Symbol and Communicative Grounding through Object Permanence with a Mobile Robot

Josue Torres-Fonseca
Boise State University
1910 W University DR
Boise, ID 83725
josuetorresfonse@u.boisestate.edu

Catherine Henry
Boise State University
1910 W University DR
Boise, ID 83725
catherinehenry@u.boisestate.edu

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

Abstract

Object permanence is the ability to form and recall mental representations of objects even when they are not in view. Despite being a crucial developmental step for children, object permanence has had only some exploration as it relates to symbol and communicative grounding in spoken dialogue systems. In this paper, we leverage SLAM as a module for tracking object permanence and use a robot platform to move around a scene where it discovers objects and learns how they are denoted. We evaluated by comparing our system's effectiveness at learning words from human dialogue partners both with and without object permanence. We found that with object permanence, human dialogue partners spoke with the robot and the robot correctly identified objects it had learned about significantly more than without object permanence, which suggests that object permanence helped facilitate communicative and symbol grounding.

1 Introduction

Communicative grounding is the process of mediating what words mean (Clark, 1996) and *symbol grounding* is the establishment of connections between language and the perceptual, physical world (Harnad, 1990). Following Larsson (2018) that explained how symbol grounding is a side effect of communicative grounding, children who are learning their first language cannot learn symbol grounding without simultaneously being engaged in communicative grounding. Consider the following example, within the physical space of a room. A child (C) picks up a ball (B) and a caregiver (P) engages in dialogue with the child about the ball:

(1) a. (C picks up a B and looks at it)
b. P: That's a ball!
c. C: ball
d. P: Ball! Very good!

Communicative grounding happens between P and C during this interaction as P offers *ball* as a word with a semantic potential and C understands B to be an extension of *ball*. At the point (1)-b symbol grounding takes place between C and B where C links the word *ball* to the object in their hand. Communicative grounding then follows when C says *ball* and receives a positive confirmation from P, resulting in knowledge that P has experienced an interaction with C when C heard and demonstrated understanding of *ball*, and C received confirmation of understanding of the word *ball* from P.

But what happens in Example (1) when C moves their attention to a different object? It is the case that the C has grounded the word *ball* using their experience with B, and P acknowledges that C has done so, but does it matter that the object is no longer in view? Prior work explored the interplay between communicative and perceptual grounding (Chai et al., 2014; Larsson, 2018), but there is very little work on how *object permanence* plays a role in the communicative and symbol grounding process. Piaget identified object permanence in the child development process within the sensorimotor stage—a period that lasts from birth to nearly two years old (i.e., beginning before children can speak) when children largely interact with and understand the world through their sensorimotor experience (Piaget, 2013; Bremner et al., 2015). Moreover, children who are learning their first words are *egocentric* in that they have not yet developed the capability of understanding another person's point of view (i.e., of an object) (Repacholi and Gopnik, 1997). A lack of object permanence means that objects that children observe, but are then out of view no longer exist, and are separate and distinct objects if the child observes them again.[1]

[1] Lack of object permanence is the common assumption that holds for most vision and language datasets, e.g., refCOCO (Yu et al., 2016) where referring expressions to ob-

Proceedings of the SIGdial 2022 Conference, pages 124–134
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 ©2022 Association for Computational Linguistics

Moore and Meltzoff (1999) suggested that as early as four months, a child begins to recognize that objects have permanence even when the child is not actively observing them—an ability that the child can leverage before they start to learn language—but this knowledge has been ignored in prior research. Therefore, in this paper, we ask the question: *Does object permanence matter for communicative grounding and symbol grounding in an automated learning spoken dialogue system?* We hypothesize that it does matter, particularly for first-language acquisition in a spoken dialogue system (SDS) that has no prior exposure to language. We test our hypothesis in a human-robot interaction (HRI) task where we task human participants to interact with a robot and observe that the robot has been able to utter words in the right context. We use a survey to measure the perceptions of the human participants in order to establish that communicative grounding took place, and we measure the number of words that the robot "learned" during the interaction to determine if communicative and symbol grounding took place. We find through our experiment that symbol and communicative grounding are affected by object permanence, leading to increased user engagement and a more responsive and effective spoken dialogue system that learns word groundings as it interacts.

In the following section, we compare our work to others then explain our method for tracking object permanence using a *simultaneous localization and mapping* (SLAM) module and the the robot-ready SDS system that we used. We then explain our experiment and conclude.

2 Background & Related Work

Object permanence is a crucial milestone in cognitive development, and it has been suggested by Moore and Meltzoff (1999) that as early as four months this milestone is reached. Tomasello and Farrar (1984) shows that as infants enter the sixth stage of object permanence development (where children understand that objects completely removed from their view still exist) they start to learn relational words. A more recent study explores the development of search behavior in 7 month old infants after they guide them in understanding the effects of their actions upon hidden objects. This indicates that object permanence is crucial in searching behavior as it leads to the understanding that infants have the ability to cause hidden objects to reappear (O'Connor and Russell, 2015).

