# Predicting Perceived Age: Both Language Ability and Appearance are Important

**Sarah Plane, Ariel Marvasti, Tyler Egan** and **Casey Kennington**
Speech, Language & Interactive Machines Group
Boise State University
Boise, Idaho, U.S.A.
firstnamelastname@boisestate.edu

## Abstract

When interacting with robots in a situated spoken dialogue setting, human dialogue partners tend to assign anthropomorphic and social characteristics to those robots. In this paper, we explore the age and educational level that human dialogue partners assign to three different robotic systems, including an un-embodied spoken dialogue system. We found that how a robot speaks is as important to human perceptions as the way the robot looks. Using the data from our experiment, we derived prosodic, emotional, and linguistic features from the participants to train and evaluate a classifier that predicts perceived intelligence, age, and education level.

## 1 Introduction

Co-located, face-to-face spoken dialogue is the primary and basic setting where humans learn their first language (Fillmore, 1981) partly because dialogue participants (i.e., caregiver and child) can denote objects in their shared environment which is an important developmental step in child language acquisition (McCune, 2008). This setting motivates human-robot interaction tasks where robots acquire semantic meanings of words, and where part of the semantic representation of those words is *grounded* (Harnad, 1990) somehow in the physical world (e.g., the semantics of the word *red* is grounded in perception of color vision). Language grounding for robots has received increased attention (Bansal et al., 2017) and language learning is an essential aspect to robots that learn about their environment and how to interact naturally with humans.

However, humans who interact with robots often assign anthropomorphic characteristics to robots depending on how they perceive those robots; for example stereotypical gender (Eyssel and Hegel, 2012), social categorizations (Eyssel and Kuchenbrandt, 2012) stereotypical roles (Tay et al., 2014), as well as intelligence, interpretability, and sympathy (Novikova et al., 2017). This has implications for the kinds of tasks that we ask our robots to do and the settings in which robots perform those tasks, including tasks where language grounding and acquisition is either a direct or indirect goal. It is important not to assume that humans will perceive the robot in the "correct" way; rather, the age and academic level appropriateness needs to be monitored, particularly in a grounding and first-language acquisition task. The obvious follow-up question here is: *Do robots need to acquire language as human children do?* Certainly, enough functional systems exist that show how language can be acquired in many ways. The motivation here, however, is that those systems could be missing something in the language acquisition process that children receive because of the way they are perceived by human dialogue partners. We cannot tell until we have a robot that is shown as being perceived as a child (current work) and use that robot for language learning tasks (future work).

We hypothesize in this paper that how a robot looks and acts will not only affect how humans perceive that robot's intelligence, but it will also affect how humans perceive that robot's age and academic level. In particular, we explore how humans perceive three different systems: two embodied robots, and one a spoken dialogue system (explained in Section 3). We show through an experiment that human perception of robots, particularly in how they perceive the robots' intelligence, age, and academic level, is due to how the robot appears, but also due to how the robot uses speech to interact.



*Proceedings of the SIGDIAL 2018 Conference*, pages 130–139,
Melbourne, Australia, 12-14 July 2018. 

## 2 Related Work

Several areas of research play into this work including seminal (Roy and Reiter, 2005) and recent work in grounded semantic learning in various tasks and settings, notably learning descriptions of the immediate environment (Walter et al., 2014); navigation (Kollar et al., 2010); nouns, adjectives, and relational spatial descriptions (Kennington and Schlangen, 2015); spatial operations (Bisk et al., 2018), and verbs (She and Chai, 2016). Previous work has also focused on multimodal aspects of human-robot interaction, including grounded semantics (Thomason et al., 2016), engagement (Bohus and Horvitz, 2009), and establishing common ground (Chai et al., 2014). Others have explored how robots are perceived differently by different human age groups such as the elderly (Kiela et al., 2015), whereas we are focused on the perceived age of the robot by human dialogue partners. Moreover, though we do not design our robots for deliberate affective grounding (i.e., the coordination effect of building common understanding of what behaviors can be exhibited, and how beahvior is interpreted emotionally) as in Jung (2017), we hypothesize that how our robots behave effects how they are perceived.

