# Collaborative Language Grounding Toward Situated Human-Robot Dialogue

*Joyce Y. Chai, Rui Fang, Changsong Liu, Lanbo She*

■ *To enable situated human-robot dialogue, techniques to support grounded language communication are essential. One particular challenge is to ground human language to a robot's internal representation of the physical world. Although copresent in a shared environment, humans and robots have mismatched capabilities in reasoning, perception, and action. Their representations of the shared environment and joint tasks are significantly misaligned. Humans and robots will need to make extra effort to bridge the gap and strive for a common ground of the shared world. Only then is the robot able to engage in language communication and joint tasks. Thus computational models for language grounding will need to take collaboration into consideration. A robot not only needs to incorporate collaborative effort from human partners to better connect human language to its own representation, but also needs to make extra collaborative effort to communicate its representation in language that humans can understand. To address these issues, the Language and Interaction Research group (LAIR) at Michigan State University has investigated multiple aspects of collaborative language grounding. This article gives a brief introduction to this research effort and discusses several collaborative approaches to grounding language to perception and action.*

A new generation of cognitive robots has emerged in recent years to provide service, care, and companionship to humans. To support natural dialogue between a human and these robots, technology enabling grounded language communication has become increasingly important. During human-robot dialogue, given a human utterance, the robot needs to understand what objects and activities in its representation of the perceived world the human is talking about. When the human issues a language command, the robot needs to identify a sequence of low-level operations to execute that command. To support these capabilities, *language grounding* addresses the problem of connecting human language to robot internal representations of the world, which include the robot's knowledge of the world, the perception of the environment, and the control of actions that can change the world. Only when such grounding is made is the robot able to understand human language, follow language instructions, and communicate with humans in language.

However, grounding language to the robot internal sensorimotor representations is extremely challenging. For exam-