Bechtle et al. (2015) worked towards developing a sense of object permanence in robots through creating a simulated experimental setup where a robot learns how the movements of its arms (one holding a shield) affect the visual detection of an object in a scene. Although, not directly related to object permanence, Platonov et al. (2019) is more closely related to grounding as they create a SDS which is able to create a 3D model of a physical block world and answer spatial questions about it. Roy et al. (2004) also explored spatial reasoning within a physical world through the creation of a robot called Ripley which performed grounding of spatial language that could not be understood under fixed-perspective assumptions.

Of similar importance in cognitive development is communicative grounding. Researchers, notably Chai et al. (2014), have investigated how the collaborative efforts of a robot in situated human-robot dialogue affects both perceived and true grounding which involved a situated setup of objects similar to our experiment. This notion of common ground and communicative grounding has also been explored in other human-robot interaction work (Kiesler, 2005; Powers et al.; Stubbs et al., 2007, 2008; Peltason et al., 2013) and work involving human interactions with virtual agents (Pustejovsky et al., 2017). Our work extends and builds on prior work as we focus on using object permanence in a robot to improve its language learning abilities.

3 Proposed System

In this section we explain how we modeled the dialogue for language learning, integrated with robot modules. We first explain the choice of robot: Digital Dream Lab's Cozmo robot. Plane et al. (2018) showed that participants perceived Cozmo as young and with potential to learn, which is precisely the setting and perception that we want dialogue partners to have when interacting with Cozmo. Cozmo is small, has a track for movement, a lift and a head with an OLED display which allow it to display its eyes. Within the head is a small camera and a speech synthesizer (with a "young" sounding voice). For this study we make use of Cozmo's camera for object detection, track for navigation and most importantly Cozmo's built-in

jects depicted in images only offer a single visual experience (though multiple referring expressions) to the objects.

SLAM (Simultaneous Localization and Mapping) functionality for object permanence. Cozmo has no microphone, so we use an external microphone.

The system outlined in this paper uses the incremental framework ReTiCo (Michael and Möller, 2019; Michael, 2020) extended for multimodal use with Cozmo (Kennington et al., 2020), leveraging existing modules as well as the newly developed Object Permanence module. The full SDS is depicted in Figure 3. The modules include: Object Detection, Feature Extraction, Automatic Speech Recognition, Natural Language Understanding, Grounded Semantics, Action Management (Navigation & Speaking), and Object Permanence.

Object Detection The Object Detection module uses YOLO object detection (Redmon et al., 2016). The model we used was pre-trained on the MSCoco dataset (Lin et al., 2014) containing 91 object types with a total of 2.5 million labeled objects in 328 thousand images. We apply this model as a means for object region classification in order to draw bounding boxes around objects in images received from Cozmo's Camera. We discard the labels and only use the bounding box information as to avoid the use of a pretrained vocabulary since children are born without linguistic knowledge. The output of this module is the bounding box information of the objects in view to Cozmo.

Feature Extraction The Feature Extraction module uses CLIP (Radford et al., 2021) a neural network trained on a variety of (image, text) pairs. This module takes an image and bounding box information, extracts each sub-image containing each object, then passes those through CLIP's image encoder which returns image features encoded by the vision portion of the CLIP model. This module outputs a vector of size 512 for each detected object, for each frame. In our case only one object will be detected in an image, though as the robot shifts and moves, multiple frames of the object will results in multiple CLIP vector representations of that object. Taken together, the Object Detection and Feature extraction modules provide a way of isolating and extracting features from objects; children likewise have experienced objects physically (i.e., visual, tactile) before they learn that words denote objects. Both modules use models that were trained using language data which certainly affects functionality of the modules. We ignore the language aspects of the models, and leave for future work developing models (e.g., object region detection) that are trained without language data.

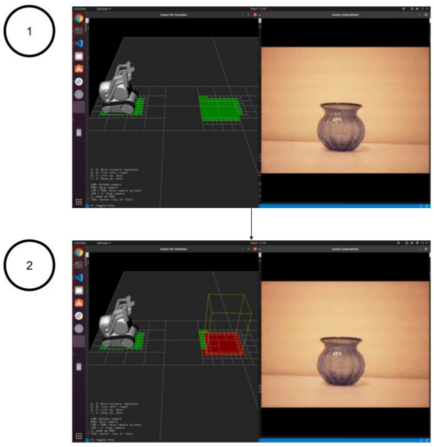


Figure 1: Visualization of the creation of a custom object in SLAM. In 1, the object is not yet observed, but in 2 the object is placed in the SLAM space.

Automatic Speech Recognition The Automatic Speech Recognition (ASR) module transcribes user speech. We use Google's speech to text API. The output is the word-level transcription.

