BOISE STATE UNIVERSITY

SPEECH, LANGUAGE & INTERACTIVE MACHINES

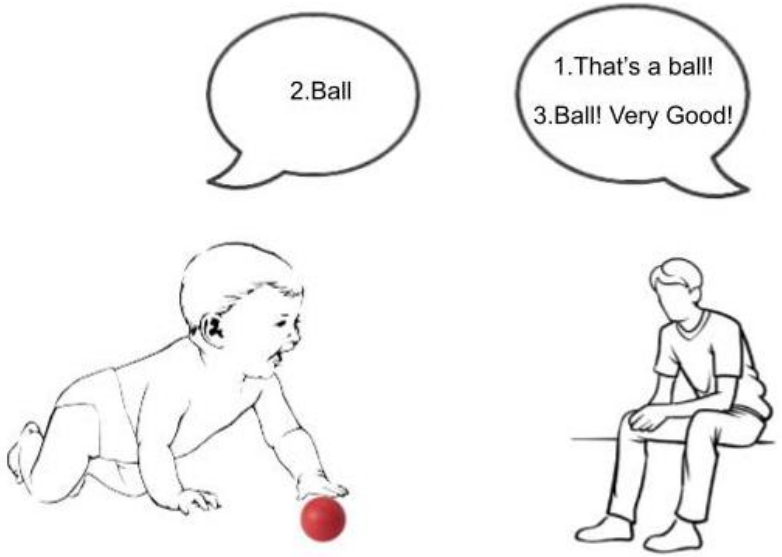
Symbol and Communicative Grounding Through Object Permanence with a Mobile Robot

JOSUE TORRES-FONSECA, CATHERINE HENRY, DR. CASEY KENNINGTON

DEPARTMENT OF COMPUTER SCIENCE, BOISE STATE UNIVERSITY

Communicative and Symbol Grounding

- 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)
- Symbol grounding is a side effect of communicative grounding



Object Permanence

- Ability to form and recall mental representations of objects even when they are not in view
- Piaget identified object permanence in sensorimotor stage (birth to 2 yrs old) where children interact with and understand the world through sensorimotor experience
- It has been suggested that as early as four months, a child begins to recognize that an object has permanence (Moore and Meltzoff, 1999)
- Children learning their first language are egocentric (Repacholi and Gopnik, 1997)
- **Research Question: Does object permanence matter for communicative grounding and symbol grounding in an automated spoken dialogue system?**