  ISSN 0738-4602

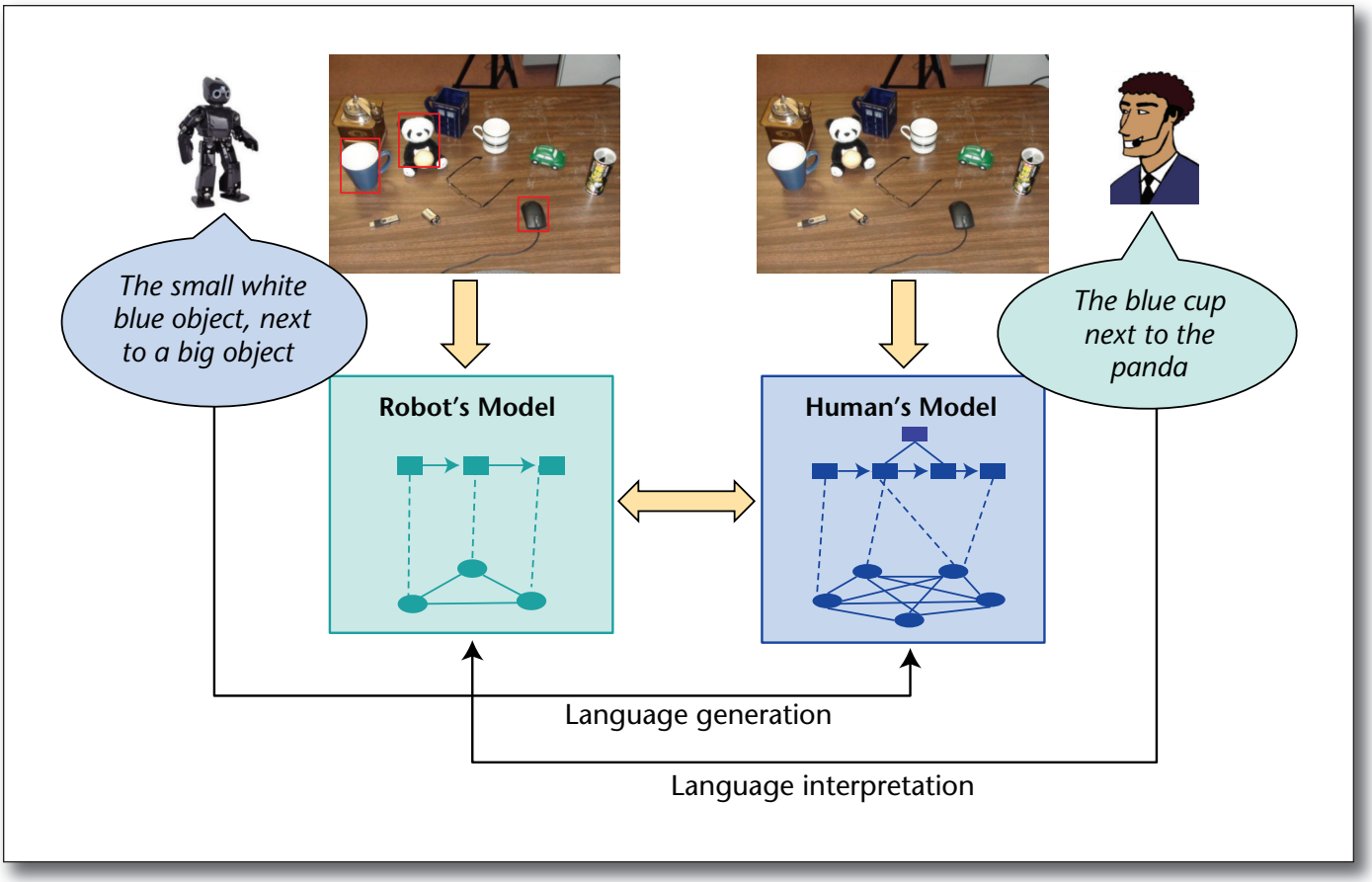



*Figure 1. Collaborative Language Grounding to Establish a Joint Perceptual Basis.*

ple, grounding language to perception requires the robot to process perceived visual signals and infer high-level concepts and structures (for example, objects, actions, and others). How to ground linguistic expressions to the underlying numerical visual features and how to acquire and learn grounded semantic models become critical questions. Similarly, grounding language to actions requires the robot to make connections between high-level concepts (for example, expressed by a verb) and the low-level robotic control system, which is often programmed to control only a set of primitive operations (for example, move to, open gripper, and close gripper for a robotic arm). How to acquire and learn grounded meanings of concepts that can support automatic planning for primitive robotic operations is another critical question.

Grounding language to perception and action is further complicated by the lack of common ground in representing the shared world. Compared to human partners, robots have mismatched capabilities in perception and reasoning. As shown in figure 1, although a human and a robot are copresent in a shared physical world, their representations of the shared environment and joint tasks are significantly misaligned. The human, who has higher capabilities, may have enriched representations, while the robot with lower capabilities may only have impoverished representations of the same world. Thus, although they are copresent, they don't share joint perceptual experience, which will jeopardize the common ground of the shared environment. Communication between humans and robots will become difficult. For example, the human may specify "the blue cup next to the panda," but the robot may not be able to understand which object in its representation is being referred to if the cup cannot be recognized correctly. This is a typical problem of grounded language interpretation. Similarly, for language generation, suppose the robot wants to refer to the blue cup (but cannot recognize it correctly). The robot may use language "the small white blue object, next to a big object." It will be difficult for the human to understand which object in his/her representation is being talked about. Because of these difficulties, language grounding often cannot be succeeded by one attempt, but rather is achieved by a collaborative process between humans and robots based on multiple iterations (for example, through refashion, clarification, and others). Therefore, algorithms for language grounding in human-robot dialogue will need to take collaboration into account.

To address these issues, the Language and Interaction Research (LAIR) group at Michigan State University has investigated several aspects of collaborative language grounding toward situated human-robot dialogue. This article gives an overview of this research effort. It starts with a simulation experiment to examine collaborative effort in mediating a shared perceptual basis between human partners with mismatched perceptual capabilities. It then describes approaches that incorporate collaboration to ground language to perception as motivated by observations from human communication studies. It further shares two experiments that demonstrate the advantage of collaborative effort from the robot in language grounding to mitigate perceptual differences. In addition to discussing grounding language to perception, this article also provides a brief introduction to grounding language to action, focusing on representations of grounded verb semantics and their acquisitions through collaborative step-by-step language instructions.