Natural Language Understanding The Natural Language Understanding (NLU) module takes in the transcribed speech from the ASR and determines the dialogue act (i.e., *intent*) of the user using RASA (Bocklisch et al., 2017) an open source NLU library. Specifically, we use RASA to categorize user speech into 5 different dialogue acts:

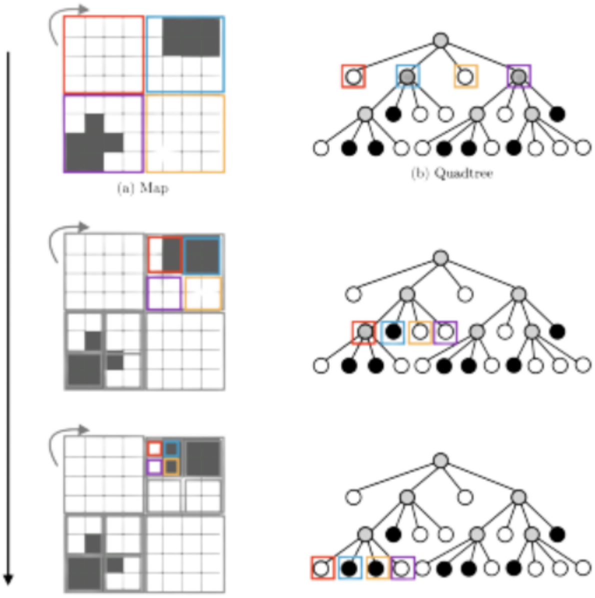
- positive user feedback (e.g., *yes*)
- negative user feedback (e.g., *no*)
- where questions (e.g., *where is the can?*)
- what questions (e.g., *what is that?*)
- statements (e.g., *that is red.*)

The positive and negative user feedback is used to document the number of questions that Cozmo answered correctly and incorrectly from the participant. We categorize *where* and *what* questions so that Cozmo can differentiate between initiating finding behavior (*where* questions) and answering questions using the best known word about an object (*what* questions). This signals to our system when it should be in a state of learning how words ground into images, or whether it should be exploiting what it knows in order to locate and identify an object it has seen. This fairly simplistic ontology of dialogue acts is in line with child development; children can infer intent of positive and negative feedback, as well as simple questions like location before they are able to speak, albeit often through extra-linguistic information such as prosody and affective displays (see (Locke, 1995), chapters 3-5). We trained RASA on 19 hand-crafted examples of positive user feedback, 10 examples of negative user feedback, 25 examples of what questions, 22 examples of where questions, and 747 examples of statements that we extracted from random samples of text from Wikipedia. We train on 747 examples of statements because statements are the most difficult to identify as there are many different variations of statements therefore requiring many training examples.

Grounded Semantics The Grounded Semantic Module performs symbol grounding by mapping heard words (though the ASR) to observed objects. The module makes use of the the Words as Classifiers (WAC) model (Kennington and Schlangen, 2015). In the WAC model, each word is represented by its own classifier trained on positive and negative examples of real-world referents and has been shown to learn words with only a few examples, which is critical for our task that is intended to mimic how children fast-map words to objects. The module learns as it "hears" a word (i.e., a recent update from the ASR module) and is currently observing an object. The WAC model associates words with the detected objects (i.e., represented as CLIP vectors) as positive examples. The module systematically trains individual logistic regression classifiers for each word as it hears words and associates those words with objects. Negative examples for training are randomly sampled from positive vectors associated with other words (the system must have heard at least two words and associated some objects with them in order to train). The classifiers are trained every time an utterance is spoken and after observing an object every 20 added frames.

The Grounded Semantics module has two modes: explore and exploit. In the explore mode, the module associates words with objects and trains the individual word classifiers as explained above. In the exploit mode, the module instead uses the recently heard words and either attempts to identify the object that is the best fit for the description or it attempts to determine which word is the best fit for an object that is currently under observation. The module's mode is determined by the speech act as signalled by the NLU module, explained above.

Action Management For dialogue (and robot action) management we use PyOpenDial (Jang et al., 2019). This module acts as a broker of the entire dialogue state to map from states to actions. In our case, the primary actions are `explore` when the robot drives around looking for objects, `find` when the dialogue partner asks about an object, `learn` when the robot should be associating words with objects, and `answer` when the robot should utter something in response to a dialogue partner's *where* or *what* dialogue act. The *explore* action is the default. In the *explore* state, Cozmo randomly drives in front of one of the 7 different objects (see Figure 4).


(a) Map

(b) Quadtree

Figure 2: Quad-tree Maps are space efficient alternatives to an an occupancy map where open space is compressed into a single "unoccupied" cell. White means no cells are occupied, grey means some, and black means all. The root node represents the entire map and children are arranged in clockwise order. The first child node corresponds to the 4x4 grid in the upper left.

Object Permanence The Object Permanence module is an application of a SLAM module that is part of the Cozmo robot's functionality. The goal of the SLAM module is to track the position and location of observed objects in a 3-dimensional space. The surface that the robot can drive on is a 2-dimensional plane that the SLAM module breaks into very small cells. The SLAM module then uses *quad-tree maps* (Finkel and Bentley, 1974) to determine which cells are occupied and which ones are free. Representing the space as a quad-tree map allows SLAM to store and retrieve object lo-