Background & Related Work

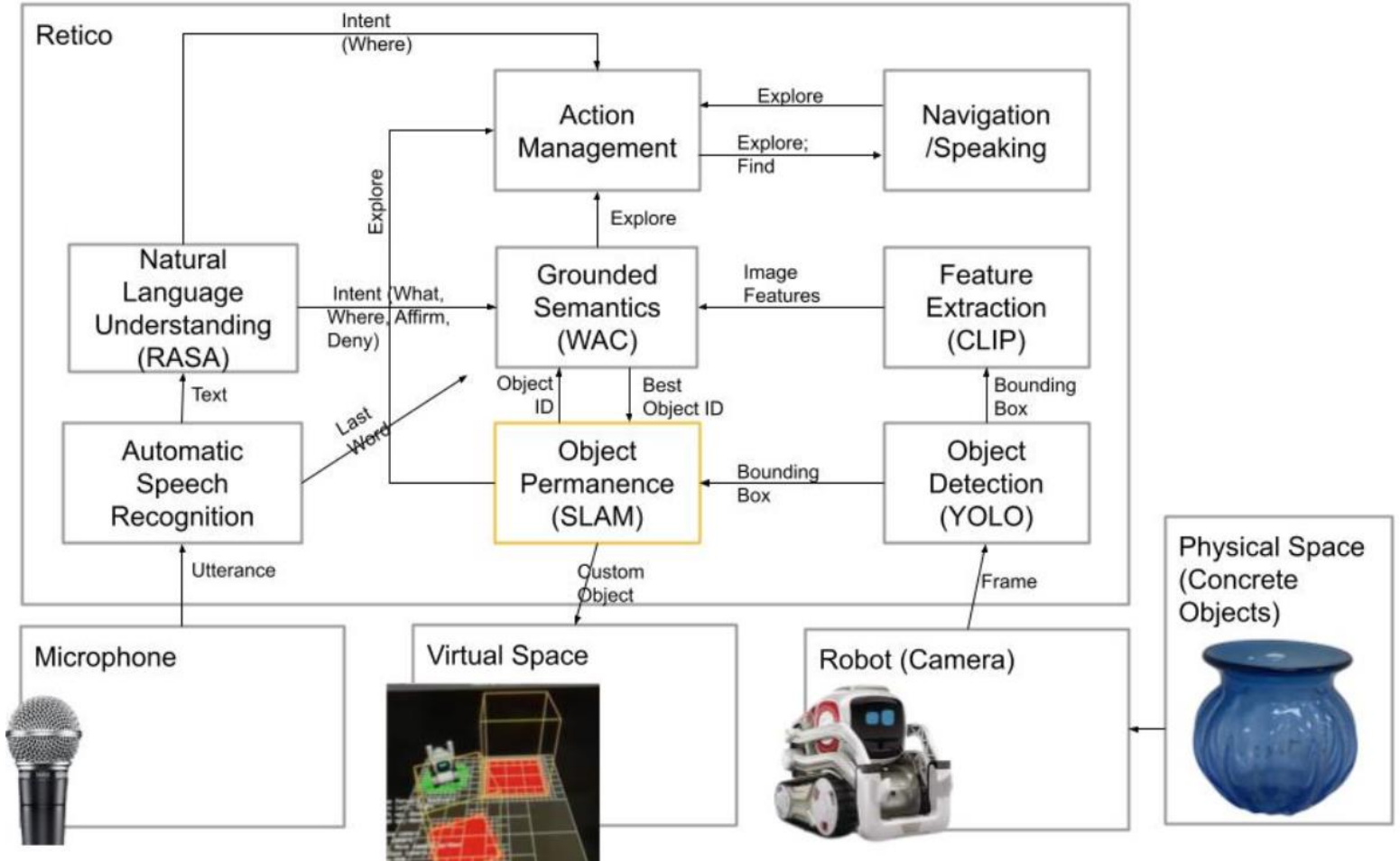
- As infants understand that objects removed from their view exist they start to learn relational words (e.g. above, below, behind) (Tomasello and Farrar, 1984)
- Object permanence is crucial in searching behavior as infants understand that they have the ability to cause hidden objects to reappear (O' Connor and Russell, 2015)
- Platonov et al. (2019) created a Spoken Dialogue System able to create a 3D model of a physical block world and answer spatial questions about it.
- Chai et al. (2014) has shown that collaborative efforts have a positive affect on communicative and symbol grounding

Cozmo

- Cozmo
  - Track for movement
  - Lift
  - Head with OLED Display for Eyes
  - Camera (used for Object Detection)
  - Speech Synthesizer
  - SLAM (Simultaneous Localization and Mapping; used as object permanence)

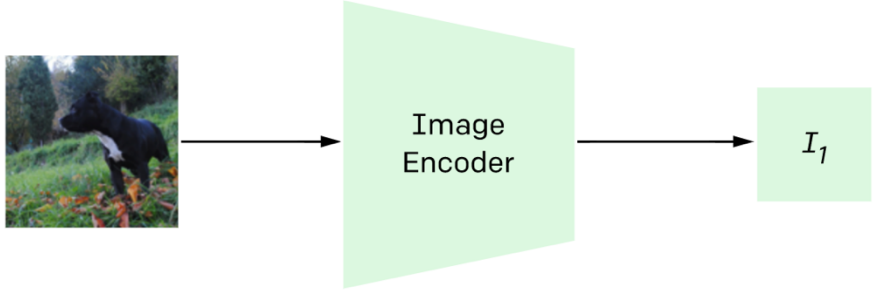
System Overview

- Uses incremental framework ReTiCo(Michael and Möller, 2019; Michael, 2020)
- Extended for Multimodal use using Cozmo (Kennington et al., 2020)

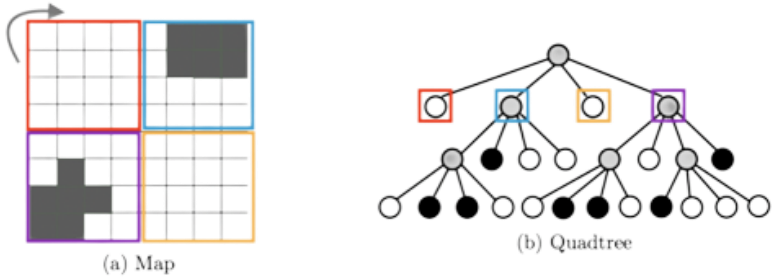


Object Detection & Feature Extraction

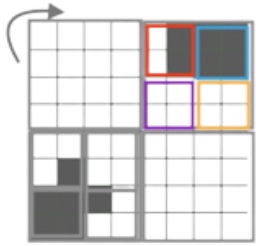
- YOLO (You Only Look Once) (Redmon et al., 2016)
  - Pretrained on MSCOCO (Lin et al, 2014)
  - Applied as a means for object region classification and drawing bounding box around objects
- CLIP (Radford et, al 2021)
  - Neural network trained on (image, text) pairs
  - Takes image and bounding box and extracts sub-image proceeding to pass through CLIP's image encoder and outputs feature vector of size 512

