Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and **Raymond J. Mooney**

Department of Computer Science, University of Texas at Austin

Austin, TX 78712, USA

{jesse, jsinapov, maxwell, pstone, mooney}@cs.utexas.edu

Abstract

Grounded language learning bridges words like 'red' and 'square' with robot perception. The vast majority of existing work in this space limits robot perception to vision. In this paper, we build perceptual models that use haptic, auditory, and proprioceptive data acquired through robot exploratory behaviors to go beyond vision. Our system learns to ground natural language words describing objects using supervision from an interactive human-robot "I Spy" game. In this game, the human and robot take turns describing one object among several, then trying to guess which object the other has described. All supervision labels were gathered from human participants physically present to play this game with a robot. We demonstrate that our multi-modal system for grounding natural language outperforms a traditional, vision-only grounding framework by comparing the two on the "I Spy" task. We also provide a qualitative analysis of the groundings learned in the game, visualizing what words are understood better with multi-modal sensory information as well as identifying learned word meanings that correlate with physical object properties (e.g. 'small' negatively correlates with object weight).

1 Introduction

Robots need to be able to connect language to their environment in order to discuss real world objects with humans. Mapping from referring expressions such as "the blue cup" to an object referent in the world is an example of the *symbol grounding problem* [Harnad, 1990]. Symbol grounding involves connecting internal representations of information in a machine to real world data from its sensory perception. *Grounded language learning* bridges these symbols with natural language.

Early work on grounded language learning enabled a machine to map from adjectives and nouns such as "red" and "block" to objects in a scene through vision-based classifiers [Roy, 2001]. We refer to adjectives and nouns that describe properties of objects as language *predicates*. Most
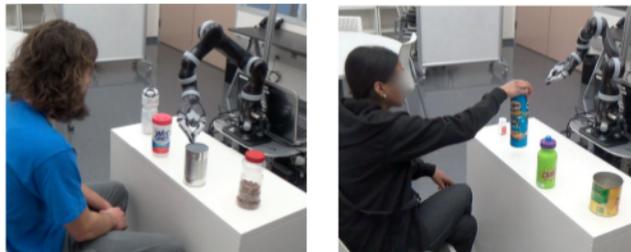
Figure 1: **Left**: the robot guesses an object described by a human participant as "silver, round, and empty." **Right**: a human participant guesses an object described by the robot as "light," "tall," and "tub."

work has focused on grounding predicates through visual information. However, other sensory modalities such as haptic and auditory are also useful in allowing robots to discriminate between object categories [Sinapov *et al.*, 2014b]. This paper explores grounding language predicates by considering visual, haptic, auditory, and proprioceptive senses.

A home or office robot can explore objects in an unsupervised way to gather perceptual data, but needs human supervision to connect this data to language. Learning grounded semantics through human-robot dialog allows a system to acquire the relevant knowledge without the need for laborious labeling of numerous objects for every potential lexical descriptor. A few groups have explored learning from interactive linguistic games such as "I Spy" and "20 Questions" [Parde *et al.*, 2015; Vogel *et al.*, 2010]; however, these studies only employed vision (see Section 2).

We use a variation on the children's game "I Spy" as a learning framework for gathering human language labels for objects to learn multi-modal grounded lexical semantics (Figure 1). Our experimental results test generalization to new objects not seen during training and illustrate both that the system learns accurate word meanings and that modalities beyond vision improve its performance.

To our knowledge, this is the first robotic system to perform natural language grounding using multi-modal sensory perception through feedback with human users.

2 Related Work

Researchers have made substantial progress on grounding language for robots, enabling tasks such as object recognition and route following from verbal descriptions. Early work used vision together with speech descriptions of objects to learn grounded semantics [Roy and Pentland, 2002].

In the past few years, much of this work has focused on combining language with visual information. For grounding referring expressions in an environment, many learn perceptual classifiers for words given some pairing of human descriptions and labeled scenes [Liu *et al.*, 2014; Malinowski and Fritz, 2014; Mohan *et al.*, 2013; Sun *et al.*, 2013; Dindo and Zambuto, 2010; Vogel *et al.*, 2010]. Some approaches additionally incorporate language models into the learning phase [Spranger and Steels, 2015; Krishnamurthy and Kollar, 2013; Perera and Allen, 2013; Matuszek *et al.*, 2012]. Incorporating a language model also allows for more robust generation of robot referring expressions for objects, as explored in [Tellex *et al.*, 2014]. In general, referring expression generation is difficult in dialog [Fang *et al.*, 2014]. Since we are focused on comparing multi-modal to vision-only grounding, our method uses simple language understanding and constructs new predicate classifiers for each unseen content word used by a human playing "I Spy", and our basic generation system for describing objects is based only on these predicate classifiers.

