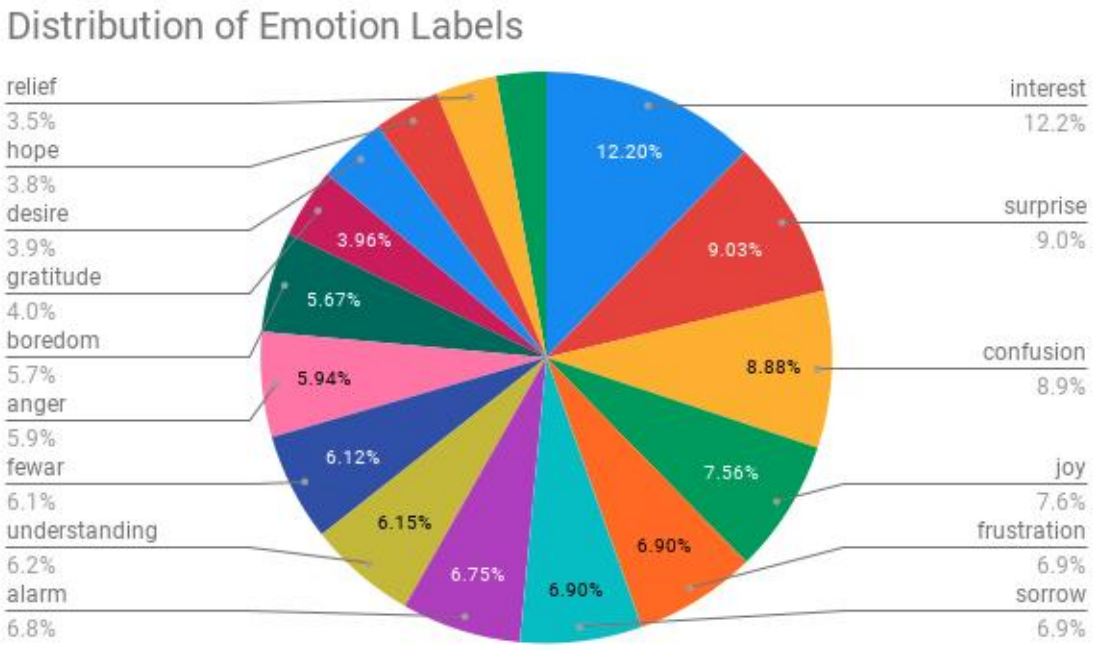

Figure 3.3: Distribution of emotion labels as assigned by the workers.

To further measure the challenge that people have in interpreting the robot's affective display, we calculated inter-annotator agreement using Cohen's Kappa statistic [10]. As each recording received labels from two different workers, we treated the two workers as two different annotators with one important proviso: when two workers agreed on at least one label, we marked the two annotations as agreed upon. This resulted in a Kappa score of 0.26, which is considered in the "fair agreement" range. This agreement was not higher because the annotators could choose from among sixteen difference choices. That this value is not below or equal to zero tells us that there is some agreement in how people perceive a robot's affective display.

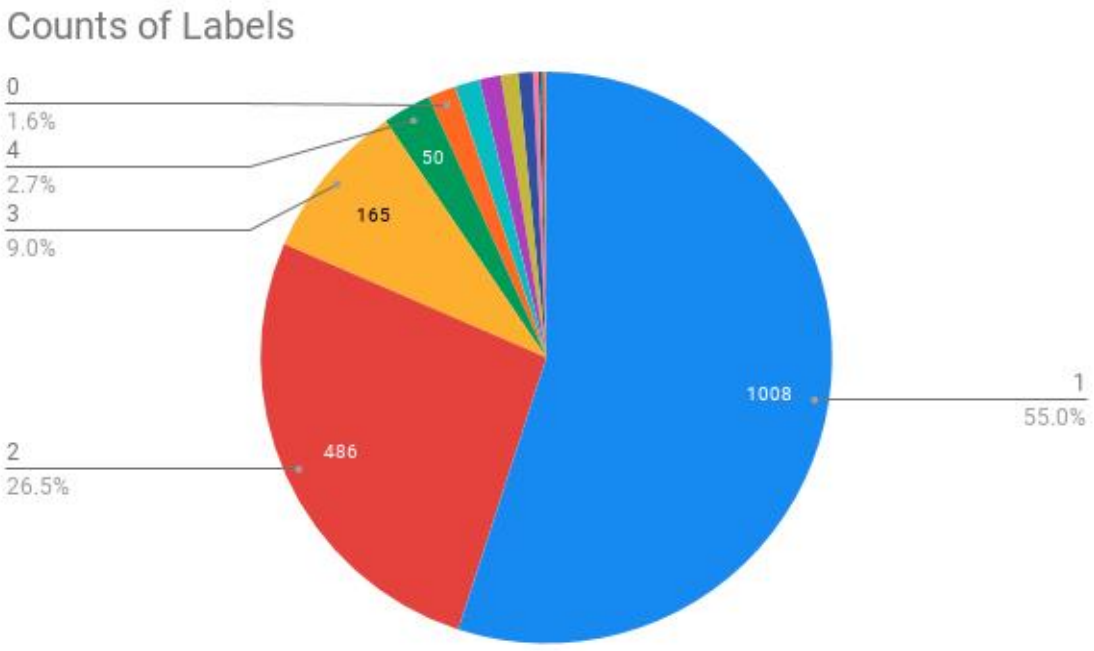

Figure 3.4: Total counts of labels per task.

3.3 Emotional Interpretation of Robot Displays

Though it is not possible to fully recover the intent of the designers who created the animations, we can estimate the intended affective display of Cozmo from the designer-written animation names. Below are some examples of these animation names:

- `bored_event`
- `greeting_happy`
- `explorer_driving01_loop_01_head_angle`
- `rollblock_fail`

Note that some names have words that denote affective displays, while others only focus on the function of the animation and not how it might be interpreted as an emotion or affect. By taking the individual word tokens (i.e., between the underscores) we identified the common words that we interpreted to denote affective displays: bored, celebration, fail, focused, frightened, frustrated, happy, determined, lose, neutral, success, surprise, upset, win; 145 of the 940 animation names had at least one of these tokens in them.

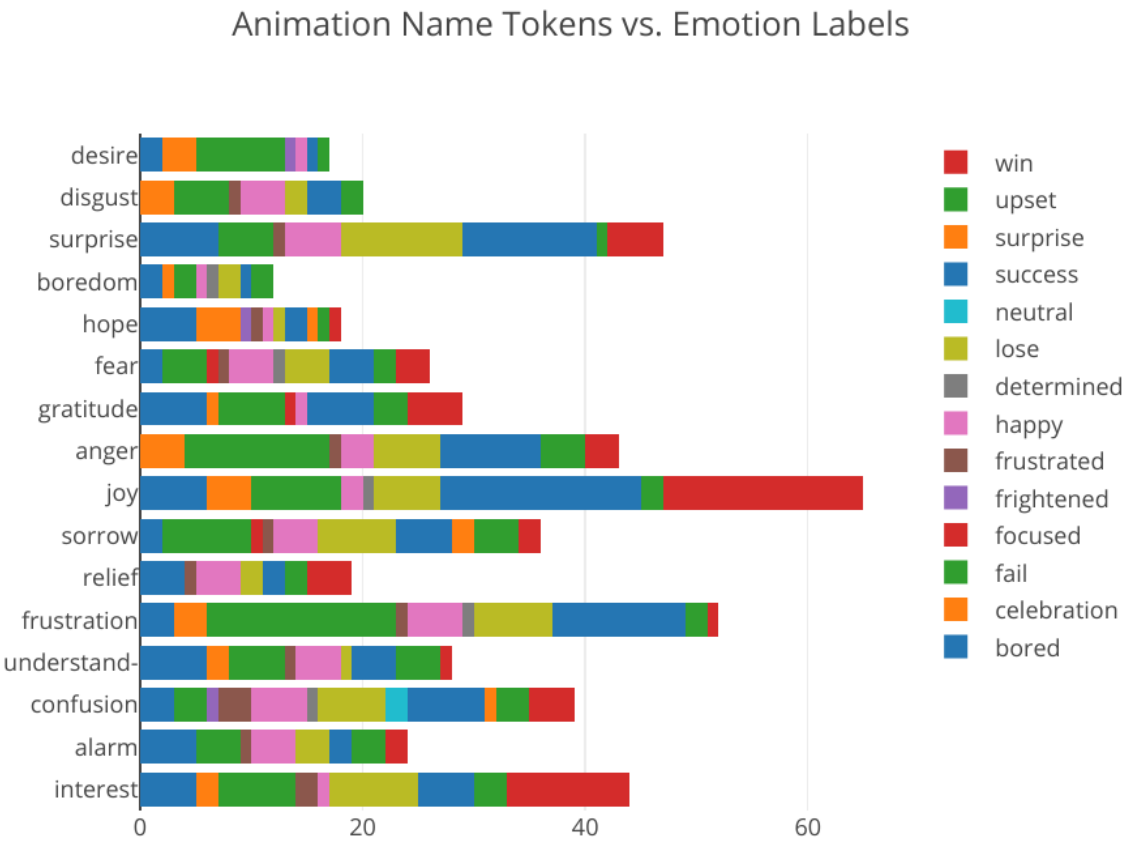

Figure 3.5: Taken from [27]: Common designer name tokens compared with annotator emotion labels; lower tokens map to counts on the left side of the bars.