## Collaborative Effort in Mediating Shared Perceptual Basis

In conversation, participants coordinate their mental states based on their mutual understanding of their intention, goals, and current tasks (Austin 1962, Grice 1975, Clark 1996). An important notion, which is also critical to the success of communication, is common ground. According to Clark (1996), "two people's common ground is, in effect, the sum of their mutual, common, or joint knowledge, beliefs, and suppositions." What allows participants to successfully establish common ground largely depends on the shared bases: their joint experiences, events, or episodes. One type of shared basis, which is extremely important for situated communication, is the perceptual basis, which describes joint perceptual experiences. Clark (1996) explains that:

> Each of us lives in a world of perceptible things, entities we can look at, feel, hear, smell, taste. At any moment, we have perceptual access, with more or less effort, to only part of that world, our perceptual shell. You and I have distinct perceptual shells, but when we are together, they overlap. But having overlapping perceptual shells isn't sufficient for perceptual copresence. You and I manage to attend to the same things and to become confident that we have done so in the right way.

As shown in figure 1, although humans and robots are copresent, their perceptual shells no longer overlap because of their mismatched capabilities. This is quite different from human-human communication. Thus one critical question is how humans with mismatched perceptual capabilities communicate with each other to establish a joint perceptual basis. While previous works have studied the role of mismatched spatial reasoning capabilities and diverse culture background in human-human communication, mismatched perceptual capabilities have not been addressed. As it is difficult to recruit human subjects with measurable differences in perceptual capabilities, a simulation system was developed to conduct experiments as shown in figure 2 (Liu, Fang, and Chai 2012).

At one end of the system is a director who is assumed to have higher perceptual capabilities; and at the other end is a matcher who is assumed to have lower perceptual capabilities. To simulate differences in perception, the director is given an original image and the matcher is given an impoverished image, which is rendered by applying a simple computer vision algorithm on the original scene. Some objects available to the director have a unique name. The director's goal is to communicate the names of the objects to the matcher so that the matcher comes to an understanding of which object has what name. Using this system, we conducted several user studies and collected a set of conversation data that demonstrated how the director and the matcher strived to mitigate perceptual differences and reach a mutual understanding of object names.

Figure 2 also shows an example of collaborative dialogue between the director (D) and the matcher (M). This example demonstrates similarity in collaborative behaviors as identified in collaborative referential communication. Clark and Wilkes-Gibbs (1986) describe referential communication as a collaborative process following the principle of least collaborative effort: "speakers and addressees try to minimize collaborative effort, the work both speakers and addressees do from the initiation of the referential process to its completion." As shown in figure 2, the partners make extra efforts to refer and ground their references. For example, instead of directly going to the "yellow pepper" and convey its name "Brittany," the director takes an extra effort by starting with "a cluster of four objects in the upper left." The director then goes through step-by-step installments (Clark and Wilkes-Gibbs 1986) and waits for the matcher's acceptance before taking another step that leads closer to the targeted object (that is, the yellow pepper). The matcher also makes an extra effort. Not only does the matcher provide feedback and acknowledgement at each step, but the matcher also provides additional descriptions about what the matcher perceives from the environment. Spatial expressions are commonly used to describe objects in the environment, which include not only the binary relations (for example, "the one to the left of the blue cup"), but also the group-based descriptions (for example, "a cluster of four objects in the upper right"). Thus computational models for both language interpretation (for example, reference resolution) and language generation (for example, referring expression generation) will need to model different spatial relations and take collaboration into account.