Outside of robotics, there has been some work on combining language with sensory modalities other than vision, such as audio [Kiela and Clark, 2015]. Unlike that line of work, our system is embodied in a learning robot that manipulates objects to gain non-visual sensory experience.

Including a human in the learning loop provides a more realistic learning scenario for applications such as household and office robotics. Past work has used human speech plus gestures describing sets of objects on a table as supervision to learn attribute classifiers [Matuszek *et al.*, 2014; Kollar *et al.*, 2013]. Recent work introduced the "I Spy" game as a supervisory framework for grounded language learning [Parde *et al.*, 2015]. Our work differs from these by using additional sensory data beyond vision to build object attribute classifiers. Additionally, in our instantiation of the "I Spy" task, the robot and the human both take a turn describing objects, where in previous work [Parde *et al.*, 2015] only humans gave descriptions.

3 Dataset

The robot used in this study was a Kinova MICO arm mounted on top of a custom-built mobile base which remained stationary during our experiment. The robot's perception included joint effort sensors in each of the robot arm's motors, a microphone mounted on the mobile base, and an Xtion ASUS Pro RGBD camera. The set of objects used in this experiment consisted of 32 common household items including cups, bottles, cans, and other containers, shown in Figure 2. Some of the objects contained liquids or other contents (e.g., coffee beans) while others were empty. Contemporary work gives a more detailed description of this object

Figure 2: Objects used in the "I Spy" game divided into the four folds discussed in Section 6.1, from fold 0 on the left to fold 3 on the right.


grasp   lift   lower

drop   press   push
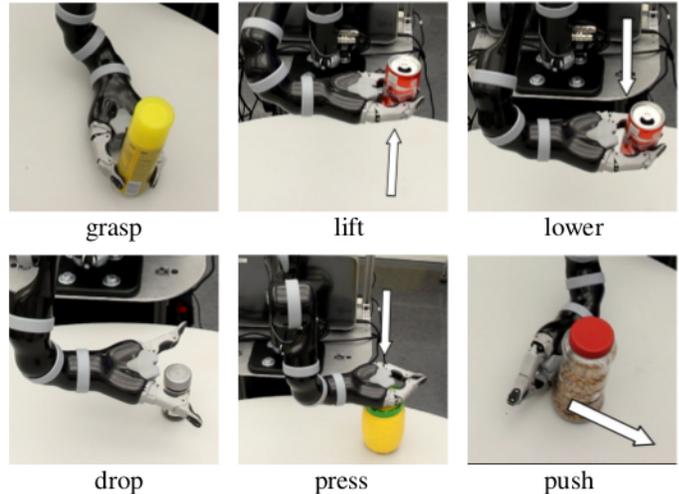
Figure 3: The behaviors the robot used to explore the objects. The arrows indicate the direction of motion of the end-effector for each behavior. In addition, the *hold* behavior (not shown) was performed after the *lift* behavior by simply holding the object in place for half a second.

dataset [Sinapov *et al.*, 2016], but we briefly describe the exploration and modalities below.

3.1 Exploratory Behaviors and Sensory Modalities

Prior to the experiment, the robot explored the objects using the methodology described by Sinapov et al. [2014a], and the dimensionality of the raw auditory, haptic, and proprioceptive data were reduced comparably (final dimensionality given in Table 1). In our case, the robot used 7 distinct actions: *grasp*, *lift*, *hold*, *lower*, *drop*, *push*, and *press*, shown in Figure 3. During the execution of each action, the robot recorded the sensory perceptions from *haptic* (i.e., joint efforts) and *auditory* sensory modalities. During the *grasp* action, the robot recorded *proprioceptive* (i.e., joint angular positions) sensory information from its fingers. The joint efforts and joint positions were recorded for all 6 joints at 15 Hz. The auditory sensory modality was represented as the Discrete Fourier Transform computed using 65 frequency bins.