Kiela et al. (2015) compared tutoring sequences in parent-child and human-robot interactions with varying verbal and demonstrative behaviors, and Lyon et al. (2016) brought together several areas of research relating to language acquisition in robotics. We differ from this previous work in that we do not explcitely tell our participants to interact with the robots as they would a child, effectively removing the assumption that participants will treat robots in an age-appropriate way. Another important difference to their work is that we opted not to use an anthropomorphically realistic child robot because such robots often make people feel uncomfortable (Eberle, 2009). Our work is similar in some ways to, but different from work in paralinguistics where recognition of age given linguistic features is a common task (Schuller et al., 2013) in that we are make use of exra-linguistic features. Our work primarily builds off of Novikova et al. (2017) who used multimodal features derived from the human participants to predict perceived likability and intelligence of a robot. We use similar multimodal features to predict the perceived age and academic level. An important difference to their work is that we designed the experiment with three robots to vary the appearance and two language settings to vary the behavior and linguistic factors of the robots.


Figure 1: The two physical robots in our study: KOBUKI with a mounted MS Kinect and COZMO.

## 3 Experiment

The primary goal of our experiment is to determine what factors play into how humans perceive a robot's age and academic level. We used the three following robotic systems in our experiment:

- Kobuki Base Robot with a Microsoft Kinect on top (denoted as KOBUKI)
- Anki Cozmo robot (denoted as COZMO)
- Non-physical "robot" (i.e., a non-embodied spoken dialogue system) which was just a camera and speaker (denoted as SDS)

**Robot Appearance** The COZMO has a head and animated eyes and is noticeably smaller than the KOBUKI. The robots did not move during the experiments, though they were clearly activated (e.g., the KOBUKI had a small light and COZMO's eyes were visible and moved at random intervals, which is the default setting). Figure 1 shows the KOBUKI and COZMO robots as seen by the participants. We chose these three robots because they were available to us and we assume that, based solely on appearance, participants would perceive the robots differently. We chose a spoken dialogue system (SDS) as one of the "robots" because we wanted to explore how participants perceive a system that is unembodied in direct comparison to embodied systems.

**Robot Behavior** The COZMO robot has a built-in speaker with a young-sounding synthetic voice. We used two adult voices for the KOBUKI and SDS robots from the Amazon Polly system (the Joey and Joanna voices) which we played on a small speaker.[1] We vary the language setting of the robots by assigning each robot one of two possible settings: *high* and *low*. In the high setting,

[1] https://aws.amazon.com/polly/

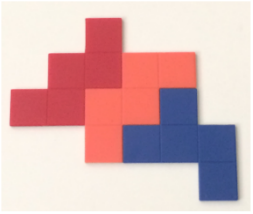



Figure 2: Example puzzle made up of three colored pentomino tiles with a specified name.

the following responses were possible: *sure*; *okay*; *yeah*; *oh*; *I see*; *uh huh*; ___ (where the robot repeats a word spoken by the participant) and any combination of those responses in a single uttered response; and for the low setting, the following responses were possible: *yes*; *okay*; *uh*; ___ (where the robot repeats a word spoken by the participant). In the high setting, the robot would produce responses more often than in the low setting. These responses are characteristic of different levels of *feedback*; the high setting contains feedback strategies that signaled understanding to the participant, whereas the low setting only signaled phonetic receipt. This corresponds to previous work (Stubbe, 1998) which investigated various feedback strategies employed in human-human dialogue termed *neutral minimal responses* (corresponding to our low setting) and *supportive minimal responses* (corresponding to our high setting).

With this setup, there are 6 possible settings: high and low for each of the three robots. Our hypothesis is that participants will perceive KOBUKI as older and more intelligent than COZMO overall (in both high and low settings) despite being less anthropomorphic, perceive COZMO as very young in the low setting, and that SDS will be perceived as older than COZMO in the high setting, but similar to COZMO in the low setting.

### 3.1 Task and Participants

The experimenter gave each participant consent and instruction forms to complete before the experiment. The participant was then given three colored pentomino puzzle tiles and a sheet of paper with three goal shapes (example in Figure 2), each composed from the corresponding tiles. The experimenter instructed the participant to sit at a table where they would see a robot. Their task was to explain to the robot how to use the tiles to construct the three goal shapes and tell the robot the name of each shape. The experimenter did not specify how to accomplish this task or give examples of the kinds of things that the robot might understand. The experimenter then left the room, leaving the participant with the robot to complete the task. The robots only responded verbally in the *low/high* setting as explained above and their responses were controlled by the experimenter (i.e., in a Wizard-of-Oz paradigm). The robots produced no physical movement. When the participant completed each task, they uttered a keyword (i.e., *done*), then the experimenter returned and administered a questionnaire. This process was followed for each of the three robots.