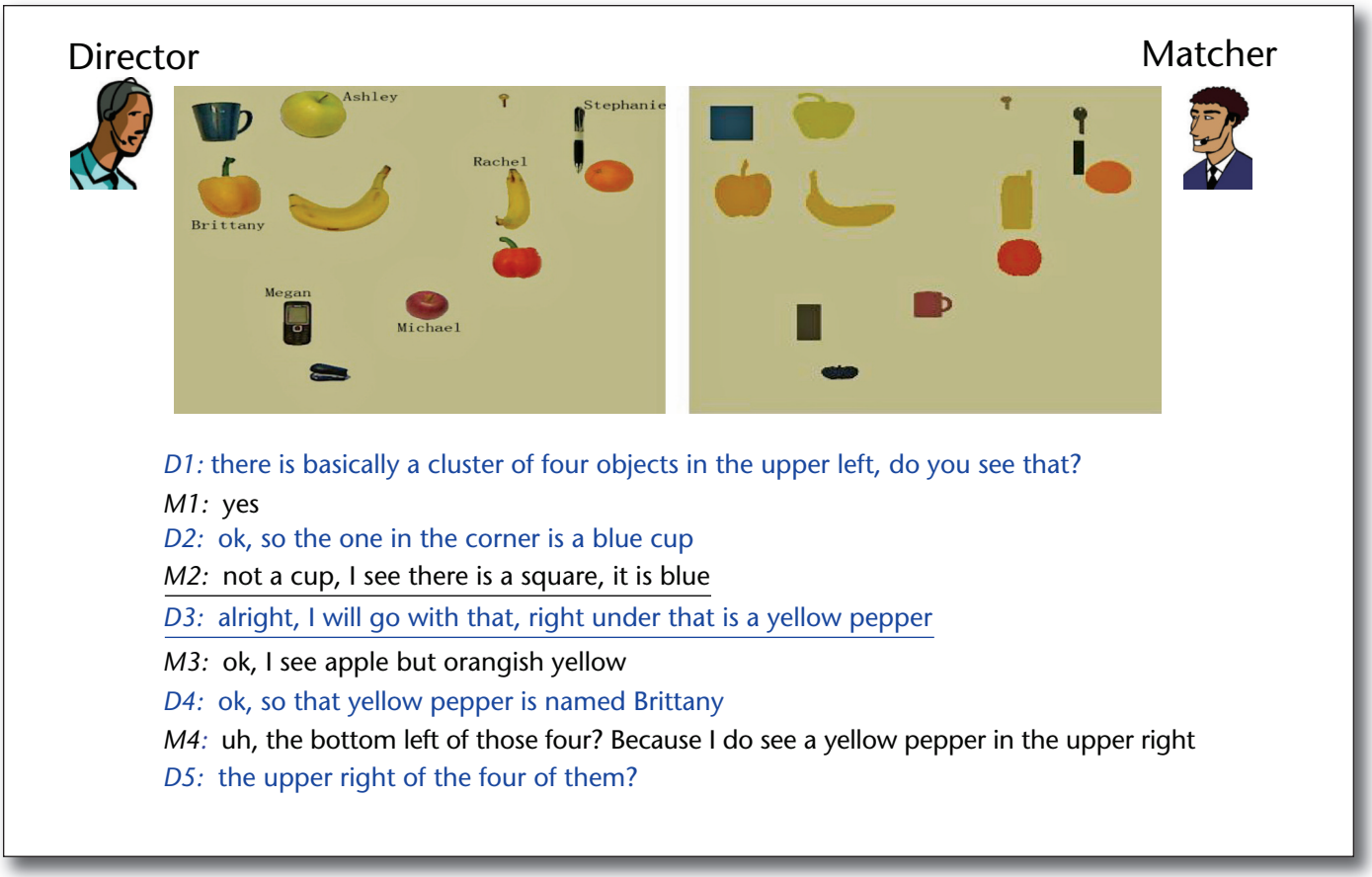



*Figure 2. A Simulated Environment to Study Human-Human Communication Under Mismatched Perceptual Basis.*

## Grounding Language to Perception

Recent years have seen an increasing amount of work on grounding language to visual perception (Tellex et al. 2011; Matuszek et al. 2012; Yu and Siskind 2013; Naim et al. 2015; Yang et al. 2016). At the LAIR lab, we focus on collaboration in grounding language to perception. Starting with simplified scenes, we have developed collaborative models for language grounding that are essential for both language interpretation and language generation.

### Grounded Language Interpretation

Grounded language interpretation refers to the problem of grounding human language to the robot's internal representation of the perceived world. This often involves grounding referring expressions (for example, "a cluster of four objects," "a blue cup," and others from the director's utterances in figure 2) to the external world perceived by the robot. Understanding which object(s) are talked about is challenging given mismatched underlying representations. However, shown in figure 2, as the collaborative discourse unfolds, referring expressions can be refashioned and more evidence (for example, through different spatial or group relations) can be provided by the director. Thus it is important to keep track of different relations and use these relations to help identify target objects. An effective way to capture relations is through graphs. This section introduces graph-based approaches to interpreting and grounding language to agents' perceptions.