We compared the annotator-labeled affects with the affective tokens in those corresponding animations. This comparison is shown in Figure 3.5, where the labels are

on the y-axis and the count of tokens of intended affect interpretation is represented in the bars. In some cases there are clear analogs to the emotion list we used from [33], i.e., bored=boredom, frightened=fear, frustrated=frustration, happy=joy, and surprise=surprise, but even for those pairings, affect was interpreted in many different ways. The label for surprise, for example, was used to identify animations with bored, fail, frustrated, happy, lose, success, upset, and win name tokens. In this case, surprise as a token in an animation name was never actually interpreted as surprise by the annotators. On the other hand, the token win was interpreted by workers as nearly every affective display (see the red/rightmost items in each bar).

This answers the first question, by confirming that people do interpret emotions from the robot's behavior.

3.4 Consistent Interpretation of Displays

Though interpretation of affective display is not mutually exclusive, as shown in the above section, certain affects can be treated as opposites. We therefore break apart the task of classification of the 16 possible affective displays into 8 binary classifiers for valence pairings, following [33] by coupling the emotions into positive and negative valence pairs as shown in in Table 3.1. We hypothesize that doing so will allow us to consider which modalities influence which affects and valence pairs more directly. We can then make use of the individual binary classifiers to make graded predictions about what affect a human would interpret.

Taking this approach, we were able to predict the likelihood that a user would interpret the Cozmo's robots emotional display as either representing "confusion" or "understanding" with over 58% accuracy. For the understanding-confusion valence

positive valence	negative valence
interest	alarm
understanding	confusion
relief	frustration
joy	sorrow
gratitude	anger
hope	fear
surprise	boredom
desire	disgust

Table 3.1: Valence of 16 specific affects.

pair, only considering the face animations for a feature-set worked as well as using the face animations and Cozmo's sound effects. Clearly, some modalities have information that is sometimes contributory or inhibitive when considered in conjunction with other modalities. We interpret this to mean that when determining if a robot is displaying understanding vs. confusion, showing some kind of display in a "face" (even if this display only involves animated eyes) plays an important role.

By only including facial animations, this understanding-confused binary prediction model scored a reasonable correlation of 0.24, when compared against the annotations of human labelers for randomly generated animations.

3.5 Connections to Present Work

This result shows that our understanding-confused classifier yields reliable predictions when compared to humans for novel animations. Though we did not use this model specifically in this thesis, this classifier can be used in a specific tasks in which the distinction between understanding and confused emotional displays is relevant to the task, e.g., a task in which the robot is learning from a casual human user, and the predictive accuracy of the robot's model may not be apparent to the user, and a

minimal level of social engagement with a person is important. A model that predicts high confusion would need to alter its behavior if this did not reflect the robot's model, i.e., the robot could predict with great accuracy the correct word label to describe an object.

We found those animations that both Amazon Mechanical Turk workers assigned understanding and confusion; from these, we surveyed our research team to find the animations that were the least ambiguous – that is, along a Likert scale from Confused to Understanding, we selected those animations clustered at either end of the ratings. We took the bottom ten as our Confused animations and the top ten for our Understanding animations for the experiment.

CHAPTER 4

METHODS

In this chapter, we explain the methodology of our experiment to answer the question, *do emotional displays on robots help engagement for language acquisition?* The focus of this methodology is not the implementation of any one module, but how those modules can be used in combination to give a robot the ability to learn words as it interacts with a human user.

This methodology was refined over the course of a short pilot study, completed using volunteers from the author's research group.

In this methodology, there are two areas where learning takes place:

- **Learning the grounded meaning of words**: the robot builds a Words-As-Classifiers (WAC) Grounded Semantics model as it interacts with the human user and also as it interacts with the world – the world here represented as those objects the robot detects and approaches. WAC has not been tested in this type of interactive scenario before.

- **Learning which valence of an emotion to display**: using a reinforcement learning strategy, we author a learning policy for the robot to learn whether it should display confusion or understanding based on a state variable – *Robot Confidence* – that is affected by what the robot hears from the human par-

ticipant. The specifics for this learning are explored in the Automatic Speech Recognition section below.

The rest of this section describes the integration of the various modules to ensure the robot could operate independently of any externally controlled commands for the duration. These modules include

1. Visual Perception
2. Object Detection
3. Feature Extraction
4. Automatic Speech Recognition
5. Grounded Semantics
6. Robot Actions
    - Navigation
    - Emotional Displays
    - Word proposals
7. Reinforcement Learning

Before examining these modules, we must introduce the concept of *Robot Confidence*. *Robot Confidence* is a state variable tracked across all of the modules, which is affected by positive and negative user feedback, as well as the number of times Cozmo trains its Grounded Semantics classifiers. Specifically, *Robot Confidence* is an integer instantiated at 0 and capped at -10 and +10. *Robot Confidence* is referred to by several modules to determine the robot's present level of success in the interaction (or 'utility,' for the Q-Learning algorithm in the Reinforcement Learning module).

4.1 Visual Perception

The Visual Perception module handles the event of a new image being captured by Cozmo's camera. This is passed to the Object Detector.[1]

4.2 Object Detection

This object detection module is a Mask RCNN graph [15] adapted taken from the tensorflow library. This graph is pre-trained on a dataset of sixty separately labeled grocery items. We apply this configuration of the Mask RCNN model towards drawing bounding boxes around pentomino blocks in images captured from the Cozmo camera, from the Visual Perception module. We take the top (i.e., the most confidently identified) object returned by the Mask RCNN graph and use this object's bounding box to guide the robot's navigation and feature extraction for the Grounded Semantic module. Though the Mask RCNN model also provides object labels, we ignore those and only use the bounding box information.

This module is the bottleneck for the interaction – the average lag in processing a cropped image through the Mask RCNN graph is 0.5 seconds. This was the reason for limiting the number of images processed by the object detection module to one, for each phrase transcribed from the user by the Automatic Speech Recognition module.

4.3 Feature Extraction

The Feature Extraction module contains an image classification model built on the Keras implementation of VGG19. This model is trained using the ImageNet corpus

[1]This module tries to place the latest image in the Image Queue, a thread-safe queue with a size limited to one, to be processed by the Object Detection module that is running in a separate thread. If this fails, the except block runs a pass statement.

weights.

This model is used to make predictions of what this cropped image contains – the second-to-last layer is used as the feature representation of the object and returned to train the classifiers of the Grounded Semantics module.

4.4 Automatic Speech Recognition

The *Automatic Speech Recognition* (ASR) module parses transcribed user speech. User speech can be classified according to three exclusive dialogue acts: 1) descriptions of objects that can act as labels for the Grounded Semantics model; 2) positive feedback affirming the robot's most recent action; or 3) negative feedback disapproving of the most recent action made by the robot.

This parsing is accomplished by comparing inputs captured from the Google ASR to two lists to determine if an input can be classified as either positive or negative feedback. These feedback words are excluded from the Grounded Semantics model (which does not prevent the model from training classifiers based on homophones (e.g., learning "know" based on instances when the participant says, "no")), and used to determine which word models should be preserved or destroyed, as well as determining if the robot should make a confused or an understanding emotional display, according to the Q-Learning reinforcement learning module.

State variable *Robot Confidence* is affected by the ASR module. In the case of positive feedback that is parsed from the input stream, *Robot Confidence* is boosted by two; in the case of negative feedback parsed from the input stream, *Robot Confidence* drops by four. Otherwise, for those inputs that are successfully matched to perceived objects in the robot's environment, *Robot Confidence* increases by one. In the ex-

perimental condition, these changes in *Robot Confidence* are communicated to the Reinforcement Learning module using a thread-safe queue of size one, the Stimulus Queue. In turn, these 'stimuli' influence the training of the Q-Learning model and the next emotion that is selected.

Due to the nature of the Google ASR and the user study, occasionally words of an obscene or distracting nature are heard and then proposed by Cozmo. Words are considered distracting due to the context of a robot speaking to a participant, in which the robot's statements could be construed as referring to the participant themselves. We created a list of obscenities to prevent Cozmo from proposing words that would interrupt the interaction and distract from the participant's goal. This list is built on an ad-hoc basis over the course of the studies, according to those words that presented themselves.

Those inputs that don't match any of the values present in the three lists are passed on as labels to the Grounded Semantics module (along with the time that they were heard), along with the object collected from the Match Queue at the beginning of the process.

The feedback lists are defined in the appendix.

4.5 Grounded Semantics

The Grounded Semantics module follows the Words-As-Classifiers (WAC) [21] design for grounded lexical semantics. WAC allows the robot system to possess a shared grounding of a word label with the study participant. This means that each word heard by a user becomes a classifier model trained on positive "is" or negative "is not" examples of that word. Those labels that are spoken by a user and heard by

Cozmo are assumed to be positive reference to the object that is contained within the bounding box, if the time that the label is heard is less than 10 seconds after the object is seen. This time period was chosen after a period of trial and error spent testing the robot, and considering the lag introduced to its interaction by the ASR, object detection, and feature extraction models. The largest square area outside of the bounding box in the image is used as a negative example to train the word's classifier.

Both the features for the negative and positive examples are obtained from the Feature Extraction module. These features and matching labels are used to construct a temporary "language cache," whose size is checked each time a match is made. If the language cache has more than three matches, it is used to construct the corresponding WAC classifiers. These classifiers are the scikit-learn implementation of a logistic regression classifier. If a word's classifier is already present in the WAC model's dictionary of classifiers, these new examples are used to update that classifier's weights; otherwise, a new classifier is constructed. Each time words are learned in this manner the WAC model is saved as a pickle file and the temporary language cache is cleared; also, the *Robot Confidence* state variable is incremented. This pipeline of data inputs and transformations for the Grounded Semantics module is illustrated in Figure 4.1.