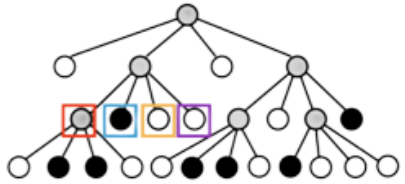


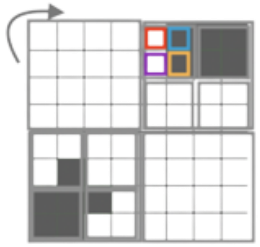
SLAM (object permanence)

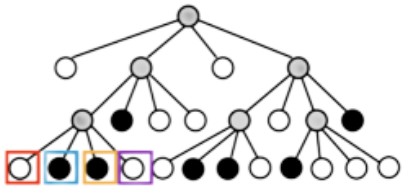

(a) Map

(b) Quadtree

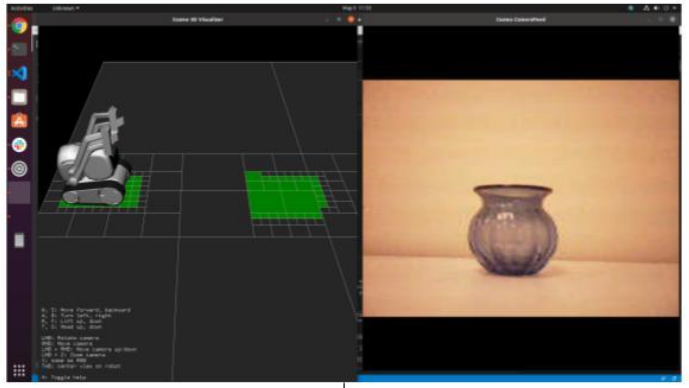




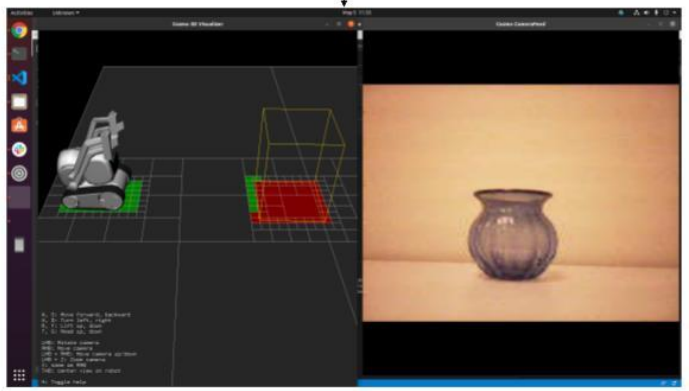




1



2



Grounded Semantics

- Words as Classifiers (Kennington and Schlangen, 2015)

Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution

Casey Kennington
CITEC, Bielefeld University
Universitätsstraße 25
33615 Bielefeld, Germany
ckennington@cit-ec.uni-bielefeld.de

David Schlangen
CITEC, Bielefeld University
Universitätsstraße 25
33615 Bielefeld, Germany
david.schlangen@uni-bielefeld.de

Abstract

An elementary way of using language is to refer to objects. Often, these objects are physically present in the shared environment and reference is done via mention of perceivable properties of the objects. This is a type of language use that is modelled well neither by logical semantics nor by distributional semantics, the former focusing on inferential relations between expressed propositions, the latter on similarity relations between words or phrases. We present an account of word and phrase meaning that is perceptually grounded, trainable, compositional, and 'dialogue-plausible' in that it computes meanings word-by-word. We show that the approach performs well (with an accuracy of 65% on a 1-out-of-32 reference resolution task) on direct descriptions and target/landmark descriptions, even when trained with less than 800 training examples and automatically transcribed utterances.

1 Introduction

The most basic, fundamental site of language use is co-located dialogue (Fillmore, 1975; Clark, 1996) and referring to objects, as in Example (1), is a common occurrence in such a co-located setting.