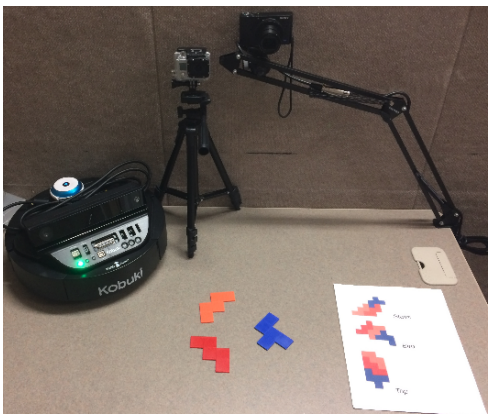
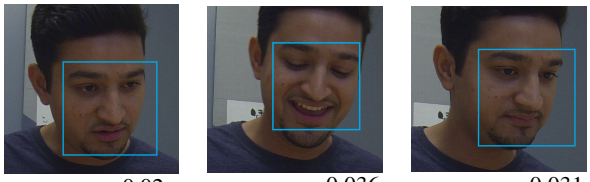



Figure 3: Example setting using the KOBUKI robot. The participants were to show the robot how to construct the three goal objects on the paper using the corresponding physical tiles. The additional cameras were used to record audio and video of the participant.

The following aspects of the experiment were randomly assigned to each participant: the order of robot presentation, the puzzle tiles and corresponding goal shapes for each robot, the language setting (i.e., *high* or *low*) which remained the same for all three robot interactions for each participant, and for KOBUKI and SDS the adult voice (either Joey or Joanna). We recruited 21 English-speaking participants (10 Female, 11 Male), most of whom were students of Boise State University. The interaction generally took about 30 minutes; participants received $5 for their participation.

### 3.2 Data Collection

We recorded the interactions with a camera that captured the face and a microphone that captured the speech of each participant. We automatically transcribed the speech using the Google Speech API (we manually checked an accented female



| | | | | | |
|---|---|---|---|---|---|
| anger | 0.02 | anger | 0.036 | anger | 0.031 |
| contempt | 0.009 | contempt | 0.019 | contempt | 0.182 |
| disgust | 0.009 | disgust | 0.018 | disgust | 0.04 |
| fear | 0.002 | fear | 0.005 | fear | 0.000 |
| happiness | 0.005 | happiness | 0.831 | happiness | 0.066 |
| neutral | 0.909 | neutral | 0.006 | neutral | 0.649 |
| sadness | 0.03 | sadness | 0.093 | sadness | 0.027 |
| surprise | 0.011 | surprise | 0.006 | surprise | 0.002 |

Figure 4: Examples of results as processed by the MS Emotions API.

voice which achieved an estimated WER of 30.0) and segmented transcriptions into sentences after 1 second of detected silence, which is a longer pause duration than the average pause duration for adult-adult conversation (though adults tend to take longer pauses when interacting with children (DePaulo and Coleman, 1986)). This resulted in video, audio, and transcriptions for each participant, for each robot interaction. We also collected 58 questionnaires (we had to remove several because they were missing data; i.e., some participants did not answer some of the questionnaire questions), one for each robot interaction, from each participant.

## 4 Data Analysis

Using the data collected from the experiment, we derived subjective measures from the questionnaires and we derived a number of objective measures from the video, audio, and transcriptions. In this section, we explain what methods we used to derive and analyze those measures.

**Emotion Features** Using the video feed of the participants, we extracted an image of the participants' faces every 5 seconds. We used the Microsoft Emotion API for processing these images to calculate an average distribution over 8 possible emotion categories for each image: *happiness, sadness, surprise, anger, fear, contempt, disgust,* and *neutral*. Figure 4 shows an example of face snapshots taken from the video in the task setting and the corresponding distributions over the emotions as produced by the API.

**Prosodic Features** From the audio, we calculated the average fundamental frequency of speech (F0) of the participant over the entire interaction between the participant and the robot for each robot setting.