The Grounded Semantic module is where the Reinforcement Learning's policy is set; i.e., the utility function that weighs the "quality" for an emotional display given the current state (defined by *Robot Confidence*), based on changes in values to *Robot Confidence*, which itself is dependent on the types of feedback the user provides. Here, *Robot Confidence* is capped at positive ten and negative ten – based on the likelihood of hearing positive or negative feedback (and the associated weights assigned to these

events (-4 for negative feedback; +2 for positive feedback)), along with the fact that *Robot Confidence* increments each time the Grounded Semantics module builds an association between a user input and the extracted features of a cropped image, the value of two was selected as a threshold for when the Cozmo robot could make word proposals of its own (i.e., the robot would utter the word) based on the classifiers from the Grounded Semantics model. Based on this policy, two was selected as a threshold to allow for the robot to make proposals quickly, but also to enforce a delay in proposals upon receiving negative feedback. Once *Robot Confidence* is greater than the value of two, it enables the Robot Actions module to make proposals using the WAC classifiers constructed in the Grounded Semantics module. The labels captured by the ASR module immediately following a prediction are interpreted as feedback to that prediction. Any input that does not match one of the values from the positive feedback list is interpreted as the participant's disapproval of the proposed word, and taken as a cue to destroy the existing classifier for that word in order to discourage repeated incorrect proposals. Feedback matching an item from the positive feedback list results in an addition of positive five to *Robot Confidence*.

4.6 Robot Actions

As many of the robot action calls are relegated to this module as possible, to prevent sharing the robot across threads and bogging down the interaction with competing calls to the single robot object.

The Robot Actions module is wrapped with an instance of a Dialogue Manager, a decision modeling object defined within the PyOpenDial framework [18]. The Robot Actions module is called whenever the Dialogue Manager is triggered, which occurs

when the Dialogue Manager is passed state updates. This occurs at the conclusion of each call to the Robot Actions module. The state variables that are passed to the Dialogue Manager at the conclusion of each episode are "near object" and "found object," whose truth values determine the next action to be chosen by the Dialogue Manager. This chosen action is then sent to the Robot Action module by the subsequent call to trigger.

At the end of every call to the Robot Actions module – what we term an "episode" – the robot orients itself to the participant's face by using a random search pattern. An overview of the order of the actions produced by the robot in an episode are given in Figure 4.2.

Navigation The three possible actions that are defined by the task action model are: "find object," "go to object," and "point to object." Depending on which action is sent to the Robot Actions module, the robot either enters a random search pattern ("find object") until the Navigation Queue is populated by an object found by the robot's object detection model; it drives towards an object until certain conditions (detailed in the Appendix) are met that determine the object has been reached ("go to object"); or the robot raises and lowers its lift repeatedly as a means of gesturing to an object which it has already found and approached in the preceding episodes ("point to object").

To restrict the robot's view to only include those objects that make it obvious to the participant what the robot is looking at, the Cozmo resets its head to its lowest angle at the beginning of each episode in order to minimize the potential to be distracted by bounding boxes encapsulating entities that are not present directly in front of it on the table and are therefore more difficult for a participant to discern.

Additionally, the robot's lift is raised to it maximum height so as not to block the camera's view, misdirect the object detection model, or influence the training of the classifiers in its Grounded Semantics model.

Bounding box information for the nearest relevant object is obtained from the Navigation Queue and passed to a method that checks if the bounding box passes certain conditions that would ensure it encapsulates an object that would be of real relevance to the robot, and not simply a random entity on the horizon that the model has mistakenly perceived to be an object, and which an observing participant would not be able to tell what the robot is looking at. These conditions are that the bounding box is less than 40% of the area of the screen, the top of the bounding box is greater than 1% of the screen-size away from the top of the screen, and the bottom of the bounding box is less than 60% of the screen-size away from the top of the screen.

Once these conditions are met, the coordinates are given to another function that determines the amount that the robot should turn and drive in order to place the center of the bounding box near the center of its field of view. If the x-coordinate of the center of the box is within 56 to 44% the width of the screen, the robot stops turning. Otherwise, 0.5 is subtracted from this value, the difference is then multiplied by -50 degrees (Cozmo's API determines a negative degree value to signal a turn to the left), and this is divided by four to create more conservative movements on the part of Cozmo, who – due to the time required by its Object Detection module – is operating on at least a half-second lag behind the participant. If the y-coordinate of the center of the bounding box is within 36 to 48% the height of the screen (these values being less than 50% to account for Cozmo's relatively short lens, which makes objects that are nearby appear further away), then Cozmo stops driving to the object. If not,

0.42 is subtracted from this y-coordinate (0.42 instead of 0.5 because of Cozmo's short lens which makes objects that are quite near appear further removed), the difference is multiplied by -10 millimeters, the product is raised to the power of four to account for parallax (objects that are further away move toward the center of Cozmo's viewfinder more slowly than those that are closer), and divided by two to create more conservative movements for the same reasons stated for the turn decision. If the Robot Actions module is processing the "find object" action, Cozmo will only implement the turn decisions, ignoring the drive commands entirely. If the module is processing the "go to object" action, Cozmo implements both the turn and drive decisions. In either case, if the bounding box sent to the object check function fails three times, the action is abandoned and the state variables "near object" and "found object" are passed in a system update. However, after the first bounding box passes the object check, this fail-counter is reset back to zero to account for any real object that may have entered the frame. If more than two bounding boxes in the same episode pass the object check, or the decision for the robot's turn angle is less than two degrees, then "found object" is set to True and the "find object" action is completed. If the selected action is "go to object", one additional condition must be met: the chosen drive distance must be less than 5 millimeters. If this is the case, then "near object" is also set equal to True and its value passed to the Dialogue Manager in a state update.

Emotional Displays In the experimental condition, the robot performs an emotional display once per episode following the completion of the navigation task within the Robot Actions module. The Robot Actions module gets the emotion from the queue and the robot performs the emotional display. Emotional displays are

chosen within the Reinforcement Learning module. The displays are selected from the Cozmo's broad repertoire of pre-scripted animations to evince either "confusion" or "understanding," using the process explained in the Data Analysis section of the Preliminary Work chapter.

Emotional Displays are determined by the Reinforcement Learning module, which runs a separate Dialogue Manager on a different thread. Updates in the robot's confidence level are passed to the Reinforcement Learning module using a thread-safe queue of size one, the Stimulus Queue, from the Automatic Speech Recognition module, as it parses inputs for feedback.

Proposals Following an emotional display in the experimental condition, or the navigation action in the control condition, the robot then checks its confidence level to ascertain if it has received sufficient positive feedback – or if it has learned enough recent words – in order to make a word prediction. The robot must have a confidence greater than two in order to make a prediction. This low threshold is to encourage early prediction of words, while also preventing the robot from speaking if it receives negative feedback.

4.7 Reinforcement Learning

Our robot follows a Q-Learning algorithm in which the agent navigates between states (an integer called *Robot Confidence*, which fluctuates according to the most recent user feedbacks, as outlined in the ASR and Grounded Semantics sections above) by using actions (either an understanding or confused emotional display). These actions are weighted according to the present state, and the human partner's previous feedback to that state (positive or negative).

The Reinforcement Learning module is reset for each trial. This is done in order to properly frame the cold-start problem and model an agent that is beginning an interaction with no prior knowledge of how to communicate with a human user.

The valence of the emotional display is decided by the Reinforcement Learning module, implemented using the PyOpenDial framework and a separate Dialogue Manager object that is running a Q-Learning algorithm (implemented as a Dynamic Bayesian Network; specifically a Markov Decision Process) that is training simultaneously on a separate thread. Once the valence has been chosen by the Q-Learning algorithm (either "understanding" or "confused"), the specific animation is selected at random from the corresponding list of animations (understanding and confused animation lists are in the appendix).