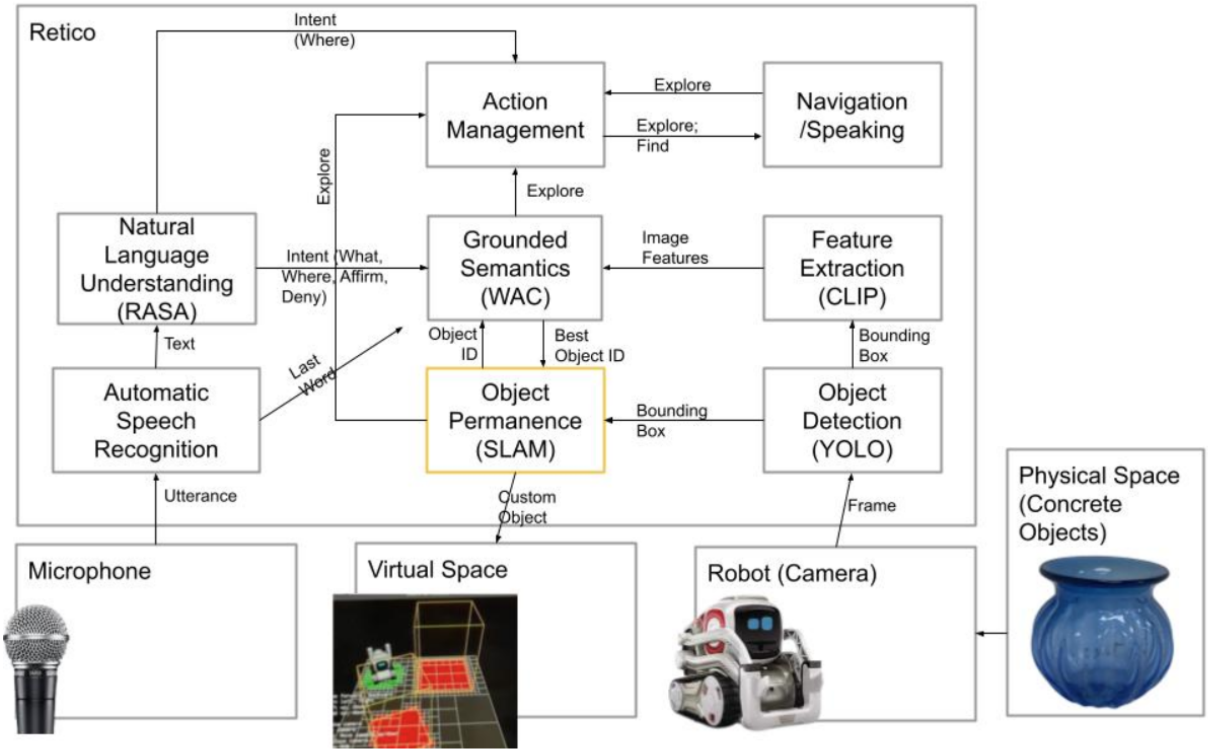


Figure 3: Schematic of our system.

cations efficiently. An example of a quad-tree map is shown in Figure 2. While we do not argue that humans use quad-tree maps for organizing object permanence, it serves as a functional approximation of what object permanence affords: the ability to remember objects and their locations.

When the system is first invoked, there is initially no history of observed objects and the robot's starting point becomes the point of reference for everything that the robot will observe. As the robot moves (i.e., drives forward or backward, and turns left or right) the SLAM module can track precisely how far and in which direction the robot has moved from its origin. The original SLAM module for Cozmo is designed to track specific objects based on a marker code (i.e., three blocks each with QR-like symbols). We extended the functionality of the SLAM module to include any object that is observed by the Object Detection module described above. We use that module's bounding box information (See Figure 1) and current observed location relative to where Cozmo is facing to infer the object's location and uniqueness. The uniqueness here is important because if the robot moves away from the object then returns to it later, the robot should be able to identify the object as one that has been seen before, not as a new object. For each new object that the robot observes, the Object Permanence module assigns a unique identifier. The unique identifier is shared with the Grounded Semantics module so it can associate specific objects (i.e., their CLIP vectors) with words that were used to describe those specific objects.

Traditional symbol grounding generally only visualizes representations of objects and associates those with referring expressions or descriptions, but the identity of objects is discarded during testing. Here, the WAC model not only learns word groundings through experience as it observes words uttered in association with observing objects, but it also uniquely identifies each object and keeps a history of its visual experience with each object regardless of how they were referred to or described. Importantly, the SLAM functionality does not just identify unique objects, it gives the robot the ability to return directly to that object without colliding with other objects because it tracks all objects that Cozmo has observed.

System Task Behavior The default action for Cozmo is `explore` which is done by randomly choosing a position at one of the drawn squares in front of all seven objects shown in 4 and centering its camera to the closest object. Once in front of an object, Cozmo waits up to 10 seconds for an utterance from a dialogue partner. If the partner utters

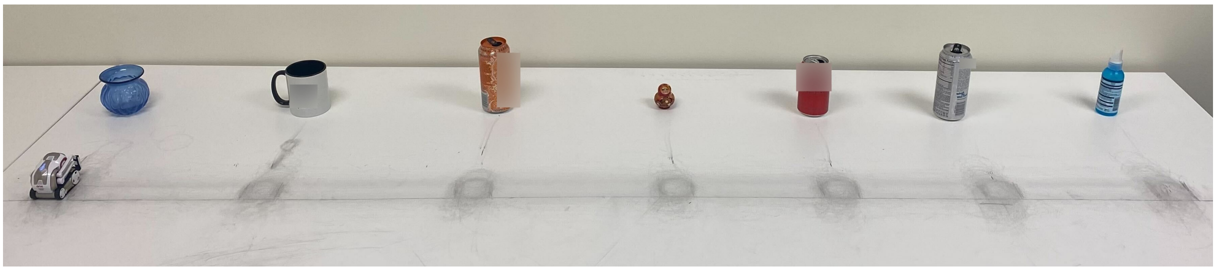


Figure 4: The seven objects used for our experiment

something, Cozmo assumes the words are about the object in view and the Grounded Semantics module learns by associating the last uttered word with the object. If no utterance is given, Cozmo moves away from the object and continues to `explore`. If the dialogue partner continues to speak, Cozmo remains in front of the object.