**Linguistic Features** Using the automatically transcribed text, we follow directly from Novikova et al. (2017) to derive several linguistic measures, with the exception that we did not derive dialogue-related features because, though our robots were engaging in a kind of dialogue with the participants, they weren't taking the floor in a dialogue turn; i.e., our robots were only providing feedback to signal either phonetic receipt or semantic understanding (*low* and *high* settings, respectively). We used the Lexical Complexity Analyser (Lu, 2009, 2012), which yields several measures, two of which we leverage here: lexical diversity (LD) and the mean segmented type-token ratio (MSTTR), both of which measure diversity of tokens; the latter averaging the diversity over segments of a given length (for all measures, higher values denote more respective diversity and sophistication in the measured text). The Complexity Analyser also produces a lexical sophistication (LS) measure, also known as lexical rareness which is the proportion of lexical word types that are not common (i.e., not the 2,000 most frequent words in the British National Corpus).

For syntactic variation, we applied the D-Level Analyser (Lu, 2009) using the D-Level scale (Lu, 2014). This tool builds off of the Stanford Part-of-Speech Tagger (Toutanova and Manning, 2000) and the Collins Parser (Collins, 2003) and produces a scaled analysis. The D-Level scale counts utterances belonging to one of 8 levels (Levels 0-7), where lower levels such as 0-1 include simple or incomplete sentences; the higher the level, the more complex the syntactic structure. We report each of these levels along with a mean level.

**Godspeed Questionnaire** We used the Godspeed Questionnaire (Bartneck et al., 2009) which consists of 21 pairs of contrasting characteristics in areas of anthropomorphism (e.g., *artificial* vs. *lifelike*), likability (e.g., *unfriendly* vs. *friendly*), intelligence (e.g., *incompetent* vs. *competent*), and interpretabilitiy (e.g., *confusing* vs. *clear*) each with a 5-point scaling between them. In addition to those questions, we included the following:

- Have you ever interacted with a robot before participating in this study?
- If you could give the robot you interacted

with a human age, how old would you say it was?

- What level of education would be appropriate for the robot you interacted with?

For the question asking about human age, answers could be selected from a set of binned ranges (under 2 years, 2-5, 6-12, 13-17, 18-24, 25-34, 35 and older), and for the education question, answers could be selected from preschool, kindergarten, 1-12 (each grade could be selected), undergraduate, graduate, post-graduate.

### 4.1 Analysis

In this section, we analyze the results of the data for the emotional, prosodic, and linguistic measures. We also provide correlations between those measures and the Godspeed Questionnaire. At the end of this section, we provide a discussion of the overall analysis.

**Emotion Analysis** The most common emotional response as produced by the MS Emotions API was *neutral* for all settings, ranging from 73-87% (avg 81%). The next most common emotions were *happiness* (avg 11.1%), *sadness* (avg 3.7%), *surprise* (2%), and *contempt* (avg 1%). We show in Figure 5 the average distribution over those four emotions for all of our settings. All other emotions were negligible.

**Prosodic Analysis** Table 1 shows the the average F0 scores for each setting. In general, in the low linguistic setting participants averaged a higher F0 across all robots. This was the case also for individual robots. By a wide margin, COZMO averaged a higher F0 than the other two robots under both low and high settings.

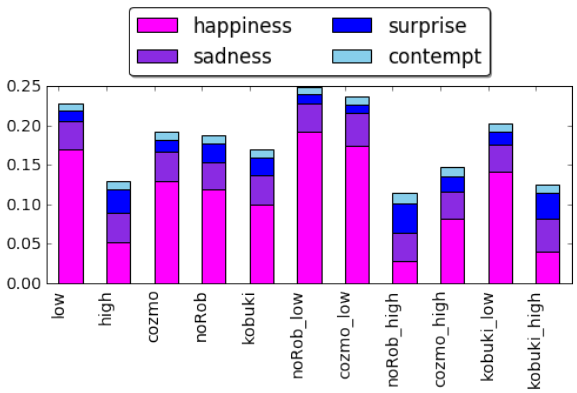



Figure 5: Average emotion (happiness, sadness, surprise, contempt) values for all settings.

| | cozmo | kobuki | noRob | all |
|---|---|---|---|---|
| **low** | 173.39 | 164.32 | 158.49 | 164.32 |
| **high** | 166.86 | 153.18 | 153.15 | 157.73 |
| **both** | 170.28 | 157.32 | 155 | |

Table 1: Prosodic analysis: average F0 values for each setting.

| setting | LD | LS | MSTTR |
|---|---|---|---|
| low | 0.45 | 0.32 | 0.62 |
| high | 0.48 | 0.34 | 0.64 |
| cozmo | 0.46 | 0.29 | 0.62 |
| noRob | 0.45 | 0.3 | 0.63 |
| kobuki | 0.46 | 0.28 | 0.63 |
| cozmo low | 0.46 | 0.23 | 0.61 |
| noRob low | 0.45 | 0.26 | 0.62 |
| kobuki low | 0.45 | 0.26 | 0.63 |
| cozmo high | 0.47 | 0.27 | 0.66 |
| noRob high | 0.47 | 0.27 | 0.63 |
| kobuki high | 0.49 | 0.23 | 0.64 |

Table 2: Linguistic analysis: LD, LS, and MSTTR values for all settings.