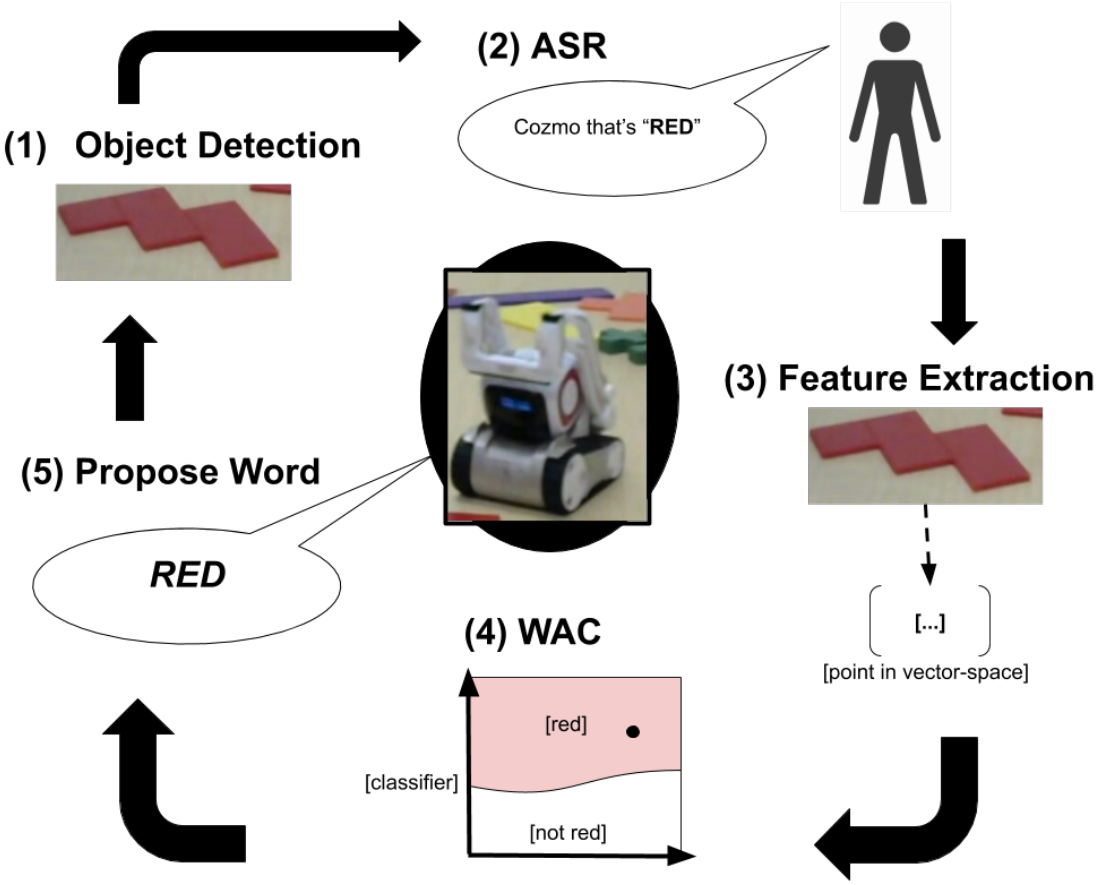

Figure 4.1: The pipeline of data inputs and transformations for the Grounded Semantics module. **Step (1) Object Detection**: taking an image captured with the Cozmo robot's camera, this module's Mask RCNN model draws a bounding box around the most likely object in the image. **Step (2) ASR**: from the audio inputs provided by the laptop microphone, the Google ASR model creates a text transcription of the most likely words said by the user. **Step (3) Feature Extraction**: using the bounding box from Object Detection, crop the image from the Cozmo robot's camera and pass it to a pre-trained Image Classifier – the VGG19 – and take the second-to-last layer as a vector representation for the object. **Step (4) WAC**: use the vector representation from **(3)** as feature data to represent the word label obtained from the ASR; with this labeled data, build a classifier trained on this example (in addition to a negative example created in the same manner, but using the area *outside* the bounding box found in **(1)**. **Step (5) Propose word**: after training this classifier with two or more examples of the same word (and their negative examples), use this classifier to predict the best word label that describes the current object the Cozmo robot is looking at.

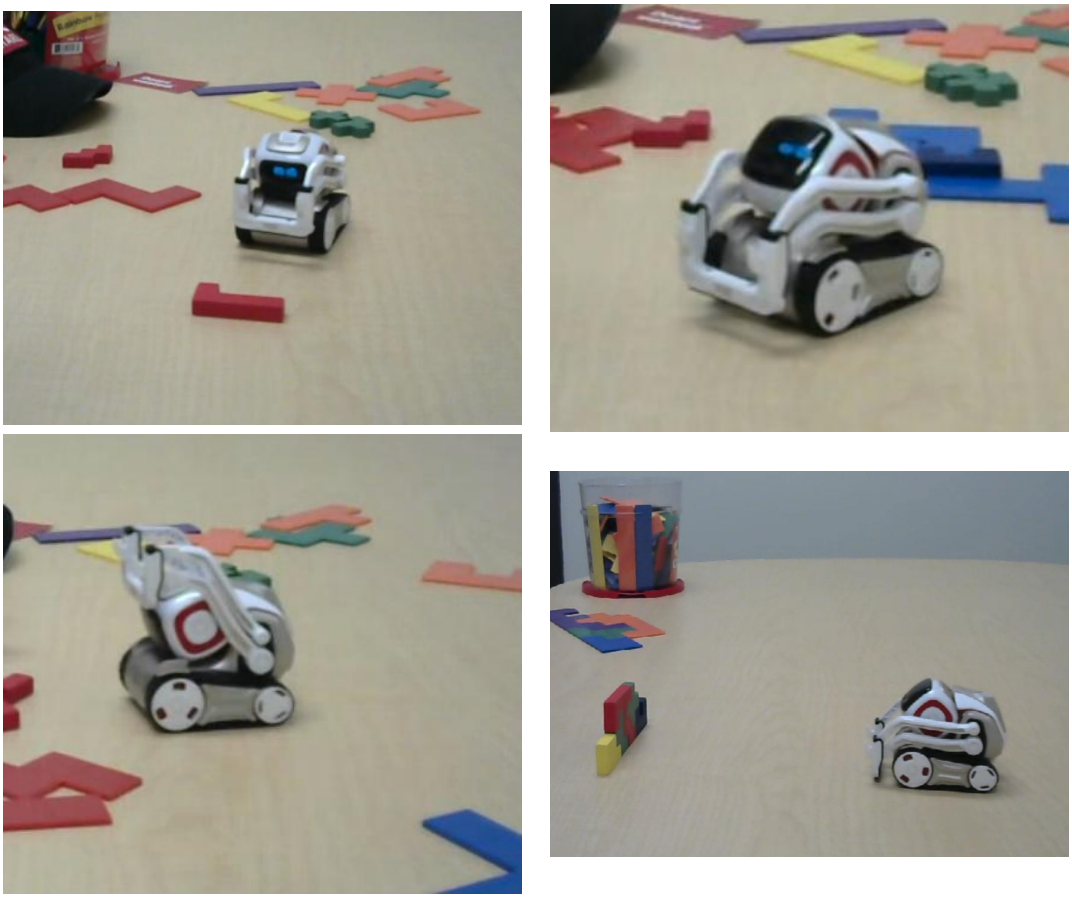

Figure 4.2: The four stages of an "Episode" for the Robot Actions module, moving clockwise and beginning from the top-left: 1) Task action: find object, go to object, or point to object; 2) Emotional display: emote understanding or confusion; 3) Word proposal: receive positive or negative feedback; 4) Locate human face.

CHAPTER 5

EXPERIMENT

In this chapter we explain the steps taken to conduct a live experiment with users and ascertain the influence of emotional displays in a language learning task between an embodied social robot and a casual human participant. We employ a within-group study design, meaning that each participant goes through the same study twice, one time in which the independent variable (i.e., with emotional display) is present, and again when it is absent (i.e., without emotional display). To mitigate learning effects, the order in which the test condition is presented is alternated.

Following [7], we investigate if a social robot's use of emotional displays in a language acquisition task mediates the robot's acquisition of new words. To test emotional displays, we conduct a live interactive study (explained in Section 3.1) between a Cozmo robot and a study participant. We use a Reinforcement Learning framework to enable Cozmo to learn which emotional display (understanding or confusion) is more useful depending on the user's most recent feedback to the robot. The goal of this study is to answer the question of whether we need to consider the user's perception of the robot's emotional state in future natural language tasks, or if this factor can be ignored.

5.1 Participant Recruitment

We recruited twenty-one study participants to interact with the Cozmo robot for two fifteen-minute periods over the course of a single session. Following each fifteen-minute interaction the participant is asked to answer every question of the same augmented Godspeed Questionnaire (found in the appendix) [4]. The Godspeed Questionnaire is 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. The entire study takes approximately one-hour; in exchange for their time participants are paid eight U.S. dollars.

Study participants are 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 are women; thirteen are men.

5.1.1 Study Setting

5.1.2 Study Task

First, the Cozmo robot is introduced to the participant, along with its affordances:

- Cozmo has a camera that can see them and the world.
- Cozmo has a microphone that can hear them.
- Cozmo doesn't know anything, but is "curious" to learn more about the world.
- 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. The entire study takes approximately one-hour; in exchange for their time participants are paid eight U.S. dollars.

- If Cozmo gets off-track, they are allowed to pick Cozmo up and move it around.
- When Cozmo is looking up, it is looking for their face.
- 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 faster.

The researcher 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 researcher was permitted to offer a constrained set of coaching tips to the participant during the interaction, if the participant needs 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 the interaction the user moves to the researcher's seat and completes the augmented Godspeed questionnaire on the researcher's laptop. We justify our use of the Godspeed Questionnaire from Weiss and Bartneck (2015) [37]. The browser is set to full-screen and the user is monitored by the researcher to ensure that the user only sees the survey. Following the completion of both interactions and subsequent surveys, the participant is paid eight dollars and signs a form acknowledging receipt of payment.

5.1.3 Hypothesis

We employed a "within-group" design because it more effectively controlled for individual differences in study participants (if one individual is more attentive / empathetic to the robot) and allowed us to recruit fewer study participants. For an involved participatory study like the one we conducted, these strengths outweigh any potential external learning effects, which can be easily controlled.

We hypothesize that if the robot produces emotional displays the user will use more words (i.e., be more engaged), in addition to producing more instances of positive or negative feedback; as a result, the robot will have more reliable classifiers in its Grounded Semantics model (i.e., it will have "learned" more words) to identify the objects it perceives using the pipeline described in the methods section (Visual Perception to Object Detection to Feature Extraction). The indicator of this improvement in the classifiers can be seen in the number of proposals the robot makes.

5.2 Evaluation

5.2.1 Language data

We evaluate the robot's performance in the task based on the number of words it hears from the participant, the number of word proposals it makes, the number of instances of positive feedback it hears, and the number of instances of negative feedback it hears.

5.2.2 Participant Surveys

We also evaluate the robot based on survey responses written by the study participants following the both trial sessions of the study. These task-specific questions are prepended to the standard Godspeed Questionnaire; the augmented Godspeed questionnaire is attached to appendix B:

- 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?

5.3 Results

Table 5.1 shows the results of the effect that emotional displays had on heard words, positive feedbacks, negative feedbacks, and proposals. 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. This is likely due to the valence of the emotional displays presented to the user.