In addition to the 7 interactive behaviors, the robot also performed the *look* action prior to grasping the object which produced three different kinds of sensory modalities: 1) an RGB color histogram of the object using 8 bins per channel; 2)

Behavior	Modality		
	color	fpfh	vgg
look	64	308	4096
	audio	haptics	proprioception
grasp	100	60	20
drop, hold, lift, lower, press, push	100	60	

Table 1: The number of features extracted from each *context*, or combination of robot behavior and perceptual modality.

Fast point feature histogram (*fpfh*) shape features [Rusu *et al.*, 2009] as implemented in the Point Cloud Library [Aldoma *et al.*, 2012]; and 3) deep visual features from the 16-layer VGG network [Simonyan and Zisserman, 2014]. The first two types of features were computed using the segmented point cloud of the object while the deep features were computed using the 2D image of the object.

Thus, each of the robot's 8 actions produced two to three different kinds of sensory signals. Each viable combination of an action and a sensory modality is a unique sensorimotor context. In our experiment, the set of contexts $\mathcal{C}$ was of size $2 \times 3 + 6 \times 2 = 18$. The robot performed its full sequence of exploratory actions on each object 5 different times (for the *look* behavior, the object was rotated to a new angle each time). Given a context $c \in \mathcal{C}$ and an object $i \in \mathcal{O}$, let the set $\mathcal{X}_i^c$ contain all five feature vectors observed with object i in context c.

4 Task Definition

In our "I Spy" task,[1] the human and robot take turns describing objects from among 4 on a tabletop (Figure 1). Participants were asked to describe objects using attributes. As an example, we suggested participants describe an object as "black rectangle" as opposed to "whiteboard eraser." Additionally, participants were told they could handle the objects physically before offering a description, but were not explicitly asked to use non-visual predicates. Once participants offered a description, the robot guessed candidate objects in order of computed confidence (see Section 5.2) until one was confirmed correct by the participant.

In the second half of each round, the robot picked an object and then described it with up to three predicates (see Section 5.2). The participant was again able to pick up and physically handle objects before guessing. The robot confirmed or denied each participant guess until the correct object was chosen.

"I Spy" gameplay admits two metrics. The **robot guess** metric is the number of turns the robot took to guess what object the participant was describing. The **human guess** metric is the complement. Using these metrics, we compare the performance of two "I Spy" playing systems (**multi-modal** and **vision-only**) as described in Section 6. We also compare

the agreement between both systems' predicate classifiers and human labels acquired during the game.

5 Implementation

To play "I Spy", we first gathered sensory data from the set of objects through robot manipulation behaviors (described in Section 3). When playing a game, the robot was given unique identifying numbers for each object on the table and could look up relevant feature vectors when performing grounding.

During the course of the game, the robot used its RGBD camera to detect the locations of the objects and subsequently detect whenever a human reached out and touched an object in response to the robot's turn. The robot could also reach out and point to an object when guessing. We implemented robot behaviors in the Robot Operating System[2] and performed text-to-speech using the Festival Speech Synthesis System.[3]

5.1 Multi-Modal Perception

For each language predicate p, a classifier G_p was learned to decide whether objects possessed the attribute denoted by p. This classifier was informed by context sub-classifiers that determined whether p held for subsets of an object's features.

The feature space of objects was partitioned by context, as discussed in Section 3.1. Each context classifier $M_c, c \in \mathcal{C}$ was a quadratic-kernel SVM trained with positive and negative labels for context feature vectors derived from the "I Spy" game (Section 5.2). We defined $M_c(\mathcal{X}_i^c) \in [-1, 1]$ as the average classifier output over all observations for object $i \in \mathcal{O}$ (individual SVM decisions on observations were in $\{-1, 1\}$).

Following previous work in multi-modal exploration [Sinapov *et al.*, 2014b], for each context we calculated Cohen's Kappa $\kappa_c \in [0, 1]$ to measure the agreement across observations between the decisions of the M_c classifier and the ground truth labels from the "I Spy" game.[4] Given these context classifiers and associated κ confidences, we calculate an overall decision, $G_p(i)$, for $i \in \mathcal{O}$ for each behavior b and modality m as:

$$G_p(i) = \sum_{c \in \mathcal{C}} \kappa_c M_c(\mathcal{X}_i^c) \in [-1, 1] \qquad (1)$$

The sign of $G_p(i)$ gives a decision on whether p applies to i with confidence $|G_p(i)|$.