Attributed relational graphs are particularly suitable for representing the perceived environment and conversation discourse (Chai, Hong, and Zhou 2004; Liu, Fang, and Chai 2012; Fang, Liu, and Chai 2012). Figure 3 shows an example of the graph representation. As the conversation unfolds, a language graph is constructed where each node captures a referring expression (for example, "yellow pepper") and the desired attributes for the grounded referents (for example, type: pepper; color: yellow). Various relations such as spatial relations between different referring expressions are captured by the edges in the graph. Similarly, the perceived objects and their spatial relations can be represented by a vision graph. Each node in the vision graph captures the lower-level numerical features associated with different dimensions of perception. The edges in the vision graph capture the spatial relations between every pair of

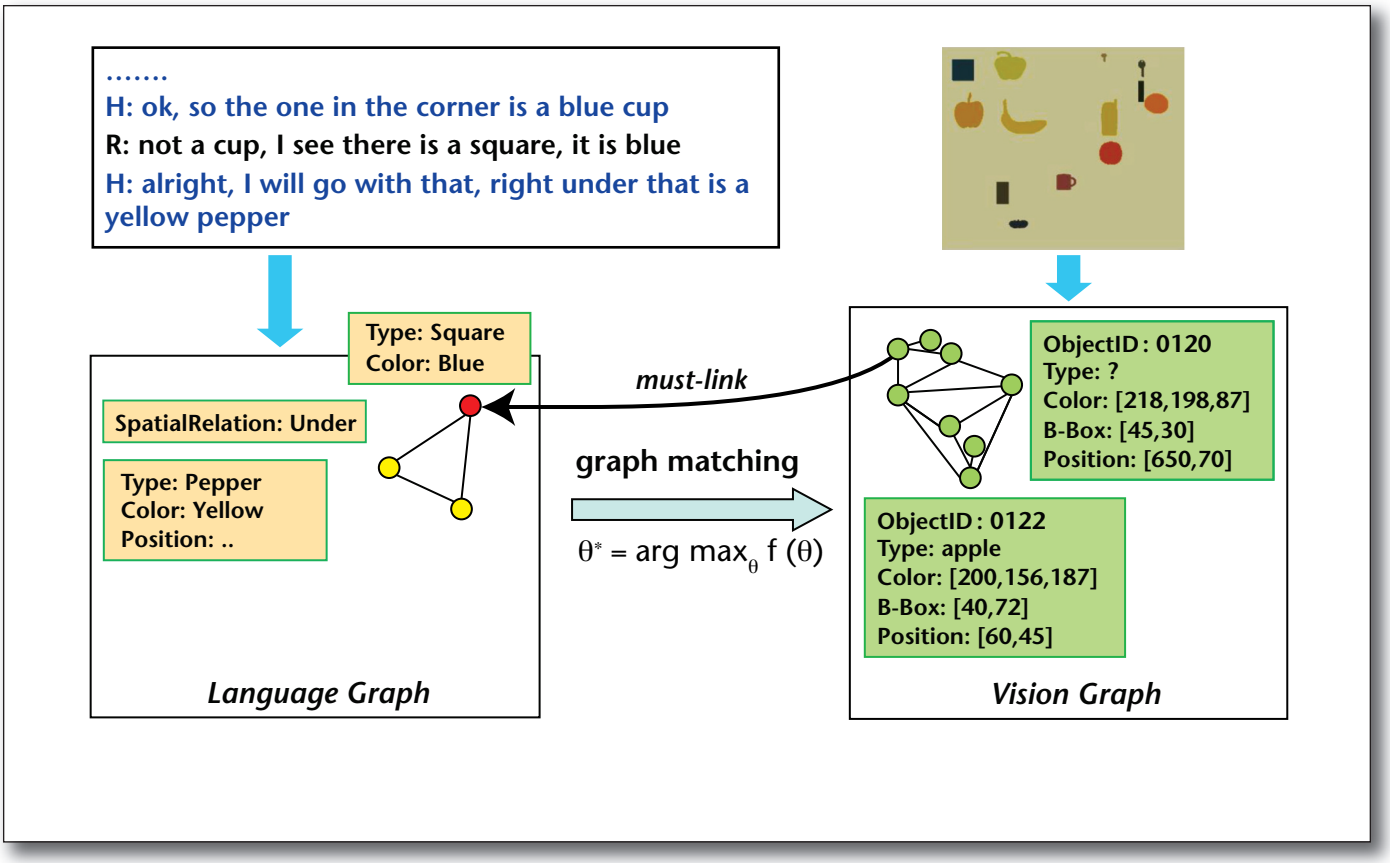



*Figure 3. Graph-Matching for Interpreting Referring Expressions.*

perceived objects. Thus, given a language graph and a vision graph, the problem of interpreting referring expressions becomes a graph matching problem that finds a best match (θ) between nodes and edges in the language graph to the nodes and edges in the vision graph that achieves the maximum compatibility between the two graphs. The graph compatibility can be measured based on node compatibility and edge compatibility. Node/edge compatibilities can be further decomposed into the match between a set of language descriptors (for example, the color "yellow" or the size "big") and the lower-level numerical features (for example, the color histogram or the size of the bounding box). These matches between language descriptors and corresponding features are also referred to as semantic grounding functions. Different weights can be associated with semantic grounding functions to indicate the strength of a particular dimension of perception applied to the overall matching. Given this graph-matching formulation, different approaches can be applied to obtain a solution, for example through greedy beam search algorithms as described in Liu, Fang, and Chai (2012) or probabilistic matching using the graduated assignment algorithm (Fang, Liu, and Chai 2012). The graph-based approaches have several advantages as will be illustrated.

### Modeling Rich Relations to Compensate Visual Recognition Errors

Graph-based approaches can effectively capture rich relations among objects and compensate for perceptual errors on individual objects. The matching algorithm does not enforce all language descriptors (for example, captured in the language nodes) completely match the lower-level perceived features (for example, captured in the vision nodes). Instead, semantic grounding functions return a real number indicating compatibility between a symbolic term and the perceived visual features. The algorithm relies on all the nodes/edges in the graphs to find the best approximation. By relying on these relations, the graph-based approach can compensate for visual recognition errors and mitigate perceptual differences between humans and agents.

A graph-based approach using greedy beam search was applied to the director-matcher communication data collected from our studies (described earlier). It has demonstrated some promising results. For example, when most of the objects (85 percent) could not be correctly recognized (by a simple computer vision algorithm), the graph-based approach successfully grounded to 66 percent of these misrecognized objects, leading to an over 27 percent performance gain compared to the approach without modeling