Table 5.1: The effect of emotional displays on a language-acquisition task

[Mean values]	All trials	Without emotions	With emotions
Heard Words	62.3	58.0	**66.0**
Positive Feedbacks	13.3	11.8	**14.7**
Negative Feedbacks	6.7	**7.7**	6.0
Proposals	8.3	7.8	**8.9**

Exploring the effect of the positive or negative valence in Table 5.2 shows that there is a marked difference between those experimental trials in which a majority of the robot's emotional displays were either positive ("Mostly understanding") or negative ("Mostly confused"). This table reveals that each emotional display has its own, distinct impact on the language-acquisition task, perhaps modulated by the intensity of the display itself. Those trials in which the robot was deemed "Mostly confused" occurred when the total number of confused displays was higher than the median

(more than four confused displays), and the number of confused displays was also greater than the number of understanding displays (amounting to 10 trials in total). The trials in which the robot was deemed to be "Mostly understanding" occurred in those trials when the total number of understanding displays was greater than the median (more than two understanding displays), and the number of understanding displays was also greater than the number of confused displays (amounting to seven trials in total).

Table 5.2: The effect of valence on a language-acquisition task

[Mean values]	All emotion trials	Mostly confused	Mostly understanding
Heard Words	**66.0**	62.8	45.4
Positive Feedbacks	**14.7**	9.7	12.3
Negative Feedbacks	**6.0**	5.5	5.9
Proposals	8.9	**9.8**	8.7

Another important takeaway from Table 5.2 is that exhibiting "Mostly understanding" or "Mostly confused" emotional displays is not as effective a strategy as a more even split of positive- and negatively valenced emotions, as evidenced by Table 5.2. This supports the idea that the success of an emotional display is dependent on context, as is the case for any form of communication. In overcoming the cold-start problem, this context is informed by the success of the robot and human partner in the language-learning task.

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 [32] which 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, sense, and knowledge, as

seen in Figures 5.2, 5.3, and 5.4. In these figures, we interpret the total number of heard and proposed words by the robot as a proxy for the participant's overall level of success in the interaction (the participant's own estimation of their success as a teacher is a confounding factor outside the scope of this study to investigate). What these figures demonstrate is that participant success in the teacher role (i.e., interacting with a robot that heard more words from them and made more word proposals as a result) did not significantly effect their estimation of the robot's intelligence, sense, and knowledge.

Where a more successful interaction did influence participants was in their estimation of the robot's pleasantness, their attachment to the robot, how interesting the robot was, if they liked the robot, and if they would like to spend more time with the robot (Figures 5.5, 5.6, 5.7, 5.8, and 5.9). A participant's success in the teacher role (an increased number of heard and proposed words) appears to have a more significant effect on participants' positive estimations of the robot than the presence of emotional displays. This effect is aided, however, by the boost that emotional displays lends to the task itself, as can be seen by the higher peaks in the number of heard and proposed words for those users.

In our Reinforcement Learning module, the Q-Learning algorithm learned to weigh in favor of one emotional display to the exclusion of the other after only a few episodes, without consideration of *Robot Confidence*. This may have been due to the training batch size and training time for the Q-Learning algorithm being set significantly lower than the default values set by PyOpenDial (10 max samples and a 5 ms sample rate, instead of 3000 max samples and a 250 ms sample rate – the default values slowed training to the point of the model was making imperceptible progress). This was done in order to allow for live training of the model. In future work, we would want to

re-write the Q-Learning algorithm to be specifically tailored to this task, for efficiency, but also for readability and to allow for easier debugging.

Figure 5.1: The setting for the interaction. Top: Cozmo sees an object. Bottom: What the participant see upon entering the room.

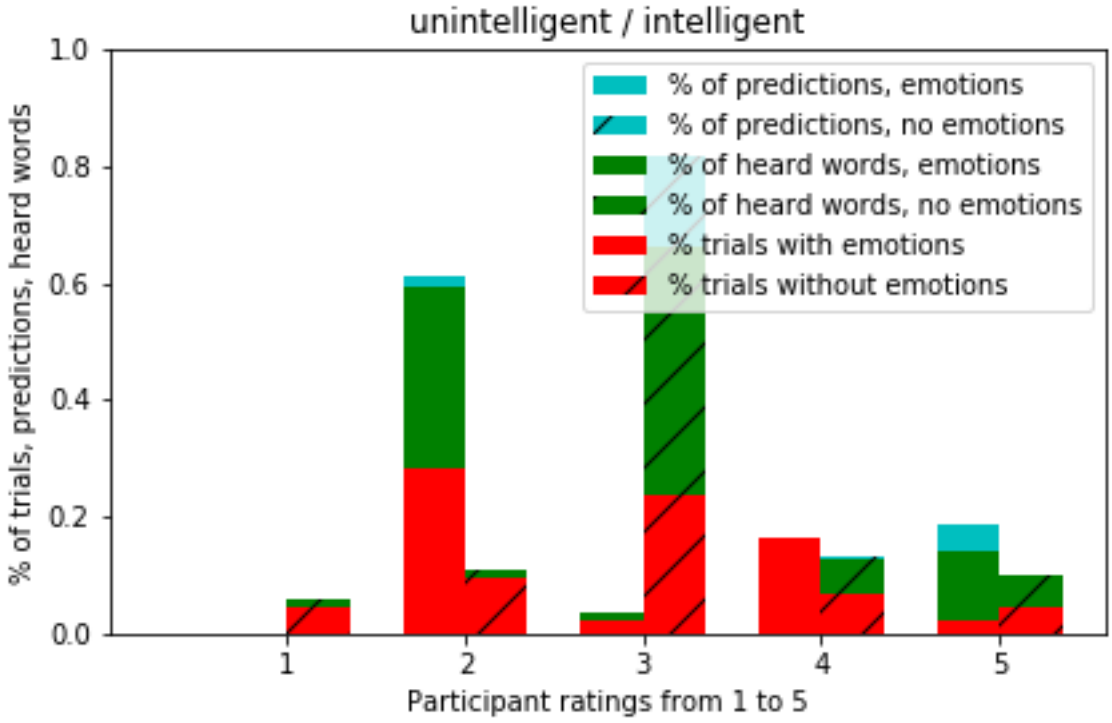

Figure 5.2: X-axis: Participant ratings from 1: unintelligent to 5: intelligent. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials.

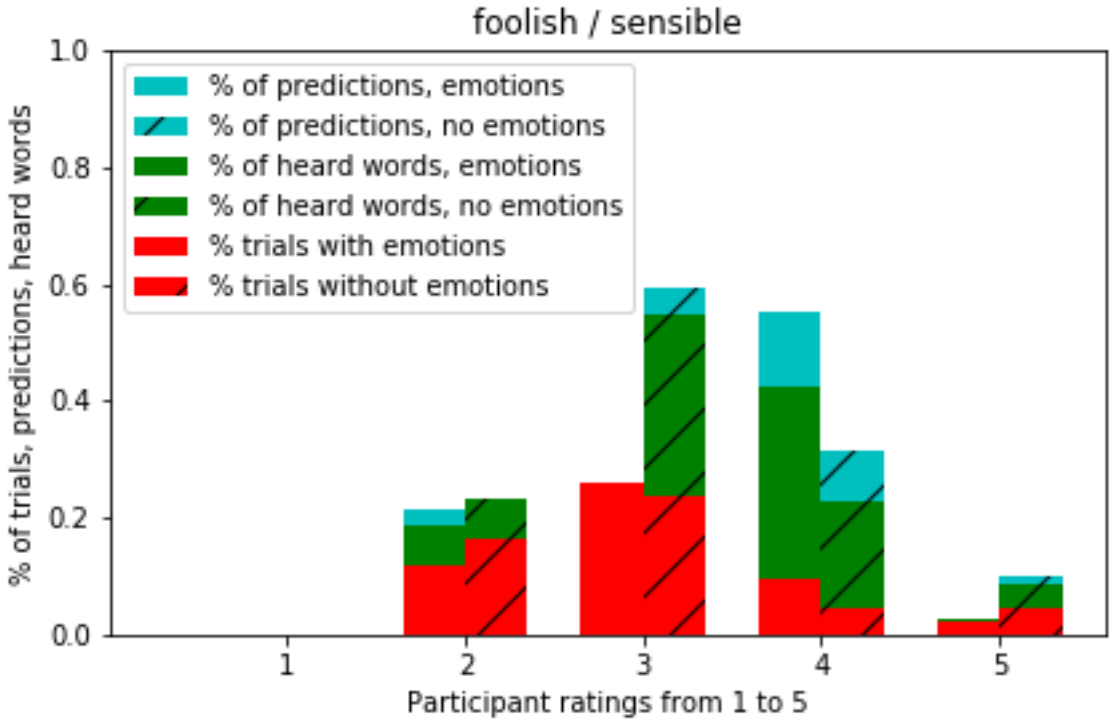

Figure 5.3: X-axis: Participant ratings from 1: foolish to 5: sensible. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials.