For example, a classifier built for 'fat'$\in P$ could give $G_{\texttt{fat}}(\texttt{wide-yellow-cylinder}) = 0.137$, a positive classification, with $\kappa_{gr,au} = 0.515$ for the *gr*asp behavior's *au*ditory modality, the most confident context. This context could be useful for this predicate because the sound of the fingers' motors stop sooner for wider objects.

5.2 Grounded Language Learning

Language predicates and their positive/negative object labels were gathered through human-robot dialog during the "I Spy" game. The human participant and robot were seated at opposite ends of a small table. A set of 4 objects were placed on the table for both to see (Figure 1). We denote the set of objects on the table during a given game $\mathcal{O}_T$.

Human Turn. On the participant's turn, the robot asked him or her to pick an object and describe it in one phrase. We used a standard stopword list to strip out non-content words from the participant's description. The remaining words were treated as a set of language predicates, $\mathcal{H}_p$. The robot assigned scores S to each object $i \in \mathcal{O}_T$ on the table.

$$S(i) = \sum_{p \in \mathcal{H}_p} G_p(i) \tag{2}$$

The robot guessed objects in descending order by score (ties broken randomly) by pointing at them and asking whether it was correct. When the correct object was found, it was added as a positive training example for all predicates $p \in \mathcal{H}_p$ for use in future training.

Robot Turn. On the robot's turn, an object was chosen at random from those on the table. To describe the object, the robot scored the set of known predicates learned from previous play. Following Gricean principles [Grice, 1975], the robot attempted to describe the object with predicates that applied but did not ambiguously refer to other objects. We used a predicate score R that rewarded describing the chosen object i^* and penalized describing the other objects on the table.

$$R(p) = |O_T| G_p(i^*) - \sum_{j \in \mathcal{O}_T \setminus \{i^*\}} G_p(j) \tag{3}$$

The robot choose up to three highest scoring predicates $\hat{P}$ to describe object i^*, using fewer if $S < 0$ for those remaining. Once ready to guess, the participant touched objects until the robot confirmed that they had guessed the right one (i^*).

The robot then pointed to i^* and engaged the user in a brief follow-up dialog in order to gather both positive and negative labels for i^*. In addition to predicates $\hat{P}$ used to describe the object, the robot selected up to $5 - |\hat{P}|$ additional predicates $\bar{P}$. $\bar{P}$ were selected randomly with $p \in P \setminus \hat{P}$ having a chance of inclusion proportional to $1 - |G_p(i^*)|$, such that classifiers with low confidence in whether or not p applied to i^* were more likely to be selected. The robot then asked the participant whether they would describe the object i^* using each $p \in \hat{P} \cup \bar{P}$. Responses to these questions provided additional positive/negative labels on object i^* for these predicates for use in future training.

6 Experiment

To determine whether multi-modal perception helps a robot learn grounded language, we had two different systems play "I Spy" with 42 human participants. The baseline **vision only** system used only the *look* behavior when grounding language predicates, analogous to many past works as discussed in Section 2. Our **multi-modal** system used the full suite of behaviors and associated haptic, proprioceptive, and auditory modalities shown in Table 1 when grounding language predicates.

6.1 Methodology

Data Folds. We divided our 32-object dataset into 4 folds. For each fold, at least 10 human participants played "I Spy" with both the **vision only** and **multi-modal** systems (12 participants in the final fold). Four games were played by each participant. The **vision only** system and **multi-modal** system were each used in 2 games, and these games' temporal order was randomized. Each system played with all 8 objects per fold, but the split into 2 groups of 4 and the order of objects on the table were randomized.

For fold 0, the systems were undifferentiated and so only one set of 2 games was played by each participant. For subsequent folds, the systems were incrementally trained using labels from previous folds only, such that the systems were always being tested against novel, unseen objects. This contrasts prior work using the "I Spy" game [Parde *et al.*, 2015], where the same objects were used during training and testing.

Human Participants. Our 42 participants were undergraduate and graduate students as well as some staff at our university.