(1) *The green book on the left next to the mug.*

Logical semantics (Montague, 1973; Gamut, 1991; Partee et al., 1993) has little to say about this process – its focus is on the construction of syntactically manipulable objects that model inferential relations; here, e.g. the inference that there are (at least) two objects. Vector space approaches to distributional semantics (Turney and Pantel, 2010) similarly focuses on something else, namely semantic similarity relations between words or phrases (e.g. finding closeness for "coloured terms on the right of the cup"). Neither approach by itself says anything about processing; typically, the assumption in applications is that fully processed phrases are being processed.

Lacking in these approaches is a notion of grounding of symbols in features of the world (Harnad, 1990).[1] In this paper, we present an account of word and phrase meaning that is (a) perceptually grounded in that it provides a link between words and (computer) vision features of real images, (b) trainable, as that link is learned from examples of language use, (c) compositional in that the meaning of phrases is a function of that of its parts and composition is driven by structural analysis, and (d) 'dialogue-plausible' in that it computes meanings incrementally, word-by-word and can work with noisy input from an automatic speech recogniser (ASR). We show that the approach performs well (with an accuracy of 65% on a reference resolution task out of 32 objects) on direct descriptions as well as target/landmark descriptions, even when trained with little data (less than 800 training examples).

In the following section we will give a background on reference resolution, followed by a description of our model. We will then describe the data we used and explain our evaluations. We finish by giving results, providing some additional analysis, and discussion.

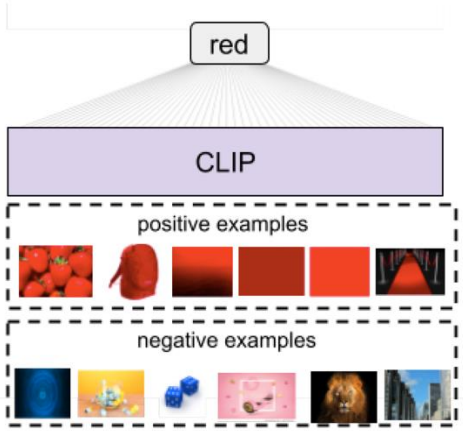
2 Background: Reference Resolution

Reference resolution (RR) is the task of resolving referring expressions (REs; as in Example (1)) to a referent, the entity to which they are intended to refer. Following Kennington et al. (2015a), this can be formalised as a function f_{rr} that, given a representation U of the RE and a representation W

[1] But see discussion below of recent extensions of these approaches taking this into account.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 292–301, Beijing, China, July 26-31, 2015.

- Words are represented as own logistic regression classifier trained on positive and negative examples of real word-referents
- Module "learns" as it hears a word and is observing an object associating words with detected objects (represented as CLIP vectors) as positive examples
- Classifiers trained every time utterance is spoken and after observing an object every 20 added frames
- Two modes: explore (explained above) and exploit (identifies best description for object or best word for object under observation)