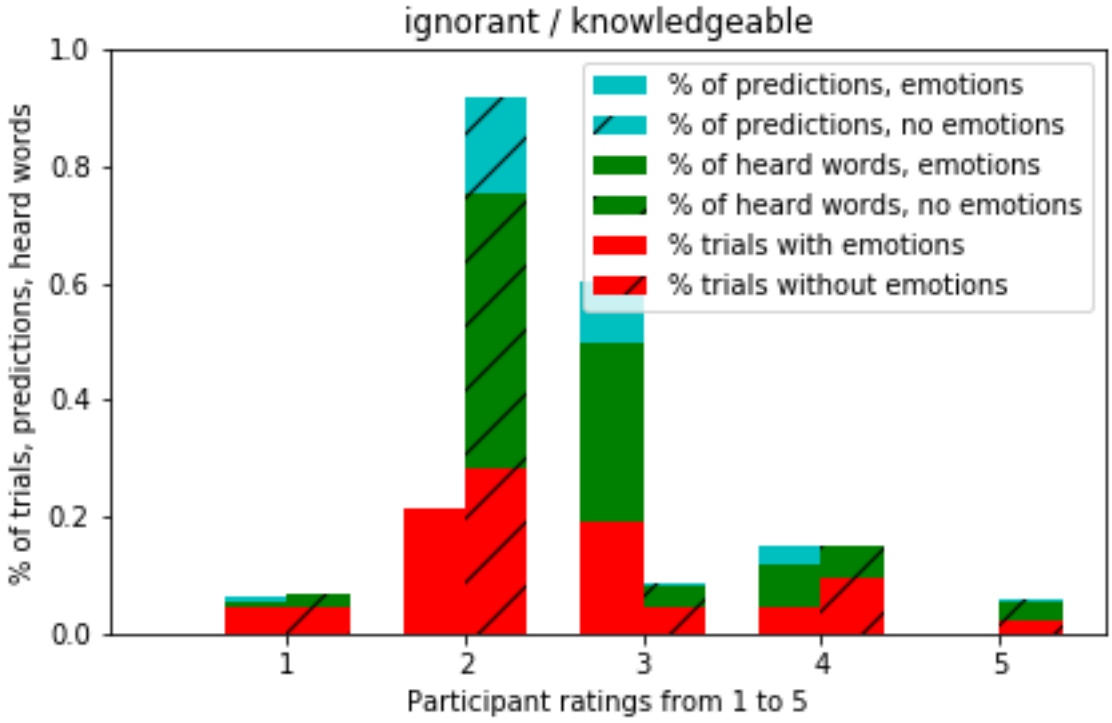

Figure 5.4: X-axis: Participant ratings from 1: ignorant to 5: knowledgeable. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials.

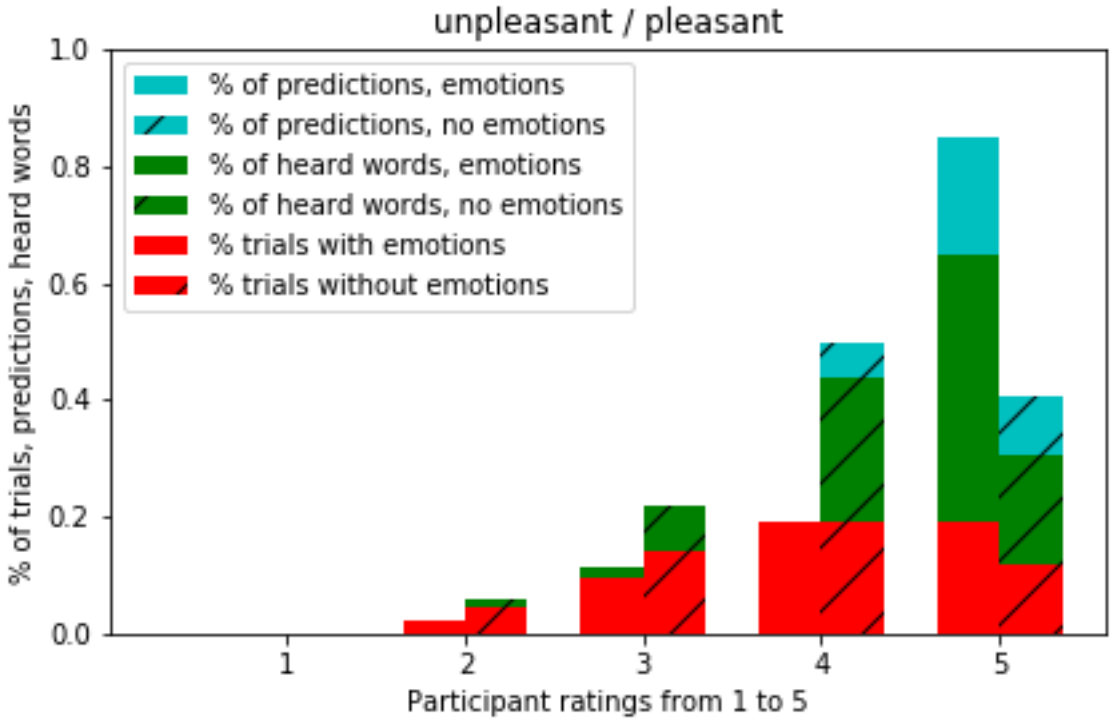

Figure 5.5: X-axis: Participant ratings from 1: unpleasant to 5: pleasant. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials..

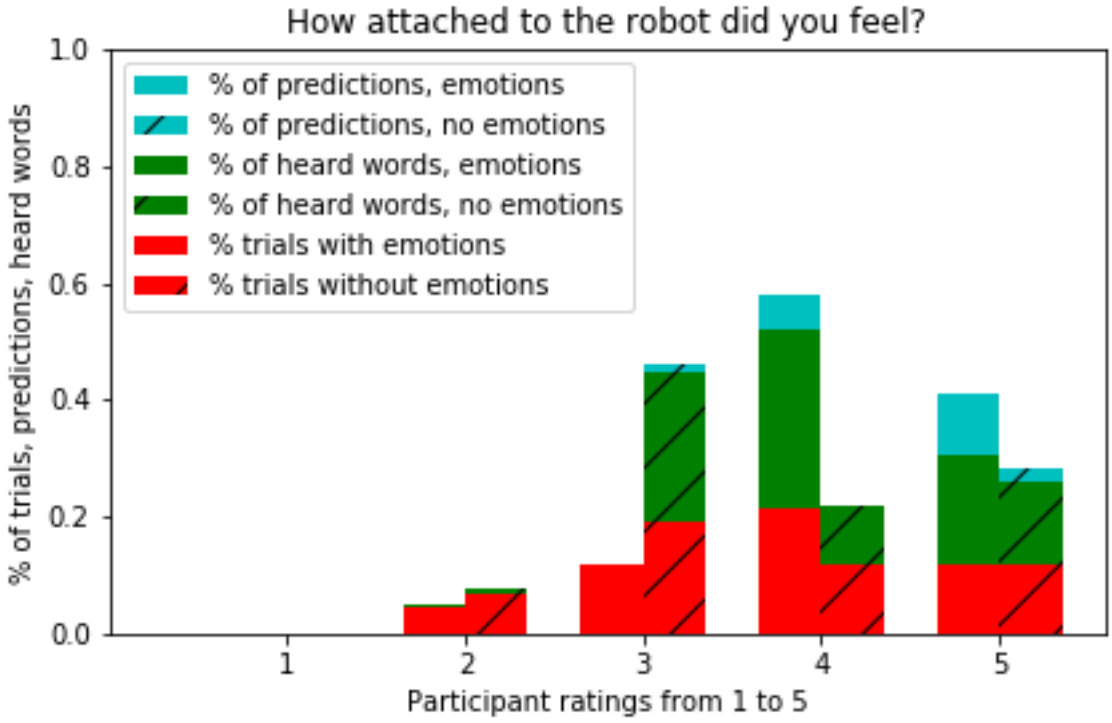

Figure 5.6: X-axis: Participant ratings from 1:not at all to 5:very much. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials.

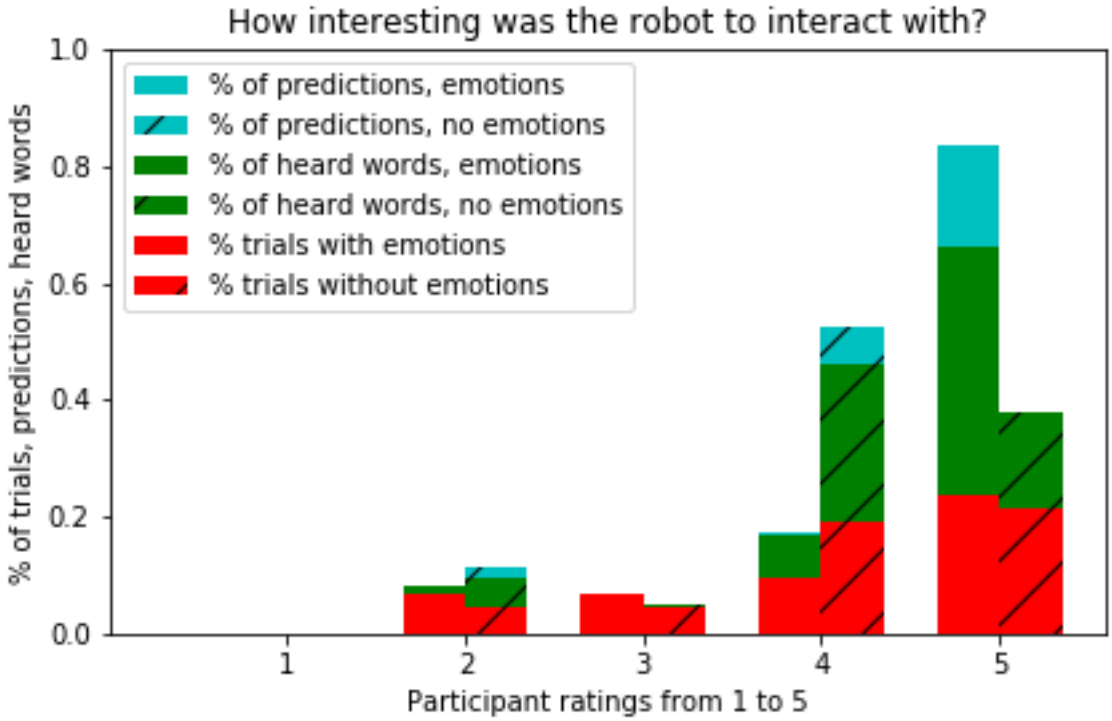

Figure 5.7: X-axis: Participant ratings from 1: not at all to 5: very much. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials.