**Linguistic Analysis** Table 2 shows the results of the linguistic analysis. The LD (lexical diversity) scores show that, on average, participants used more LD in the high settings. Figure 6 shows the results of the D-Level analysis. Level0 (i.e., short utterances) was by far the most common level which accounted for 66% of all utterances for all participants. The second most common was Level7, the level representing the most complex types of utterances. This is no surprise, as Level7 accounts for longer utterances above some threshold; i.e., all utterances of a certain length and complexity or higher fit under Level7. The *low* setting had a Level7 value of 17%, and the *high* setting had a Level7 value of 11%. This may seem surprising, but it follows previous research which has shown that, when a speaker receives fewer responses, they draw out their turns, which result longer utterances (Stubbe, 1998).

**Questionnaire Analysis** We calculated (Spearman) correlations between the prosodic, emotional, and linguistic features, and the questionnaire responses with the *low/high* settings and the robot settings. Table 3 shows the results where the correlation had a strength of 0.5 or higher. Fig-

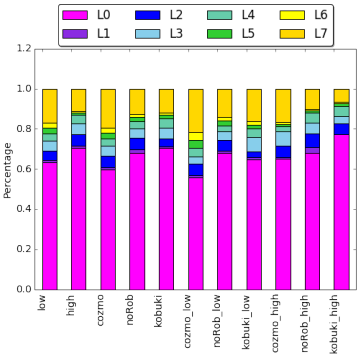



Figure 6: Percentage results for Level0-Level7 for all settings.

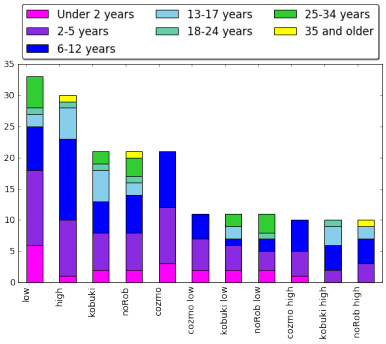



Figure 7: Questionnaire responses (raw scores) for ages.

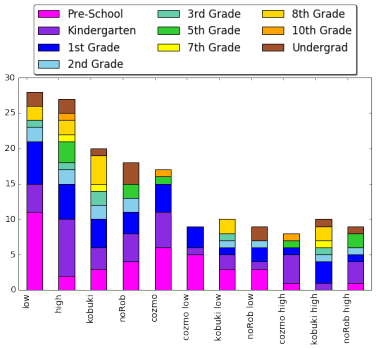



Figure 8: Questionnaire responses for academic grades.

ures 7 and 8 respectively show the age groups and academic years that the participants perceived for each robot in each setting. Overall, participants assigned low age and academic level to all robots when they produced feedback that did not signal semantic understanding (i.e., the low setting). They also assigned a lower age and academic level to COZMO for all settings (with the exception of one 10th grade assignment).

Our results confirm the Novikova et al. (2017) result which showed a strong correlation between F0 and *knowledgeable*. Interestingly, F0 only correlated knowledge with the physical robots and the SDS robot in the low setting. There is more to the F0 correlations: F0 in the low setting correlates with *conscious*, in the high setting correlates with *natural* and *human-like*, and in the COZMO robot setting with *lifelike*. There were some correlations with age and academic level: LS in the high setting correlated with the robot being perceived as age 18-24 and when interacting with COZMO, a higher F0 correlated with a perception of COZMO being 6-12 years old and in the 4th grade. Lexical diversity correlates with *sadness* and *contempt*, which indicates that participants use more diverse language (i.e., they continue speaking) when they are frustrated with the interaction (Stubbe, 1998); particularly in the high setting when they expect more from the robots. However, they increase their LS also in the high setting because they perceive the robot as more intelligent.