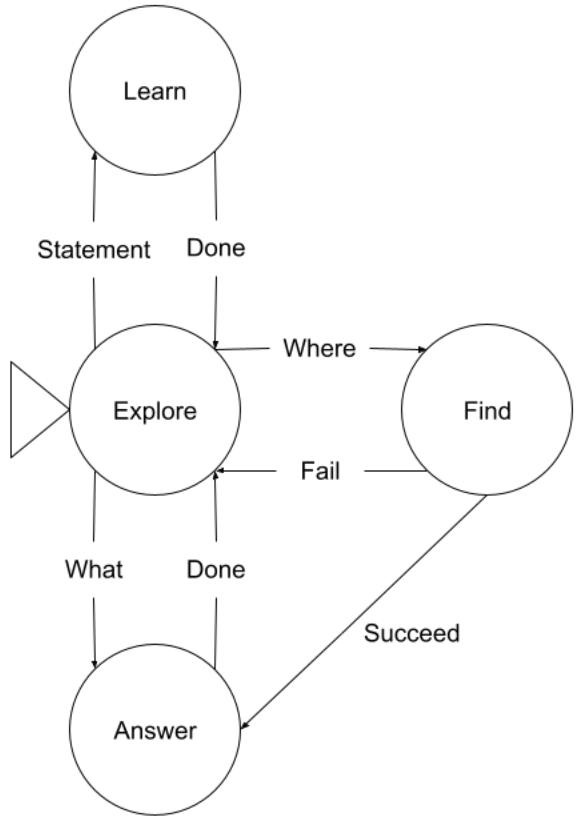


Dialogue Act Recognition

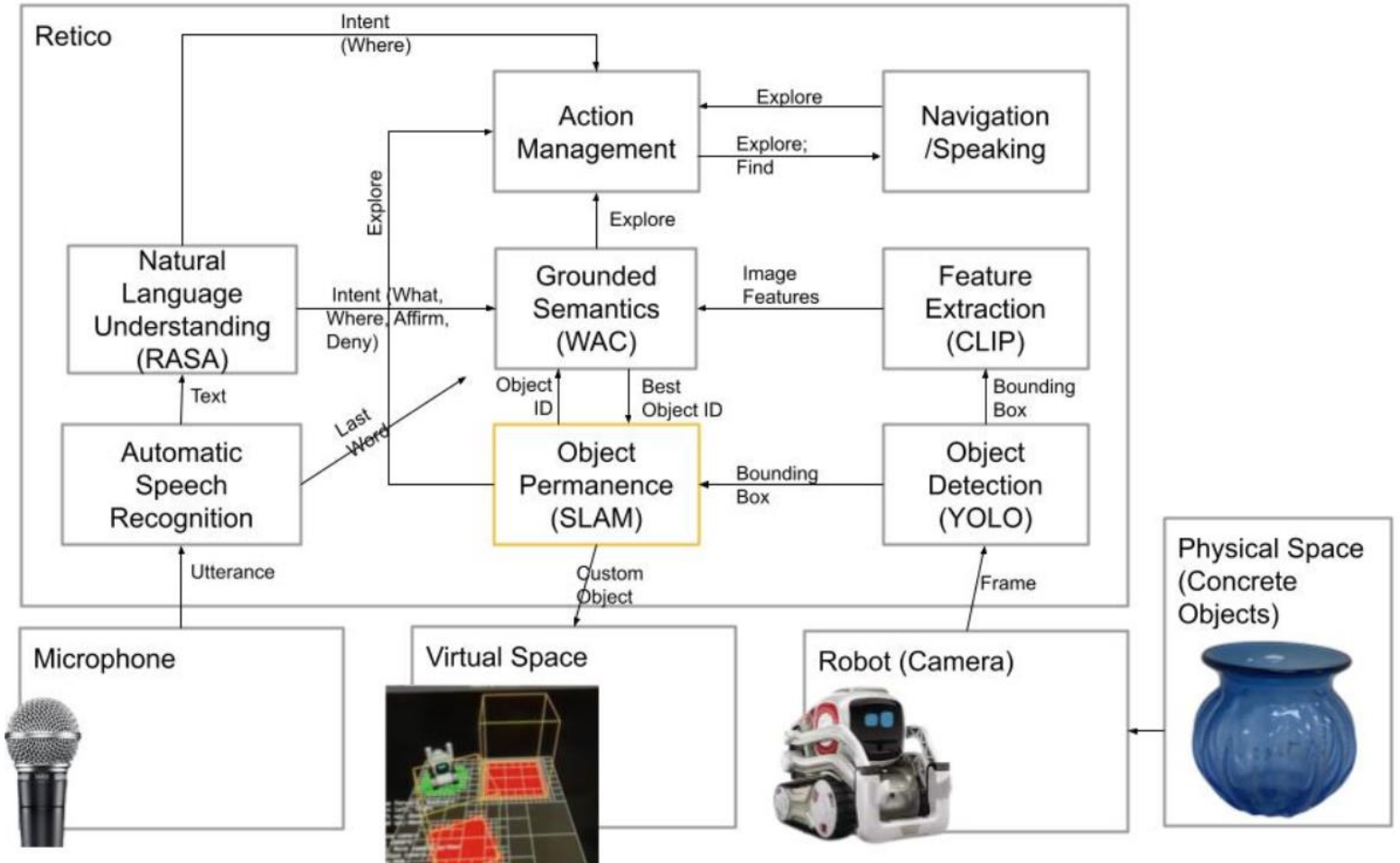
- Google's speech to text API is used for word-level transcription
- Natural Language Understanding (NLU) module takes transcription and determines dialogue act of user using RASA (Bocklisch et al, 2017)
- 5 Dialogue Acts:
  - Positive user feedback (e.g., yes)
  - Negative user feedback (e.g., no)
  - Where questions (e.g., where is the can?)
  - What Questions (e.g., what is that?)
  - Statements (e.g., that is red)