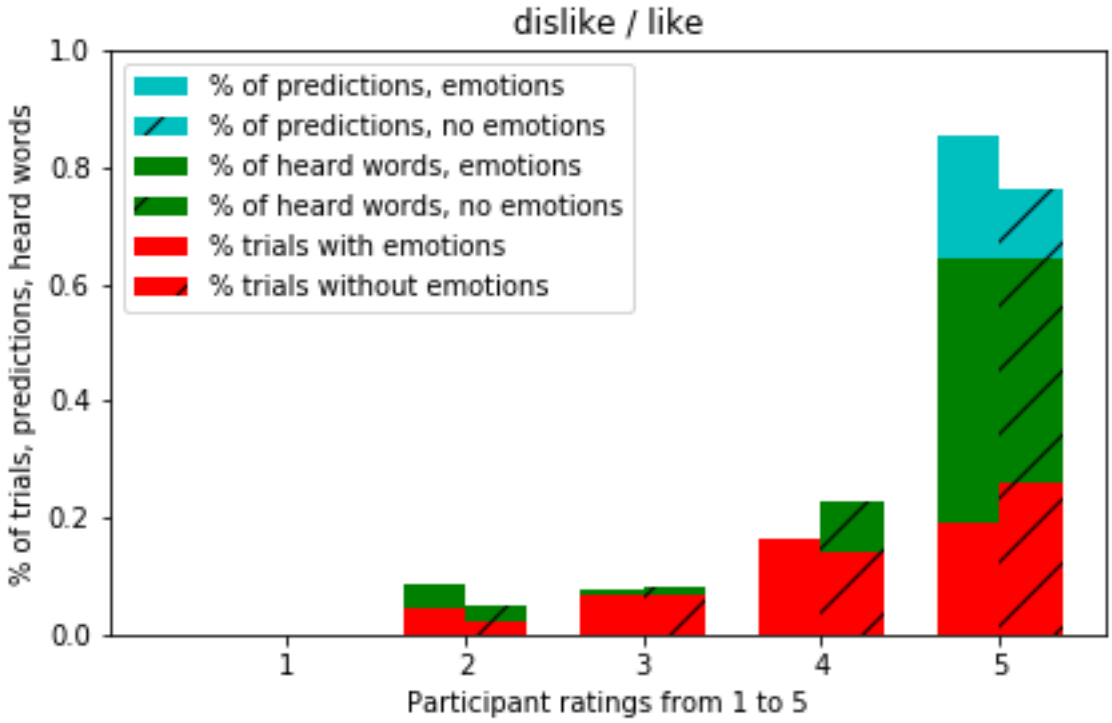

Figure 5.8: X-axis: Participant ratings from 1: dislike to 5: like. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials.

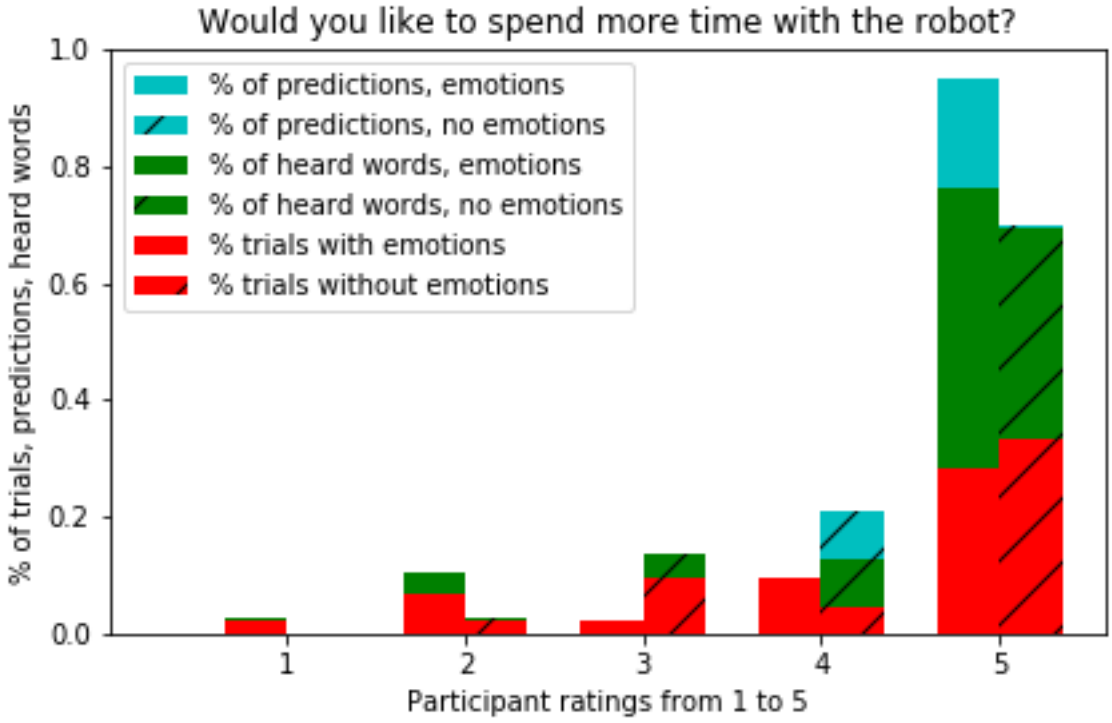

Figure 5.9: X-axis: Participant ratings from 1: not at all to 5: very much. Y-axis: the % of participants that selected those responses; the % of the robot's heard and proposed words for those trials.

CHAPTER 6

CONCLUSION

We recognized a problem in how an artificial system, such as a robot, could engage a casual human user in a language-learning task in a real-world, situated context with a cold-start requirement. Based on relevant background, we hypothesized a solution: emotional displays.

Having established that people do assign emotions to Cozmo's behaviors ([27]; see chapter 3 of this thesis), we focused on the valence pair model that was the most successful – confusion versus understanding – and also the pair of emotions most useful to demonstrate in a learning task, when a robot learner might need to communicate the state of the models that the user is training via emotional displays.

We take the lists of Understanding and Confused emotional displays created in our preliminary work. We conducted an experiment with twenty-one participants, who had to rely on the robot's movement and their own performance in a language acquisition task as context for what those emotional displays meant. 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

(orienting to objects; seeking out the user's face). This in turn led to the robot making more word proposals, which consequently led to greater engagement and more positive estimations of the robot on the part of the participant in the interaction (without leading to over-estimations of the robot's language understanding).

These results show that the presence of emotional displays can assist a robot in overcoming the cold-start problem in a dialogue with a casual human user.

6.1 Limitations and Future work

In future work, we would like to test different policies for the reinforcement learning regime. These could include a curiosity measure that would reward the robot more for hearing novel words than for hearing words it deems that it already understands, according to the accuracy of its Grounded Semantics classifiers. Additionally, this would incorporate an analysis of the number of different words the robot hears, and a measurement of how important repetition is on the part of the human teacher in the language learning task. Whether or not a user repeated the same word is indicative of their estimation of the robot's sophistication, and how over-estimations of robot language understanding could lead to fewer repetitions and weaker word comprehension.

Another aspect that demands further investigation would be the timing of emotional displays in the language learning interaction. Rather than having the robot naively produce animations at the same time in every episode, we could have the Reinforcement Learning regime take the previous task-action as input and make a decision as to when is the best time to make an emotional display (in addition to deciding the appropriate valence of emotion).

Further analysis of the exact effect of the intensity of the emotion on the success of the language-acquisition task would be enlightening. What is the difference between a very negative emotional display and a slightly negative (or closer to ambiguous) emotional display? Not every user interpreted "confusion" as confusion – many in fact, saw it as either the robot being upset with itself or them, and this lead to higher reported scores for anxiety in those trials in which the robot randomly selected those more intense animations for its "confused" displays.

We would also like to directly evaluate the Grounded Semantics classifiers that are trained by participants during their trials on novel data, as a means of better estimating the success of the language learning task.

In future work, we will survey participants to understand their own beliefs regarding their success as a teacher, and how this self-belief influences their estimation of the robot and their interaction with it as a whole.

REFERENCES

[1] Ralph Adolphs. The neurobiology of social cognition. *Current Opinion in Neurobiology*, 11(2):231–239, apr 2001.

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

[3] Lisa Feldman Barrett. Variety is the spice of life: A psychological construction approach to understanding variability in emotion, 2009.

[4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots, jan 2009.

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

[6] Mason Bretan, Guy Hoffman, and Gil Weinberg. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human Computer Studies*, 2015.

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

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

[9] Herbert H. Clark. *Using Language*. Cambridge University Press, may 1996.

[10] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 1960.

[11] Ezequiel Alejandro Di Paolo and Hanne De Jaegher. The interactive brain hypothesis. *Frontiers in Human Neuroscience*, 6:163, jun 2012.

[12] Emmanuel Ferreira and Fabrice Lefèvre. Reinforcement-learning based dialogue system for human-robot interactions with socially-inspired rewards. *Computer Speech and Language*, 34(1):256–274, nov 2015.

[13] C. J. Fillmore. *Pragmatics and the Description of Discourse*. Radical Pragmatics, 1981.

[14] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. *arXiv*, mar 2018.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pages 2980–2988. Institute of Electrical and Electronics Engineers Inc., dec 2017.

[16] Ruud Hortensius, Felix Hekele, and Emily S. Cross. The Perception of Emotion in Artificial Agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):852–864, dec 2018.