**Discussion** Taken together, the emotional, prosodic, and linguistic analyses show that participants treated the low setting with a higher average F0, less linguistic complexity, and a greater display of happiness in their facial emotions. This is useful knowledge: the way a robot speaks has an impact on the perception of that robot by the human users, regardless of whether or not that robot is embodied. Moreover, despite the fact that the robots only produced feedback as the only system behavior, the participants tended to assign a younger age and academic level to the COZMO robot. There were subtle differences in how the participants perceived the KOBUKI and SDS robots. In general, the participants seemed to perceive the SDS as being older and as having a higher academic level in the emotion, prosodic, and linguistic modalities, though those differences were small. This leads us to postulate that anthropomorphic physical features do not automatically denote intelligence in the same way as perceived ability to comprehend language. In general, participants assigned younger ages and lower academic levels for the low setting, and higher ones for the high setting. Moreover, participants generally assigned COZMO lower ages, including the most for Under 2 years. Of note is that no participant assigned COZMO an age of above 6-12 years for either of the low/high settings. The highest assigned academic level was undergrad, which was never assigned to COZMO. The KOBUKI and SDS robots were both variously assigned comparable older ages and average academic levels under all settings.

## 5 Prediction Tasks

Using the measures we derived from the collected data, we attempted to determine if we could predict the perceived age and academic level of the robots. We used the emotional features (*happiness, sadness, surprise, anger, fear, contempt, disgust,* and *neutral*), the prosody (F0 average), and the linguistic features (LS, LD, MSTTR) to predict

| low/high | feature | correlated feature | corr |
|---|---|---|---|
| low | (P) F0 avg | (Q) knowledgeable | 0.65 |
| | | (Q) conscious | 0.53 |
| | | (Q) friendly | 0.55 |
| | | (Q) intelligent | 0.57 |
| | | (Q) kind | 0.55 |
| | (L) LS | (Q) friendly | 0.51 |
| high | (P) F0 avg | (Q) natural | 0.53 |
| | | (Q) human-like | 0.5 |
| | | (Q) enjoyable | 0.51 |
| | | (Q) nice | 0.57 |
| | | (Q) sensible | 0.66 |
| | (L) LD | (E) sadness | 0.68 |
| | | (E) contempt | 0.53 |
| | (L) LS | (Q) age 18-24 | 0.56 |
| | | (Q) meets expect. | 0.63 |
| | | (Q) sensible | 0.62 |
| | | (Q) knowledgeable | 0.63 |
| | | (Q) responsive | 0.64 |
| **robot** | **feature** | **correlated feature** | **corr** |
| COZMO | (P) F0 avg | (Q) age 6-12 | 0.51 |
| | | (Q) 4th grade | 0.53 |
| | | (Q) lifelike | 0.62 |
| | | (Q) knowledgeable | 0.81 |
| | | (Q) competent | 0.64 |
| | | (Q) intelligent | 0.68 |
| | (L) MSTTR | (E) sadness | -0.55 |
| KOBUKI | (P) F0 avg | (Q) knowledgeable | 0.52 |
| | (L) MSTTR | (Q) age 2-5 | -0.53 |
| SDS | (P) F0 avg | (Q) liked | 0.51 |

Table 3: Spearman correlations between linguistic (L), prosodic (P), emotional (E), and questionnaire (Q) measures.

both the age and the academic level as separate classification tasks. We also predict intelligence, likability, and interpretability in order to compare to previous work.

### 5.1 Predicting the Perceived Age & Academic Level of Robots

**Data & Task** For predicting both age and academic level, we used the 58 data points from the participants for each interaction with each robot and applied those points to a 5-fold cross validation. We used a logistic regression classifier to perform the classification using the Python scikitlearn library. We report accuracy for our metric.

**Age** We ran the cross validation for two different settings when predicting age. In particular, we varied the labels that could be classified. We conducted a first task which treated all of the 7 possible outcomes for age as individual labels (i.e., under 2 years, 2-5, 6-12, 13-17, 18-24, 25-34, 35 and older) and a second task splitting at age 18 (i.e., younger than 18 is one label; 18 & older is the other label). The respective random baselines are 14% and 50%.

| **age** | **acc** |
|---|---|
| 7 labels | 22% |
| 2 labels (<, >=18) | 87% |
| **academic level** | **acc** |
| 14 labels | 27% |
| 2 labels (<, >= preschool) | 78% |

Table 4: Accuracies for predicting age and academic level.