Dialogue Manager and System Task Behavior

- PyOpenDial (Jang et al, 2019)
  - Rule based action movement

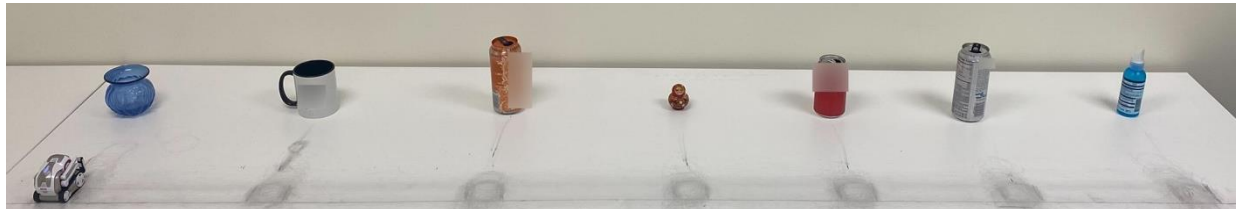


System Overview



Experiment

- Cozmo is placed in front of the leftmost object and a mic is positioned to the left and in front of the leftmost object
- Participants are explained task:
  - Teach Cozmo as many words about the objects as possible periodically quizzing Cozmo's knowledge of the objects (A/B test: 20 min with object permanence and 20 min without)
- After each interaction survey the participant fills out a survey measuring communicative grounding using Godspeed questionnaire (Bartneck et al, 2009)
- 24 participants gathered (18 male; 6 female)



Metrics

- Symbol Grounding
  - Platform for Situated Intelligence (Bohus et al, 2017) logs communication
  - Experimenter tracks number of utterances, positive and negative feedback, and number of questions asked
  - Participants keep track of correct answers (supervised by experimenter to prevent errors)
- Communicative Grounding
  - Asked questions from Godspeed questionnaire (Bartneck et al, 2009)

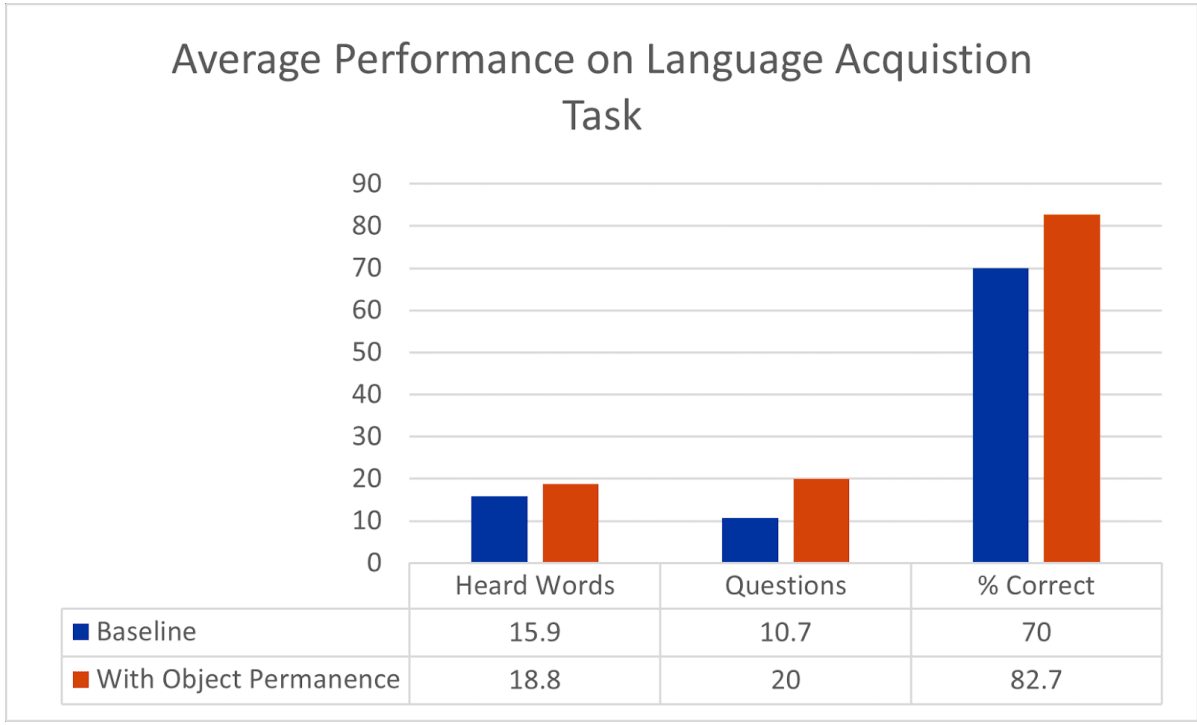
Section	Items		
Anthropomorphism	Fake	-	Natural
	Machinelike	-	Humanlike
	Unconscious	-	Conscious
	Artificial	-	Lifelike
	Moving rigidly	-	Moving elegant
Animacy	Dead	-	Alive
	Stagnant	-	Lively
	Mechanical	-	Organic
	Artificial	-	Lifelike
	Inert	-	Interactive
	Apathetic	-	Responsive
Likeability	Dislike	-	Like
	Unfriendly	-	Friendly
	Unkind	-	Kind
	Unpleasant	-	Pleasant
	Awful	-	Nice
Perceived Intelligence	Incompetent	-	Competent
	Ignorant	-	Knowledgeable
	Irresponsible	-	Responsible
	Unintelligent	-	Intelligent
	Foolish	-	Sensible
Perceived Safety	Anxious	-	Relaxed
	Agitated	-	Calm
	Quiescent	-	Surprised