At the beginning of each trial, participants were shown an instructional video of one of the authors playing a single game of "I Spy" with the robot, then given a sheet of instructions about the game and how to communicate with the robot. In every game, participants took one turn and the robot took one turn.

To avoid noise from automatic speech recognition, a study coordinator remained in the room and transcribed the participant's speech to the robot from a remote computer. This was done discretely and not revealed to the participant until debriefing when the games were over.

6.2 Quantitative Results

To determine whether our **multi-modal** approach outperformed a traditional **vision only** approach, we measured the average number of robot guesses and human guesses in games played with each fold of objects. The systems were identical in fold 0 since both were untrained. In the end, we trained the systems on all available data to calculate predicate classifier agreement with human labels.

Robot guess. Figure 4 shows the average number of robot guesses for the games in each fold. Because we had access to the scores the robot assigned each object, we calculated the *expected* number of robot guesses for each turn. For example, if all 4 objects were tied for first, the expected number of robot guesses for that turn was 2.5, regardless of whether it got (un)lucky and picked the correct object (last)first.[5]

After training on just one fold, our **multi-modal** approach performs statistically significantly better than the expected number of turns for guessing (the strategy for the untrained fold 0 system) for the remainder of the games. The **vision only** system, by contrast, is never able to differentiate itself

---

[5] 2.5 is the expected number for 4 tied objects because the probability of picking in any order is equal, so the expected turn to get the correct object is $\frac{1+2+3+4}{4} = \frac{10}{4} = 2.5$
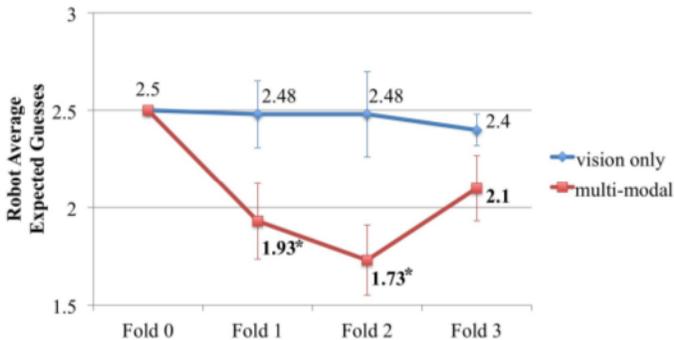
Figure 4: Average expected number of guesses the robot made on each human turn with standard error bars shown. **Bold**: significantly lower than the average at fold 0 with $p < 0.05$ (unpaired Student's t-test). *: significantly lower than the competing system on this fold on participant-by-participant basis with $p < 0.05$ (paired Student's t-test).

Metric	System	
	vision only	multi-modal
precision	.250	.378+
recall	.179	.348*
F_1	.196	.354*

Table 2: Average performance of predicate classifiers used by the **vision only** and **multi-modal** systems in leave-one-object-out cross validation. *: significantly greater than competing system with $p < 0.05$. +: $p < 0.1$ (Student's un-paired t-test).

significantly from random guessing, even as more training data becomes available. We suspect the number of objects is too small for the **vision only** system to develop decent models of many predicates, whereas **multi-modal** exploration allows that system to extract more information per object.

Human guess. Neither the **vision only** nor **multi-modal** system's performance improves on this metric with statistical significance as more training data is seen. Human guesses hovered around 2.5 throughout all levels of training and sets of objects.

This result highlights the difficulty of the robot's turn in an "I Spy" framework, which requires not just good coverage of grounded words (as when figuring out what object the human is describing), but also high accuracy when using classifiers on new objects. Context classifiers with few examples could achieve confidence $\kappa = 1$, making the predicates they represented more likely to be chosen to describe objects. It is possible that the system would have performed better on this metric if the predicate scoring function R additionally favored predicates with many examples.

Predicate Agreement. Training the predicate classifiers using leave-one-out cross validation over objects, we calculated the average precision, recall, and F_1 scores of each against human predicate labels on the held-out object. Table 2 gives these metrics for the 74 predicates used by the systems.[6]

Across the objects our robot explored, our **multi-modal** system achieves consistently better agreement with human assignments of predicates to objects than does the **vision only** system.

6.3 Qualitative Results

We explored the predicates learned by our systems qualitatively by looking at the differences in individual predicate classifier agreements, the objects picked out by these clas-

---

[6]There were 53 predicates shared between the two systems. The results in Table 2 are similar for a paired t-test across these shared predicates with slightly reduced significance.

sifiers in each system, and correlations between predicate decisions and physical properties of objects.