**Academic Levels** Similar to age, we ran the cross validation for two different settings when predicting for perceived academic level. The first task treated all of the 14 possible outcomes for academic level as individual labels (preschool, kindergarten, 1-11, undergraduate; we leave out graduate and post-graduate because they were never selected in the questionnaires, nor was 12th grade), the second task treated treated preschool and beyond preschool as a binary classification task. The respective random baselines are 7% and 50%.

**Results** The results of this prediction task are in Table 4. As might be expected, when attempting to predict using many labels, the classification task is challenging with so little data. However, the classifiers beat their respective random baselines. When classifying for age, the best performing task was a binary task splitting on 18 years at 87%, effectively making it a classifier that can determine if a human user perceives the robot as an adult or as a minor. The best performing task for the academic level classification was treating preschool and above preschool as a binary classifier. Though the data is sparse, these classifiers give us useful information: a robot can use these classifiers to determine if they are perceived as an adult by human dialogue partners, and, more importantly for our purposes, as a preschool aged child, which is the age range in which we are interested for language acquisition tasks.

### 5.2 Predicting Intelligence, Likability, and Interpretability

**Data & Task** To directly compare with Novikova et al. (2017), we also predicted perceived intelligence, likability, and interpretability using a ridge regression classifier (which is optimized to reduced standard error) while considering only certain subsets of out our feature set. We evaluated when only considering emo-

| group | emot. | pros. | non-ling. | ling. | all |
|---|---|---|---|---|---|
| like | 1.08 | **0.94** | **0.94** | 1.02 | **0.94** |
| intel | 1.95 | 1.44 | 1.44 | **0.84** | 1.44 |
| interp | 0.67 | 0.7 | 0.7 | **0.61** | 0.7 |

Table 5: Performance of prediction calculated with RMSE for likability, intelligence, and interpretability. Bold denotes the smallest RMSE for a particular feature subset (emotion, prosody, non-linguistic, linguistic, and all).

tional features, prosody, non-linguistic (in our case, emotions and prosody), linguistic, and all combined features. Our metric was root mean square error (RMSE). We average the RMSE over a 5-fold cross-validation.

**Results** Table 5 shows the results of this prediction task. We found that likability is predicted best by prosody, perceived intelligence is predicted best by linguistic features, and interpretability is predicted best by also using linguistic features. One big difference between our experiment data and that of previous work is that we did not consider dialogue features (e.g., number of turns, speech duration, number of self-repetitions, etc.), which they termed as non-linguistic features. Those features were important in predicting perceived intelligence and interpretability in their work; here, linguistic and prosodic features were the most effective in predicting all three human perceptions of the robots. This confirms the work of Novikova et al. (2017) that linguistic features are a good predictor of interpretability.

## 6 Discussion & Conclusion

In this paper, we have investigated how human dialogue partners perceive the age and academic level of three robotic systems, two of which were embodied (albeit not particularly anthropomorphically), and one unembodied spoken dialogue system. We collected data from participants as they interacted with the three robotic systems then derived prosodic, emotional, and linguistic features from that participant data, and found that those features correlate with certain age and academic perceptions of those robots, as well as a number of other subjective measures from the Godspeed Questionnaire. This work confirms what previous work has shown: that humans tend to perceive robots differently depending on different factors; in our case, varying the look and spoken reposes determined how the human participants perceived the age and academic levels, as well as intelligence, likability, and interpretability of those robots. We were then able to use these features to automatically predict perceived age (i.e., adult or minor), perceived academic level (i.e., preschool or above) and perceived intelligence, likability, and interpretabilitiy. One important result of our experiment was that human dialogue partners perceive the unembodied robot (i.e., SDS) in similar ways to embodied robots; that is, the way a robot or system speaks (i.e., in our case, produces feedback by signaling either phonetic receipt or semantic understanding) is as important to human perceptions of intelligence and likability as visual characteristics.

We cannot not simply assume that human dialogue partners would treat a robot as they would a child, which is an important aspect of tasks with realistic first-language acquisition settings. The work presented here shows that those interacting with a robot like COZMO will more likely treat COZMO as a learning child instead of as an adult. This is an important result because for future work we plan on using the COZMO robot as a platform for first language acquisition research, where the setting will be more similar to first language acquisition in humans than common language grounding tasks. The COZMO robot is an afforable way for reseachers to couple spoken dialogue systems with a robotic system; it has a Python SDK which allows researchers to access its sensors (including a color camera) and control its wheel and arm movements, as well as its speech and animated face. Our results show that human users generally like COZMO, find COZMO lifelike, competent, and intelligent; i.e., COZMO may be treated as a child, but it has potential to learn.