[17] E.J. Jacobs, J. Broekens, and C.M. Jonker. Joy, Distress, Hope, and Fear in Reinforcement Learning (Extended Abstract). *AAMAS 2014: Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems, Paris, France, 5-9 May 2014*, 2014.

[18] Youngsoo Jang, Jongmin Lee, Jaeyoung Park, Kyeng-Hun Lee, Pierre Lison, and Kee-Eung Kim. Pyopendial: A python-based domain-independent toolkit for developing spoken dialogue systems with probabilistic rules. In *PyOpenDial: A Python-based Domain-Independent Toolkit for Developing Spoken Dialogue Systems with Probabilistic Rules*, pages 187–192, 01 2019.

[19] Jason D. Williams. Integrating expert knowledge into POMDP optimization for spoken dialogue. In *Integrating expert knowledge into POMDP optimization for spoken dialogue*. Proceedings of the AAAI-08 Workshop on Advancements in POMDP Solvers, 2008.

[20] Malte F. Jung. Affective Grounding in Human-Robot Interaction. In *ACM/IEEE International Conference on Human-Robot Interaction*, volume Part F127194, 2017.

[21] Casey Kennington and David Schlangen. 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, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics.

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

[23] Ada S. Kim, Simran Bhatia, Elin A. Björling, and Dong Li. Designing a Collaborative Virtual Reality Game for Teen-Robot Interactions. In *Proceedings of the Interaction Design and Children on ZZZ - IDC '19*, pages 470–475, New York, New York, USA, 2019. ACM Press.

[24] Jacqueline M. Kory-Westlund and Cynthia Breazeal. Assessing Children's Perceptions and Acceptance of a Social Robot. In *Proceedings of the Interaction Design and Children on ZZZ - IDC '19*, pages 38–50, New York, New York, USA, 2019. ACM Press.

[25] Matthew Lewis and Lola Cañamero. Hedonic quality or reward? A study of basic pleasure in homeostasis and decision making of a motivated autonomous robot. *Adaptive Behavior*, 24(5), 2016.

[26] John Lones, Matthew Lewis, and Lola Cañamero. From sensorimotor experiences to cognitive development: Investigating the influence of experiential diversity on the development of an epigenetic robot. *Frontiers Robotics AI*, 3(AUG), aug 2016.

[27] David McNeill and Casey Kennington. Predicting Human Interpretations of Affect and Valence in a Social Robot. *RSS*, apr 2019.

[28] Joseph E. Michaelis and Bilge Mutlu. 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, 2019. ACM Press.

[29] P Ian Newcombe, Cale Campbell, Paul D Siakaluk, and Penny M Pexman. Effects of emotional and sensorimotor knowledge in semantic processing of concrete and abstract nouns. *Frontiers in human neuroscience*, 6:275, 2012.

[30] Donald A. Norman. *Design of Everyday Things, Revised edition.* Basic Books, 1990.

[31] Jekaterina Novikova, Leon Watts, and Tetsunari Inamura. Emotionally expressive robot behavior improves human-robot collaboration. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 7–12. IEEE, aug 2015.

[32] Sarah Plane, Ariel Marvasti, Tyler Egan, and Casey Kennington. Predicting Perceived Age: Both Language Ability and Appearance are Important. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 130–139, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics.

[33] David L. Robinson. Brain function, emotional experience and personality. *Netherlands Journal of Psychology*, 2008.

[34] Astrid M. Rosenthal-von der Pütten, Nicole C. Krämer, and Jonathan Herrmann. The Effects of Humanlike and Robot-Specific Affective Nonverbal Behavior on Perception, Emotion, and Behavior. *International Journal of Social Robotics*, 10(5):569–582, nov 2018.

[35] Chad Spiegel and Justin Halberda. Rapid fast-mapping abilities in 2-year-olds. *Journal of Experimental Child Psychology*, 109(1):132–140, may 2011.

[36] Evan Thompson. *Mind in life : biology, phenomenology, and the sciences of mind*. Belknap Press of Harvard University Press, 2007.

[37] Astrid Weiss and Christoph Bartneck. Meta analysis of the usage of the Godspeed Questionnaire Series. In *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, volume 2015-November, pages 381–388. Institute of Electrical and Electronics Engineers Inc., nov 2015.

APPENDIX A

PARSING FEEDBACK LISTS

Here is Appendix A. See Appendix C for a list of the Cozmo's internal features used to model its emotional displays.

positive feedback = ['yes','yeah','yep','right','correct','good','nice']

negative feedback = ['no','nope','wrong','incorrect','stop','bad','not']

APPENDIX B

AUGMENTED GODSPEED QUESTIONNAIRE

Questionnaire

1. How attached to the robot did you feel? Mark only one oval.

Not at all 1 2 3 4 5 Very

2. How interesting was the robot to interact with? Mark only one oval.

Not at all 1 2 3 4 5 Very

3. Would you like to spend more time with the robot? Mark only one oval.

Not at all 1 2 3 4 5 Very much

4. Read the statement below and select one of the given options: The robot had a goal. Mark only one oval.

Yes No

5. If you agreed, what do you think the robot's goal was? Why do you think that?

6. If you disagreed, why do you disagree? What do you think robot was doing?

7. How many years old do you think the robot is (in terms of its behavior)?

8. Please rate your impression of the robot on this scale: Mark only one oval.

Fake 1 2 3 4 5 Natural

9. Please rate your impression of the robot on this scale: Mark only one oval.

Machinelike 1 2 3 4 5 Humanlike

10. Please rate your impression of the robot on this scale: Mark only one oval.

Unconscious 1 2 3 4 5 Conscious

11. Please rate your impression of the robot on this scale: Mark only one oval.

Artificial 1 2 3 4 5 Lifelike

12. Please rate your impression of the robot on this scale: Mark only one oval.

Moving rigidly 1 2 3 4 5 Moving elegantly

13. Please rate your impression of the robot on this scale: Mark only one oval.

Dead 1 2 3 4 5 Alive

14. Please rate your impression of the robot on this scale: Mark only one oval.

Stagnant 1 2 3 4 5 Lively

15. Please rate your impression of the robot on this scale: Mark only one oval.

Mechanical 1 2 3 4 5 Organic

16. Please rate your impression of the robot on this scale: Mark only one oval.

Inert 1 2 3 4 5 Interactive

17. Please rate your impression of the robot on this scale: Mark only one oval.

Apathetic 1 2 3 4 5 Responsive

18. Please rate your impression of the robot on this scale: Mark only one oval.

Dislike 1 2 3 4 5 Like

19. Please rate your impression of the robot on this scale: Mark only one oval.

Unfriendly 1 2 3 4 5 Friendly

20. Please rate your impression of the robot on this scale: Mark only one oval.

Unkind 1 2 3 4 5 Kind

21. Please rate your impression of the robot on this scale: Mark only one oval.

Unpleasant 1 2 3 4 5 Pleasant

22. Please rate your impression of the robot on this scale: Mark only one oval.

Awful 1 2 3 4 5 Nice

23. Please rate your impression of the robot on this scale: Mark only one oval.

Incompetent 1 2 3 4 5 Competent

24. Please rate your impression of the robot on this scale: Mark only one oval.

Ignorant 1 2 3 4 5 Knowledgable

25. Please rate your impression of the robot on this scale: Mark only one oval.

Irresponsible 1 2 3 4 5 Responsible

26. Please rate your impression of the robot on this scale: Mark only one oval.

Unintelligent 1 2 3 4 5 Intelligent

27. Please rate your impression of the robot on this scale: Mark only one oval.

Foolish 1 2 3 4 5 Sensible

28. At the BEGINNING of the interaction, how did you feel on this scale: Mark only one oval.

Anxious 1 2 3 4 5 Relaxed

29. At the END of the interaction, how did you feel on this scale: Mark only one oval.

Anxious 1 2 3 4 5 Relaxed

30. At the BEGINNING of the interaction, how did you feel on this scale: Mark only one oval.

Agitated 1 2 3 4 5 Calm

31. At the END of the interaction, how did you feel on this scale: Mark only one oval.

Agitated 1 2 3 4 5 Calm

32. At the BEGINNING of the interaction, how did you feel on this scale: Mark only one oval.

Bored 1 2 3 4 5 Interested

33. At the END of the interaction, how did you feel on this scale: Mark only one oval.

Bored 1 2 3 4 5 Interested

34. Of the following relations, which do you feel describe the robot best? Mark only one oval.

Brother or Sister

Classmate

Stranger

Relative (e.g., cousin or aunt)

Friend

Parent

Teacher

Neighbor

APPENDIX C

COZMO INTERNAL FEATURES

Here is Appendix C.

left_wheel_speed

right_wheel_speed

battery_voltage

time

nav_memory_map_sizes

nav_memory_map_x

nav_memory_map_y

pose_0

pose_1

pose_4

pose_5

pose_10

pose_12

pose_13

pose_14

is_moving

is_picked_up

is_animating

lift_in_pos

head_in_pos

are_wheels_moving

is_localized

pose_angle_rads

pose_angle_degs

pose_angle_abs_rads

pose_angle_abs_degs

pose_pitch_rads

pose_pitch_degs

pose_pitch_abs_rads

pose_pitch_abs_degs

head_angle_rads

head_angle_degs

head_angle_abs_rads

head_angle_abs_degs

lift_position_height

lift_position_ratio

lift_position_angle_rads

lift_position_angle_degs

lift_position_angle_abs_rads

lift_position_angle_abs_degs

dispatcher_has_in_progress_action

accelerometer_x

accelerometer_y

accelerometer_z

gyro_x

gyro_y

gyro_z