The `find` state is activated when the NLU detects that the dialogue partner has uttered a *where* dialogue act. For example, *where is a can?* would result in a detected `find` dialogue act. This triggers the Grounded Semantics module to find the object in its history that is the most probable fit for the description (in this case, the word *can* might ground more strongly to one object compared to others). The Grounded Semantics module then signals to the Object Permanence module to drive to and face the object with the specified identifier. Once the robot reaches the object, it utters back the description (i.e., *can*). At this point the dialogue partner can utter positive or negative feedback. When the Object Permanence module is not available (i.e., our baseline system version), Cozmo randomly explores objects and the Grounded Semantics module determines if the description fits the currently observed object using the last word in the utterance. If the probability of the model is above 0.5, then Cozmo repeats what it heard to signal that Cozmo found the object. The user can then utter positive or negative feedback.

Another dialogue act is the *what* question. If the robot is currently looking at an object, then the system assumes the *what* dialogue act is about the currently observed object and looks through its history to find the best known word for the object currently in view and utters the best known word.

In the following section, we will explain how we evaluated our SDS and whether or not Cozmo's language learning abilities improved with the use of the Object Permanence module.

4 Evaluation

In this section, we explain how we evaluated our model with human participants to determine if Cozmo performs communicative and symbol grounding more effectively with Object Permanence. We compare two versions of our system: one that did not have an access to the Object Permanence module and one that did. Our evaluation included objective measures logged by the system and by the participants used to measure symbol grounding by tracking correctly "learned" words, as well as subjective measures collected using participant questionnaires used to measure communicative grounding.

Procedure Study participants met in our lab located near Boise State University's Computer Science building. The lab is setup for the participant interaction as follows. A large table is setup with 7 objects on the table as shown in Figure 4. We chose the 7 objects to vary shape and color, but wanted to have a degree of overlap for words that might be used to describe them (e.g., *can* or *blue*).

In front of each object is a straight line drawn on the table and a box.[2] Cozmo is placed in front of the leftmost object. The microphone that feeds into the ASR module is positioned in front and to the left of the table with the objects and Cozmo. The participant stands or sits (as they prefer) at the front of the table. Cozmo is not introduced to the participant until the participant has signed a consent form and the task has been explained to them. The experimenter was present to examine the state of the robot and the microphone, answer any questions the participant may have, and troubleshoot any problems that arose. The experimenter was permitted to offer a constrained set of coaching tips to the participant during the experiment, given the participant needed a reminder of their task or

[2] The line and box does not affect Cozmo, it is just there to help the participant adjust Cozmo when needed.

the instructions. Following each interaction with Cozmo, the participant was instructed to complete a questionnaire. Following the completion of the experiment and surveys, the participant was paid $10. We recruited 24 participants to interact with Cozmo for two twenty-minute periods over the course of a single session. Most study participants recruited were from the Boise State University Department of Computer Science. 18 of the participants were male; 6 were female. The entire time for each participant was approximately one-hour.

After signing the informed consent, Cozmo was introduced to the participant, with the following explanation; (1) Cozmo has a camera that can see the world; (2) Cozmo has a microphone and can hear them; (3) Cozmo doesn't know anything, but would like to know more about the world; (4) for the next 20 minutes, it is your job to teach Cozmo as many words as they can, about the seven objects in front of Cozmo; (5) Cozmo will move in front of an object. If Cozmo does not hear you speak he will move on to a new object. If Cozmo does hear you speak, then it will observe the object and repeat the word he learned. The word Cozmo learned will always be the last word you spoke. Do not teach Cozmo any more words until it repeats this word. For every word you teach it, you can write it down so you can keep track; (6) if Cozmo is not on the square in front of the object when he moves to an object you must readjust it to that square; (7) After teaching Cozmo about two different words for two different objects you can and should ask it *what* and *where* questions to check his knowledge; (8) For every question he gets right answer "yes" and put down a tally mark to record a correct answer for every question he gets wrong answer "no" only; (9) Only speak to Cozmo when he is in front of an object.

We used an A/B design, meaning that each participant went through the same procedure twice, once with Cozmo having access to Object Permanence and once without access. To mitigate priming effects, the order in which the test condition was presented was alternated.