relations (Prasov and Chai 2010). Detailed results and analysis can be found in the paper by Liu, Fang, and Chai (2012). Regular graphs can only model binary relations between two entities. But they can be extended to hypergraphs to capture higher order (*n*-ary) relations, specifically for group-based expressions such as "a cluster of four objects at upper right." Details on using the hypergraphs are described by Liu et al. (2013).

### Incorporating Collaborative Patterns to Reduce Search Space

Graph-based approaches can effectively take into consideration collaborative patterns to reduce the search space for finding the best match. As shown in figure 2, the director and the matcher collaborate with each other to strive for a common ground of the shared environment. There are different types of discourse dynamics conversation partners engage in. For example, in M2 in figure 2, the matcher says "not a cup, I see there is a square, it is blue." In this case, the matcher first rejects the director's presentation and then presents what the matcher perceives from the environment. Given this presentation from the matcher, the director immediately accepts that in D3 ("all right, I will go with that") and then extends from there to the target object ("right under that is a yellow pepper"). This matcher-present-director-accept strategy (also called robot-present-human-accept strategy) has occurred frequently in our collected data. It can be directly incorporated in the graph-based approach to improve grounding of referring expressions (Liu et al. 2013).

For example, as shown in figure 3, when the robot introduces "a blue square" into the discourse, a node representing "a blue square" is added to the language graph. One nice thing is that when the robot introduces this node, it precisely knows which object in its own representation the expression "a blue square" is grounded to. Thus as shown in figure 3, there is a "must-link" between the language node and the vision node. These "must-links" provide additional constraints that lead to a smaller and more promising subspace for searching for solutions. Our empirical results have shown that incorporating the collaborative pattern of robot-present-human-accept strategy with the hypergraphs significantly improves performance of grounding language to the target objects by an absolute gain of more than 18 percent. More details can be found in Liu et al. (2013).

### Efficient Algorithms to Produce Multiple Matching Hypotheses

Efficient graph-matching algorithms are available to provide fast solutions with multiple hypotheses for follow-up dialogue. For example, in the paper by Liu, et al. (2014), a matching algorithm based on probabilistic labeling is presented. This approach integrates different types of evidence from the collaborative discourse into a Bayesian reasoning framework to ground referring expressions. In this algorithm, nodes in the vision graph are considered as labels. Labeling refers to assigning a label (from the vision graph) to a node in the language graph. The algorithm finds a ranked list of labels to each node in the language graph in an iterative manner. It first initiates the labeling probabilities by considering only the unary attributes, which can model any prior knowledge about how a node might be related to a label. It then updates the labeling probability of each node based on the labeling of its neighbors and the relations with them. Coreference relations between linguistic entities in the language graph are also modeled during updating. A nice feature of this approach is that when the labeling probabilities converge, the system can obtain multiple hypotheses (that is, a ranked list of labels for each node in the language graph), which can be naturally incorporated into dialogue (for example, for clarification). The empirical results have shown that the probabilistic labeling approach significantly outperforms the state-space search approach in both grounding accuracy and efficiency (Liu et al. 2014).