Metrics Continued

- Perception of User
  - Asked further questions to ascertain user's perception of system and robot:
    - 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)?

Effect of object permanence on language acquisition task

(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

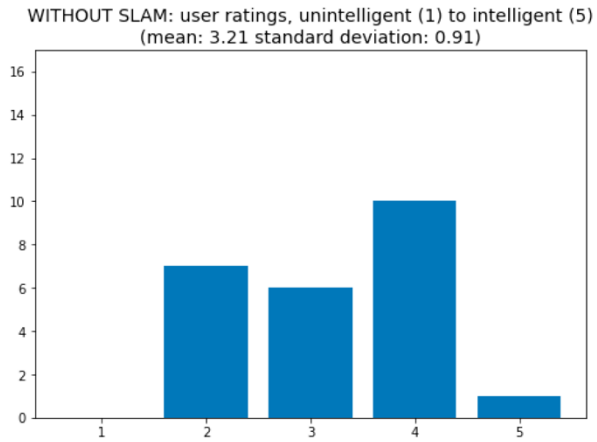


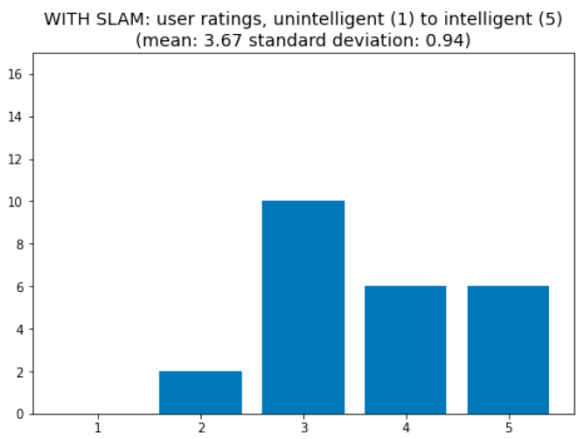
Effect of initial setting on language acquisition task

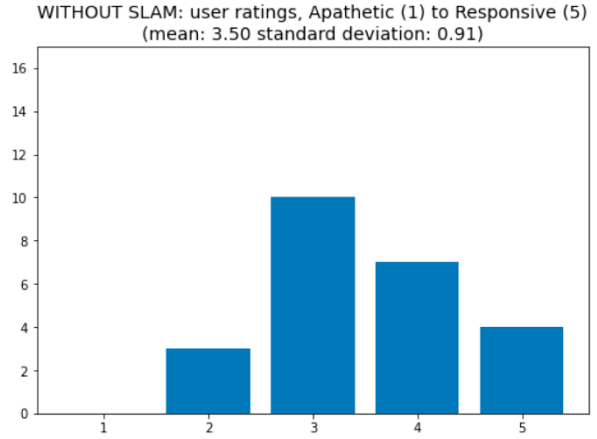
- Numbers in parentheses represent p-value