Two System Versions The test condition is the system version that had access to Object Permanence and is explained in Section 3. The baseline point of comparison for this study was a system that did not have access to the Object Permanence module. The overall functionality of the baseline system was the same as the system with access to Object Permanence, except that the system could not track and locate objects when participants asked them to. The Grounded Semantics module in this case only performs traditional symbol grounding between words and visual representations—not specific objects. This meant that the robot behavior when a `find` dialogue state was entered (i.e., after a *where* dialogue act from the participant) was different: instead of moving directly to the identified object, the robot would move towards a random object one at a time and check each object to determine if they matched the description. If an object did match, the robot would repeat the description to the participant, who then in turn offered positive or negative feedback. Under the best circumstances for the baseline system, the robot would randomly move towards an object that fit the description on the first attempt. But if the first object did not fit the description, then the robot moved towards a different object and repeated until an object matched the description. To give the baseline version a higher chance of the robot actually finding the objects, the objects were placed in a line and the robot systematically drove directly to a randomly selected object. This was designed to give the baseline system version some degree of the object permanence functionality as a stronger point of comparison.

Metrics All module communication is logged using the Platform for Situated Intelligence (Bohus et al., 2017). We specifically track the number of utterances made by the participants, including positive and negative feedback, and the number of questions asked. The participant themselves keep track of the number of questions Cozmo correctly answers (i.e., if Cozmo correctly identified an object). These metrics act as a way to measure symbol and communicative grounding, as well as engagement (i.e., more utterances means more engagement).

We evaluate the robot based on questionnaire responses filled out by the participants following each interaction to establish that communicative grounding took place. We used the Godspeed Questionnaire (Bartneck et al., 2009), a 5-point Likert-scaled questionnaire with 24 questions using negative (left side) to positive (right side) ratings of a robot's anthropomorphism, animacy, likeability, and perceived intelligence. In addition to the Godspeed questions, we asked the participants the following questions to further ascertain their perceptions of our system and robot (some items have boldface text to link them with results):

(Mean / std. dev)	baseline	with obj. perm.	p-value
Heard Words	15.9 / 3.9	18.8 / 4.7	0.02
Questions Asked	10.7 / 3.2	20.0 / 6.7	3.7e-7
% Correct	70.0 / 22.2	82.7 / 12.1	0.02

Table 1: The effect of object permanence on a language acquisition task

Interesting	0.0048
Spend more Time	0.049
Responsive	0.083
Intelligence	0.10

Table 2: Statistical Significance between values with and without Object Permanence using a t-test.

(Mean / std. dev)	1st Interaction (A)	1st Interaction (B)	2nd Interaction (A)	2nd Interaction (B)
Heard Words	19.4 / 4.7	16.2 / 2.7	18.3 / 5.0 (0.56)	15.6 / 4.9 (0.72)
Questions Asked	16.4 / 4.0	12.0 / 3.5	23.0 / 7.5 (0.01)	9.7 / 2.7 (0.13)
% Correct	83.2 / 11.5	74.7 / 20.4	82.1 / 13.0 (0.84)	64.9 / 23.7 (0.30)

Table 3: The effect of initial setting on a language acquisition task (A is for with Object Permanence and B is for without. Furthermore, the values in parentheses near the values for the second interaction represent the p-values between the values in 1st Interaction compared to 2nd Interaction for A and B)

- How attached to the robot did you feel?
- How **interesting** was the robot to interact with?
- Would you like to **spend more time** with the robot?
- How many years old do you think the robot is (in terms of its behavior)?

Results Table 1 shows the effect object permanence has on Cozmo's language acquisition abilities.[3] It is clear that with object permanence, Cozmo is perceived to learn language better than without object permanence as shown by the statistical significance values. This suggests that object permanence does appear to have an affect on symbol grounding especially as Cozmo not only hears more words on average per participant with the test condition than without, his accuracy in answering questions also increases by approximately 13%.

Relating to participant perceptions of the robot and interaction, we find that overall the mean values for the ratings were higher for the test condition than the baseline except for three questions which relate to kindness, and feelings of calmness and interest at the beginning of the interaction. Therefore, showing that overall, the test condition positively influenced users' perception of Cozmo. Furthermore, we observe that with object permanence, participants believed that Cozmo learned better than without, as seen by the overall higher intelligence and responsiveness scores in Figure (1), though note that the difference in perceived intelligence is not significant, which tells us that the baseline system was still viewed positively and therefore provided a high point of comparison.

Participants on average estimated Cozmo's age with the test condition at 3.5 years of age compared to 2.6 years of age with the baseline, suggesting that Cozmo was perceived to be more intellectually advanced with the test condition, but still an early language learning child, which also tells us that the robot did not exhibit behaviors that participants perceived as too advanced for our task. We also observe higher responsiveness in the object permanence version which likely results from participants observing that Cozmo answered questions quickly and with high accuracy, suggesting that communicative grounding was better with the object permanence version (see Appendix for more results comparing perceived Intelligence and Responsiveness).

Finally, ratings for interest and desire to spend more time with the robot are significantly higher with object permanence than without. This is especially evident when observing that the mean value for interest at the end of the interaction is 4.7 with the test condition and 4.2 without; the average increase in interest from the beginning of the interaction for the test condition is 0.54 as compared to 0.83 without. Furthermore, using questions asked as a measurement for engagement (since it shows active interest in what Cozmo is learning) we observe that with Object Permanence, Cozmo is asked

[3] Nine interactions had to be restarted due to unexpected events (e.g., Cozmo rolled off the table) which affected the SLAM map and learned words, but this happened at roughly the same frequency for both settings. Cozmo also picked up his own voice in the microphone in both settings, but this also happened at roughly the same frequency for both settings so we decided to leave it as part of the data.

approximately 9 more questions than without showing that object permanence has a significant effect on engagement (see Table 3). This is crucial, because the interaction itself needs to motivate human participants to “buy into” the robot’s language learning by spending time and effort helping it learn. See also Figure 5 in the Appendix for more results.