When multi-modal helps. We performed a pairwise comparison of predicates built in the **multi-modal** and **vision only** systems, again using leave-one-out cross validation over objects to measure performance. Table 3 shows the predicates for which the difference in f_1 between the two systems was high.

The **multi-modal** system does well on the predicates "tall" and "half-full" which have non-visual interpretations. A tall object will exert force earlier against the robot arm pressing down on it, while a half-full object will be lighter than a full one and heavier than an empty one. The color predicate "pink" seems to confuse the multi-modal grounding system using non-visual information for this purely visual predicate. This doesn't hold for "yellow", though the classifiers for "yellow" never became particularly good for either system. For example, two of the three most confident objects in the multi-modal setting are false positives.

Correlations to physical properties. To validate whether the systems learned non-visual properties of objects, for every predicate we calculated the Pearson's correlation r between its decision on each object and that object's measured weight, height, and width. As before, the decisions were made on held-out objects in leave-one-out cross validation. We found predicates for which $r > 0.5$ with $p < 0.05$ when the system had at least 10 objects with labels for the predicate on which to train.

The **vision only** system led to no predicates correlated against these physical object features.

The **multi-modal** system learned to ground predicates which correlate well to objects' height and weight. The "tall" predicate correlates with objects that are higher ($r = .521$), "small" ($r = -.665$) correlates with objects that are lighter, and "water" ($r = .814$) correlates with objects that are heavier. The latter is likely from objects described as "water bottle", which, in our dataset, are mostly filled either half-way or totally and thus heavier. There is also a spurious correlation between "blue" and weight ($r = .549$). This highlights the value of multi-modal grounding, since words like "half-full" cannot be evaluated with vision alone when dealing with closed containers that have unobservable contents.

7 Conclusion

We expand past work on grounding natural language in robot sensory perception by going beyond vision and exploring

Predicate	$f_1^{mm} - f_1^{vo}$	High Confidence Positive	High Confidence Negative
		multi-modal system	
can	0.857		
tall	0.516		
half-full	.462		
yellow	.312		
		vision only system	
pink	-.3		

Table 3: Predicates for which the difference $|f_1^{mm} - f_1^{vo}|$ between the **multi-modal** (mm) and **vision only** (vo) systems was greater than or equal to 0.3, both systems had at least 10 objects with labels for that predicate on which to train, and the system with the worse f_1 had at most 5 fewer objects with labels on which to train (to avoid rewarding a system just for having more training labels). The highest- and lowest-confidence objects for each predicate are shown. The top rows ($f_1^{mm} - f_1^{vo} > 0$) are decisions from the **multi-modal** system, the bottom row from the **vision only** system.

haptic, auditory, and proprioceptive robot senses. We compare a vision only grounding system to one that uses these additional senses by employing an embodied robot playing "I Spy" with many human users. To our knowledge, ours is the first robotic system to perform natural language grounding using multi-modal sensory perception through natural interaction with human users.

We demonstrate quantitatively, through the number of turns the robot needs to guess objects described by humans, as well as through agreement with humans on language predicate labels for objects, that our multi-modal framework learns more effective lexical groundings than one using vision alone. We also explore the learned groundings qualitatively, showing words for which non-visual information helps most as well as when non-visual properties of objects correlate with learned meanings (e.g. "small" correlates negatively with object weight).

In the future, we would like to use one-class classification methods [Liu et al., 2003] to remove the need for a follow-up dialog asking about particular predicates applied to an object to gather negative labels. Additionally, we would like to detect polysemy for predicates whose meanings vary across sensory modalities. For example, the word "light" can refer to weight or color. Our current system fails to distinguish these senses, while human participants intermix them. Additionally, in our current system, the robot needs to explore objects in advance using all of its behaviors. However, for purely visual predicates like "pink" and other colors, only the *look* behavior is necessary to determine whether an object has the property. We will work towards an exploration system that uses its learned knowledge of predicates from a game such as "I Spy" to determine the properties of a novel object while attempting to use as few exploratory behaviors as necessary.