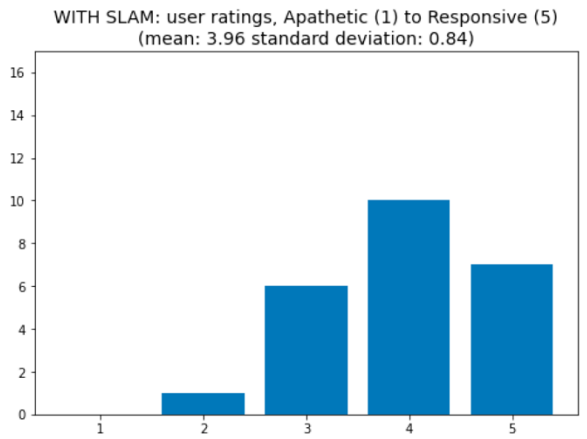
(Mean / std. dev)	1st Interaction (A)	1st Interaction (B)	2nd Interaction (A)	2nd Interaction (B)
Heard Words	19.4 / 4.7	16.2 / 2.7	18.3 / 5.0 (0.56)	15.6 / 4.9 (0.72)
Questions Asked	16.4 / 4.0	12.0 / 3.5	23.0 / 7.5 (0.01)	9.7 / 2.7 (0.13)
% Correct	83.2 / 11.5	74.7 / 20.4	82.1 / 13.0 (0.84)	64.9 /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

User Perceptions: Intelligence and Responsiveness



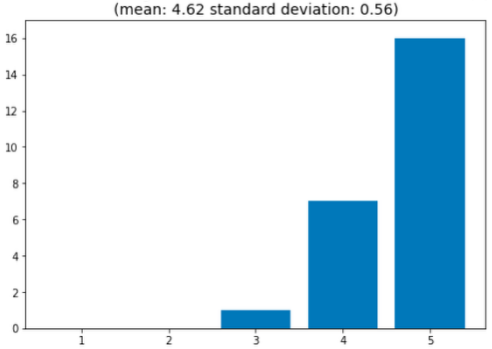




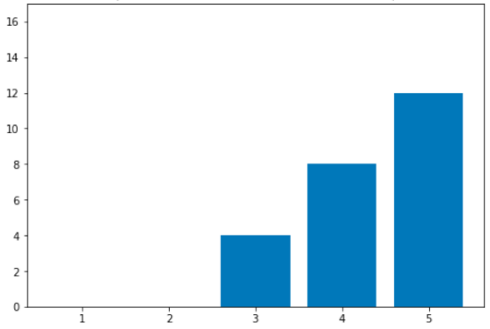


User Perceptions: Engagement

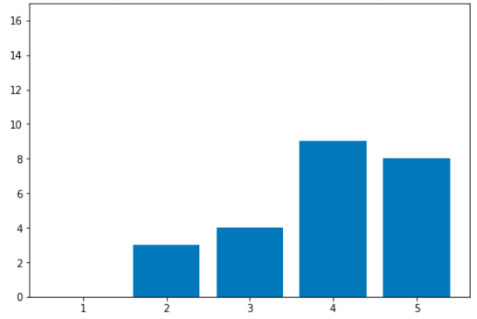
WITH SLAM: How interesting was the robot to interact with?, Not at all (1) to Very (5)
(mean: 4.62 standard deviation: 0.56)



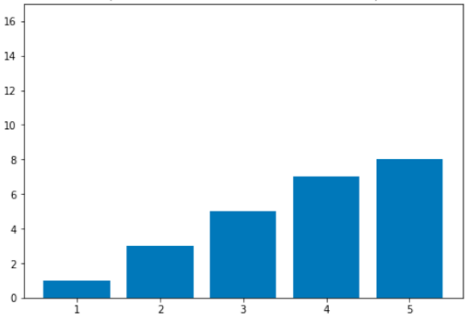
WITH SLAM: Would you like to spend more time with the robot?, Not at all (1) to Very Much (5)
(mean: 4.33 standard deviation: 0.75)



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



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)



Analysis

- Clear that object permeance is perceived to positively affect Cozmo's ability to learn language
  - Cozmo not only hears more words on average per participant with test condition than without, but accuracy also increased by approximately 13%
- Overall, mean values for ratings related to perception were higher for test condition
  - Mean value for test condition 4.7 compared to 4.2 without
  - Cozmo's age estimated at 3.5 years of age with test condition compared to 2.6 without suggesting higher intelligence with object permanence

Conclusion

- We conducted an experiment with 24 participants who performed a language acquisition task with Cozmo both with and without object permanence.
- An understanding of object permanence resulted in improved communicative and symbol grounding seen from stronger engagement and higher percentage of correct answers
- User perception improved with an understanding of object permanence
- Findings suggest that object permanence is a necessary component of any spoken dialogue system
- Future work involves designing same SDS for other robot platforms and exploring how object permanence can help examine others perspective

BOISE STATE UNIVERSITY

SPEECH, LANGUAGE & INTERACTIVE MACHINES

NSF

References