### Incremental Learning Algorithms to Support Adaptation to Perceptual Capability

Graph-based approaches allow robots to incrementally learn and assess their own perceptual capability and to use more reliable perceptual channels for better grounding. As described earlier, a key component in graph-matching is semantic grounding functions, which essentially match a linguistic descriptor (for example, the word "red") to the underlying visual features (for example, color histogram distribution). When grounding an expression (for example, "the red cup on the left") to the environment, the robot often needs to combine the semantic grounding functions for each attribute together through a weighted sum. The weight assigned to each semantic grounding function reflects how reliable the robot considers the associated semantic grounding function. Some semantic grounding functions may be more reliable than others and thus should be assigned higher weights. This is particularly important as humans and robots may have different perspectives of the visual features. For example, the robot's perception of "red" may be very different from the human's perception of red. Therefore, an important question is whether the robot can learn how reliable these semantic grounding functions are and, more importantly, whether the robot can update semantic grounding functions when it realizes they are not reliable.

To address these questions, we developed an optimization approach based on linear programming to automatically learn the weights during human-robot dialogue (Liu and Chai 2015). The idea is that, by interacting with its human partner (for example, through dialogue and feedback from the human), the robot is able to assess what dimension(s) of its own perception (for example, object recognition or color

distribution) are more aligned with the human's perception and thus the corresponding semantic grounding functions are more reliable. For example, when the robot cannot recognize objects in the environment very well, after a few rounds of interactions with humans, the robot will realize that the semantic grounding function associated with object type is not reliable and thus drop the associated weight. It will then rely on other perceptual features (and their associated semantic grounding functions) to match the linguistic expressions to its perception. The correct grounding to perception will become another training example to update the semantic grounding function itself. The detailed method and evaluations are provided by Liu and Chai (2015).

## Grounded Language Generation

The last section addresses the interpretation problem, namely, given language descriptions from human partners, how to ground them to perceived objects even though the robot only has imperfect perception of the shared environment. To enable human-robot communication, an equally important problem is how the robot can effectively generate a language description to enable its human partner understand which object is being referred to.

There has been a tremendous amount of work on referring expression generation (REG) in the last two decades (Dale 1995; Krahmer and Deemter 2012). The typical objective is to generate a single minimum language description that allows the listener to distinguish the target object from the distractors. Except for a few (Mitchell, Van Deemter, and Reiter 2013), most previous works have applied the assumption that agents and humans have access to the same domain information and the agent has a complete representation of the shared world. However, this assumption no longer holds in situated human-robot dialogue as humans and robots have mismatched representations of the shared environment. In addition, the perfect knowledge of the environment is not available to the robot ahead of time. The agent needs to make inference of the shared environment and connect lower-level visual features with words. This process is full of uncertainties and also error prone. Perceptual differences in situated human-robot dialogue pose new challenges to REG.

### Revisiting REG

To understand this new challenge, we have revisited the problem of REG in the context of mismatched perceptual basis (Fang et al. 2013). We extended a well-known graph-based approach (Krahmer, Van Erk, and Verleg 2003) that has shown effective in previous works and by incorporating uncertainties in perception into cost functions. We further extended regular graph representation into hypergraph representation to account for group-based spatial relations that are important for visual descriptions. Our empirical results have demonstrated that if the agent has a perfect perception and has a complete representation of the shared environment (the setting most previous works were based on), our hypergraph based approach achieved higher performance (84.2 percent accuracy) compared to the original approach based on regular graphs (80.4 percent). However, when the agent does not have a perfect perception of the shared environment (which is often the case in human-robot dialogue), the performance of our hypergraph has dropped significantly to 45 percent. This performance gap indicates that the current approach to generate minimum descriptions may not be applicable for situated human-robot dialogue. It calls for new solutions for REG that are capable of mediating mismatched perceptual basis. These results have motivated our work on collaborative models for REG.