References

[Aldoma et al., 2012] Aitor Aldoma, Zoltan-Csaba Marton, Federico Tombari, Walter Wohlkinger, Christian Potthast, Bernhard Zeisl, Radu Bogdan Rusu, Suat Gedikli, and Markus Vincze. Point cloud library. *IEEE Robotics & Automation Magazine*, 1070(9932/12), 2012.

[Dindo and Zambuto, 2010] Haris Dindo and Daniele Zambuto. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *International Conference on Intelligent Robots and Systems*, pages 760–796, Taipei, Taiwan, 2010. IEEE.

[Fang et al., 2014] Rui Fang, Malcolm Doering, and Joyce Y. Chai. Collaborative models for referring expression generation towards situated dialogue. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1544–1550, 2014.

[Grice, 1975] H. Paul Grice. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics*

3: Speech Acts, pages 41–58. Academic Press, New York, 1975.

[Harnad, 1990] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[Kiela and Clark, 2015] Douwe Kiela and Stephen Clark. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Emperical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal, 2015.

[Kollar *et al.*, 2013] Thomas Kollar, Jayant Krishnamurthy, and Grant Strimel. Toward interactive grounded language acquisition. In *Robotics: Science and Systems*, 2013.

[Krishnamurthy and Kollar, 2013] Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013.

[Liu *et al.*, 2003] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, , and Philip Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03)*, 2003.

[Liu *et al.*, 2014] Changson Liu, Lanbo She, Rui Fang, and Joyce Y. Chai. Probabilistic labeling for efficient referential grounding based on collaborative discourse. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 13–18, Baltimore, Maryland, USA, 2014.

[Malinowski and Fritz, 2014] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pages 13–18, Montréal, Canada, 2014.

[Matuszek *et al.*, 2012] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

[Matuszek *et al.*, 2014] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebéc City, Quebéc, Canada, 2014.

[Mohan *et al.*, 2013] Shiwali Mohan, Aaron H. Mininger, and John E. Laird. Towards an indexical model of situated language comprehension for real-world cognitive agents. In *Proceedings of the 2nd Annual Conference on Advances in Cognitive Systems*, Baltimore, Maryland, USA, 2013.

[Parde *et al.*, 2015] Natalie Parde, Adam Hair, Michalis Papakostas, Konstantinos Tsiakas, Maria Dagioglou, Vangelis Karkaletsis, and Rodney D. Nielsen. Grounding the meaning of words through vision and interactive gameplay. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 1895–1901, Buenos Aires, Argentina, 2015.

[Perera and Allen, 2013] Ian Perera and James F. Allen. Salle: Situated agent for language learning. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 1241–1247, Bellevue, Washington, USA, 2013.

[Roy and Pentland, 2002] Deb Roy and Alex Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113–146, 2002.

[Roy, 2001] Deb Roy. Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication*, 4(1), 2001.

[Rusu *et al.*, 2009] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. IEEE, 2009.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[Sinapov *et al.*, 2014a] Jivko Sinapov, Connor Schenck, Kerrick Staley, Vladimir Sukhoy, and Alexander Stoytchev. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems*, 62(5):632–645, 2014.

[Sinapov *et al.*, 2014b] Jivko Sinapov, Connor Schenck, and Alexander Stoytchev. Learning relational object categories using behavioral exploration and multimodal perception. In *IEEE International Conference on Robotics and Automation*, 2014.

[Sinapov *et al.*, 2016] Jivko Sinapov, Priyanka Khante, Maxwell Svetlik, and Peter Stone. Learning to order objects using haptic and proprioceptive exploratory behaviors. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[Spranger and Steels, 2015] Michael Spranger and Luc Steels. Co-acquisition of syntax and semantics — an investigation of spatial language. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 1909–1915, Buenos Aires, Argentina, 2015.

[Sun *et al.*, 2013] Yuyin Sun, Liefeng Bo, and Dieter Fox. Attribute based object identification. In *International Conference on Robotics and Automation*, pages 2096–2103, Karlsruhe, Germany, 2013. IEEE.

[Tellex *et al.*, 2014] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. *Proceedings of Robotics: Science and Systems*, 2014.

[Vogel *et al.*, 2010] Adam Vogel, Karthik Raghunathan, and Dan Jurafsky. Eye spy: Improving vision through dialog. In *Association for the Advancement of Artificial Intelligence*, pages 175–176, 2010.