In future work, we will apply a model of grounded semantics in a co-located dialogue setting where COZMO can learn the semantics of words as it interacts with human dialogue partners.

**Acknowledgements** This work was supported in part by the Boise State University HERC program. We would like to thank the anonymous reviewers for their comments, Hoda Mehrpouyan for use of her Kobuki robot, and the Mary Ellen Ryder Linguistics Lab at Boise State University for use of their lab for the data collection. This work was approved by Boise State University IRB 131-SB17-043.

## References


Mohit Bansal, Cynthia Matuszek, Jacob Andreas, Yoav Artzi, and Yonatan Bisk, editors. 2017. *Proceedings of the First Workshop on Language Grounding for Robotics*. Association for Computational Linguistics, Vancouver, Canada.

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

Yonatan Bisk, Kevin Shih, Yejin Choi, and Daniel Marcu. 2018. Learning Interpretable Spatial Operations in a Rich 3D Blocks World. In *Proceedings of the Thirty-Second Conference on Artificial Intelligence (AAAI-18)*, New Orleans, USA.

Dan Bohus and Eric Horvitz. 2009. Models for Multiparty Engagement in Open-World Dialog. In *Computational Linguistics*, September, pages 225–234, London, UK. Association for Computational Linguistics.

Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 33–40, Bielefeld, Germany.

Michael Collins. 2003. Head-Driven Statistical Models for Natural Language Parsing. *Computational Linguistics*, 29(4):589–637.

Bella M DePaulo and Lerita M Coleman. 1986. Talking to children, foreigners, and retarded adults. *Journal of Personality and Social Psychology*, 51(5):945–959.

Scott G Eberle. 2009. Exploring the Uncanny Valley to Find the Edge of Play. *American Journal of Play*, 2(2):167–194.

Friedericke Eyssel and Frank Hegel. 2012. (S)he's Got the Look: Gender Stereotyping of Robots1. *Journal of Applied Social Psychology*, 42(9):2213–2230.

Friederike Eyssel and Dieta Kuchenbrandt. 2012. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4):724–731.

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

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

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

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

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding Semantics in Olfactory Perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231–236, Beijing, China. Association for Computational Linguistics.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, page 259.

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

Xiaofei Lu. 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *Modern Language Journal*, 96(2):190–208.

Xiaofei Lu. 2014. *Computational methods for corpus annotation and analysis*.

Caroline Lyon, Chrystopher L Nehaniv, Joe Saunders, Tony Belpaeme, Ambra Bisio, Kerstin Fischer, Frank Förster, Hagen Lehmann, Giorgio Metta, Vishwanathan Mohan, Anthony Morse, Stefano Nolfi, Francesco Nori, Katharina Rohlfing, Alessandra Sciutti, Jun Tani, Elio Tuci, Britta Wrede, Arne Zeschel, and Angelo Cangelosi. 2016. Embodied Language Learning and Cognitive Bootstrapping: Methods and Design Principles. *International Journal of Advanced Robotic Systems*, 13(105).

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

Jekaterina Novikova, Christian Dondrup, Ioannis Papaioannou, and Oliver Lemon. 2017. Sympathy Begins with a Smile, Intelligence Begins with a Word: Use of Multimodal Features in Spoken Human-Robot Interaction. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 86–94.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language—State-of-the-art and the challenge. *Computer Speech & Language*, 27:4–39.

Lanbo She and Joyce Y Chai. 2016. Incremental Acquisition of Verb Hypothesis Space towards Physical World Interaction. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 108–117.

Maria Stubbe. 1998. Are you listening? Cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics*, 29(3):257–289.

Benedict Tay, Younbo Jung, and Taezoon Park. 2014. When stereotypes meet robots: The double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior*.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. 2016. Learning MultiModal Grounded Linguistic Semantics by Playing " I Spy ". In *Proceedings of IJCAI*, pages 3477—-3483.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics* -, volume 13, pages 63–70.

Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. 2014. A framework for learning semantic maps from grounded natural language descriptions. *The International Journal of Robotics Research*, 33(9):1167–1190.