5 Conclusion

We conducted an experiment with twenty-four participants who performed a language acquisition task with Cozmo both with and without object permanence. We analyzed our results by comparing the participants’ survey responses to measure communicative grounding and number of words heard, questions asked, and percent of questions answered correctly to measure symbol grounding between the experimental and control interactions. We found that a robot with object permanence resulted in improved communicative and symbol grounding due to stronger engagement from the participant and a higher percentage of correct answers from Cozmo. User perceptions of Cozmo with object permanence also greatly improved overall. This indicates that object permanence does in fact have a positive affect on communicative and symbol grounding. Our findings suggest that an understanding of object permanence is a necessary component of any spoken dialogue system built to reach the potential of natural dialogue between humans.

Acknowledgements Thanks to the anonymous reviewers for their very useful feedback. This material is based upon work supported by the National Science Foundation under Grant No. 2140642.

References

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.

Sarah Bechtle, Guido Schillaci, and Verena V Hafner. 2015. First steps towards the development of the sense of object permanence in robots. In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 283–284.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management.

Dan Bohus, Sean Andrist, and Mihai Jalobeanu. 2017. Rapid Development of Multimodal Interactive Systems: A Demonstration of Platform for Situated Intelligence. In *Proceedings of ICMI*, Glasgow, UK. ACM.

J. Gavin Bremner, Alan M. Slater, and Scott P. Johnson. 2015. Perception of object persistence: The origins of object permanence in infancy. *Child Development Perspectives*, 9(1):7–13.

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.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Raphael A Finkel and Jon Louis Bentley. 1974. Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1):1–9.

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

Youngsoo Jang, Jongmin Lee, Jaeyoung Park, Kyeng-Hun Lee, Pierre Lison, and Kee-Eung Kim. 2019. PyOpenDial: A python-based Domain-Independent toolkit for developing spoken dialogue systems with probabilistic rules. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 187–192, Hong Kong, China. Association for Computational Linguistics.

Casey Kennington, Daniele Moro, Lucas Marchand, Jake Carns, and David McNeill. 2020. rrSDS: Towards a robot-ready spoken dialogue system. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 132–135, 1st virtual meeting. Association for Computational Linguistics.

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.

S Kiesler. 2005. Fostering common ground in human-robot interaction. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 729–734.

Staffan Larsson. 2018. Grounding as a Side-Effect of grounding. *Top. Cogn. Sci.*

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

John L Locke. 1995. *The Child's Path to Spoken Language*. Harvard University Press.

Thilo Michael. 2020. Retico: An incremental framework for spoken dialogue systems. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 49–52, 1st virtual meeting. Association for Computational Linguistics.

Thilo Michael and Sebastian Möller. 2019. Retico: An open-source framework for modeling real-time conversations in spoken dialogue systems. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pages 134–140. TUDpress, Dresden.

M. Keith Moore and Andrew N. Meltzoff. 1999. New findings on object permanence: A developmental difference between two types of occlusion. *British Journal of Developmental Psychology*, 17(4):623–644.

Richard J O'Connor and James Russell. 2015. Understanding the effects of one's actions upon hidden objects and the development of search behaviour in 7-month-old infants. *Dev. Sci.*, 18(5):824–831.

Julia Peltason, Hannes Rieser, Sven Wachsmuth, and Britta Wrede. 2013. On grounding natural kind terms in human-robot communication. *KI-Künstliche Intelligenz*, 27(2):107–118.

Jean Piaget. 2013. *The construction of reality in the child*, volume 82. Routledge.

Sarah Plane, Ariel Marvasti, Tyler Egan, and Casey Kennington. 2018. 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, Melbourne, Australia. Association for Computational Linguistics.

Georgiy Platonov, Benjamin Kane, Aaron Gindi, and Lenhart K Schubert. 2019. A spoken dialogue system for spatial question answering in a physical blocks world. *arXiv:1911.02524 [cs]*.

Aaron Powers, Adam Kramer, Shirlene Lim, Jean Kuo, Sau-Lai Lee, and Sara Kiesler. Common ground in dialogue with a gendered humanoid robot. Accessed: 2022-5-2.

James Pustejovsky, Nikhil Krishnaswamy, Bruce Draper, Pradyumna Narayana, and Rahul Bangar. 2017. Creating common ground through multimodal simulations. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

B M Repacholi and A Gopnik. 1997. Early reasoning about desires: evidence from 14- and 18-month-olds. *Dev. Psychol.*, 33(1):12–21.

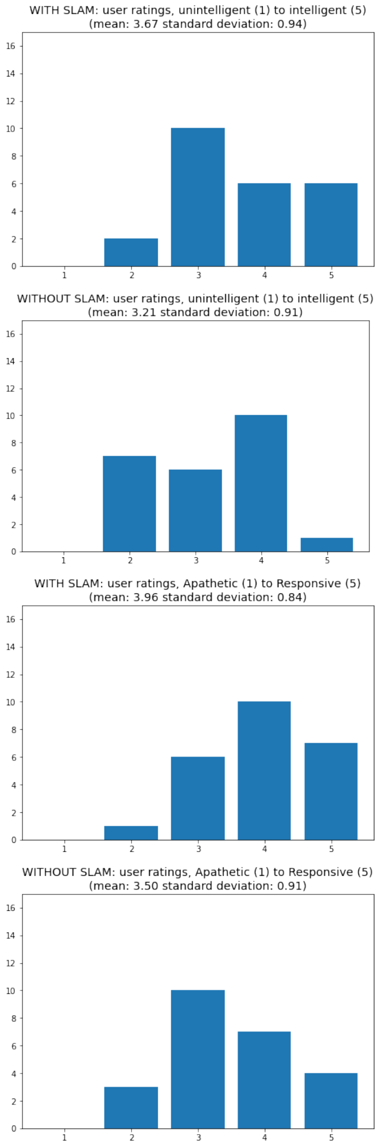
D. Roy, Kai-Yuh Hsiao, and N. Mavridis. 2004. Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(3):1374–1383.

Kit Stubbs, David Wettergreen, and Illah Nourbakhsh. 2008. Using a robot proxy to create common ground in exploration tasks. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, HRI 2008, Amsterdam, The Netherlands, March 12-15, 2008*, pages 375–382. unknown.

Kristen Stubbs, Pamela J Hinds, and David Wettergreen. 2007. Autonomy and common ground in Human-Robot interaction: A field study. *Intelligent Systems, IEEE*, 22(2):42–50.

Michael Tomasello and Michael Jeffrey Farrar. 1984. Cognitive bases of lexical development: object permanence and relational words*. *J. Child Lang.*, 11(3):477–493.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.



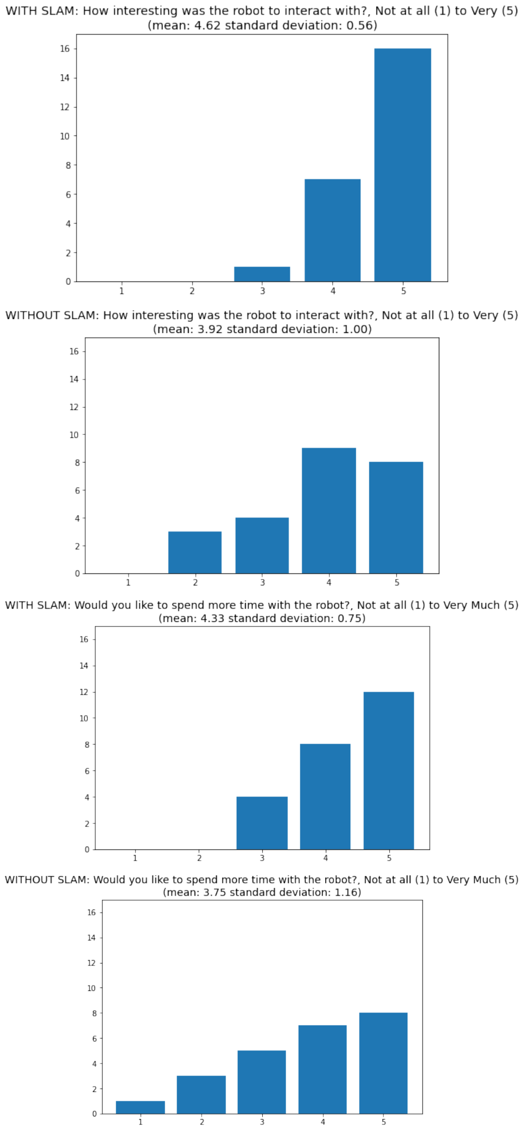
WITH SLAM: user ratings, unintelligent (1) to intelligent (5)
(mean: 3.67 standard deviation: 0.94)

WITHOUT SLAM: user ratings, unintelligent (1) to intelligent (5)
(mean: 3.21 standard deviation: 0.91)

WITH SLAM: user ratings, Apathetic (1) to Responsive (5)
(mean: 3.96 standard deviation: 0.84)

WITHOUT SLAM: user ratings, Apathetic (1) to Responsive (5)
(mean: 3.50 standard deviation: 0.91)

Figure 6: Intelligence and Responsiveness ratings for Cozmo with and without object permanence



WITH SLAM: How interesting was the robot to interact with?, Not at all (1) to Very (5)
(mean: 4.62 standard deviation: 0.56)

WITHOUT SLAM: How interesting was the robot to interact with?, Not at all (1) to Very (5)
(mean: 3.92 standard deviation: 1.00)

WITH SLAM: Would you like to spend more time with the robot?, Not at all (1) to Very Much (5)
(mean: 4.33 standard deviation: 0.75)

WITHOUT SLAM: Would you like to spend more time with the robot?, Not at all (1) to Very Much (5)
(mean: 3.75 standard deviation: 1.16)

Figure 5: Engagement ratings for Cozmo with and without object permanence