### Collaborative Models for REG

Previous work on the collaborative process for referential communication (Clark and Wilkes-Gibbs 1986) states that, to minimize collaborative effort, partners tend to go beyond issuing an elementary referring expression, but rather use other different types of expressions such as episodic, installment, self-expansion, and others. Motivated by these findings as well as observations from our own study on human-human communication described earlier, we have developed collaborative models for REG particularly to capture the following two types of collaborative behaviors.

> *Episodic Model* generates referring expressions in an "episodic" fashion by generating a sequence of smaller noun phrases, which lead to the target object, for example, as in "below the orange, next to the apple, it's the red bulb."

> *Installment Model* generates referring expressions in an "installment" fashion by generating one small noun phrase, waiting for the listener's response, and then generating another small noun phase based on the feedback. This process iterates until the target object is attained. For example:

> *Speaker:* under the pepper
> *Listener:* yes.
> *Speaker:* there is a group of three objects.
> *Listener:* OK.
> *Speaker*: there is a yellow object on the right within the group.

In order for agents to generate these episodes in an installment manner, we treated REG as a sequential decision-making problem and formulated it under the reinforcement learning framework (Fang, Doering, and Chai 2014). The idea is that, from its prior experience engaging in referential communication with a human, the agent should be able to learn a good generation policy for any given state of communication. Using terms from reinforcement learning, a policy $\pi : S \rightarrow A$ is a mapping from states ($S$) to actions ($A$). The action-value function $Q(s, a)$ is the expected return for starting in state $s \in S$ and taking

action $a \in A$. The goal is to learn an optimal policy $\pi^* = \arg\max_a Q(s, a)$ that maximizes $Q(s, a)$. In the context of REG, a state can be characterized by many factors such as the uncertainties of the perceived environment, the target object, the current landmark object, which has been confirmed by the human, and the human feedback. An action describes a generation strategy such as what object to describe for an episode and what descriptor to use (for example, whether using attributes or spatial relations to landmark objects to describe it). As the space of $S$ and $A$ can be big, it's not possible to enumerate all possible pairs. Thus function approximation is used to approximate the value function through linear regression, that is, a linear combination of weighted features. Using a simulated environment, we conducted experiments where the agent interacted with users from Amazon Mechanic Turk. The SARSA algorithm (Sutton and Barto 1998) was applied to learn the weights associated with the features for $Q(s, a)$. The learned value function can be used to identify the best generation action for any state during communication. Our experimental results have shown that, compared to the leading noncollaborative approach (based on hypergraphs), the collaborative models significantly improve the performance. In particular, the installment model (68.9 percent accuracy) has led to an absolute gain of 21 percent compared to the noncollaborative approach (47.2 percent accuracy). These results have shown that collaborative models are more effective in mitigating perceptual differences between humans and robots in referential communication. More details about the approach and empirical results are described by Fang, Doering, and Chai (2014).

## Experiments on Human-Robot Communication

Previous sections give a brief overview to collaborative models for grounded language interpretation and language generation with a goal to mediate perceptual discrepancies. This section describes two empirical studies that extend these models in human-robot communication with a particular focus on the robot's collaborative behaviors.

### Collaborative Effort in Establishing Common Ground of Shared Environment

The graph-based approach for grounding language to environment is integrated in a human-robot communication system. The first experiment investigates the role of collaborative effort from the robot in facilitating language interpretation to establish a common ground of shared environment (Chai et al. 2014). A demo of the system used in the experiment can be seen on YouTube.[1]

A 2x2 factorial design was applied in the experiment. There were two factors: perceptual difference and collaborative effort. Each factor has two levels. A high perceptual difference means the human and the robot have a high mismatch in their perceptions. In this case, we manipulated the robot vision system so that 60 percent or 90 percent of the objects in the shared environment couldn't be correctly recognized. A low perceptual difference refers to the situation where only 10 percent or 30 percent of objects couldn't be recognized correctly by the robot. A low collaborative effort refers to the robot's minimum effort in accepting or rejecting a presentation from the human through explicit confirmation. A high collaborative effort models the matcher's behavior in human-human communication. In this setting, the robot makes an extra effort in proactively describing what it perceives from the shared environment in addition to an explicit confirmation. Figure 4 shows some examples of these two different settings.

| Low Collaborative Effort | High Collabrative Effort |
|---|---|
| **Accept**<br>*"Got it"* | **Accept and Describe**<br>*"Sure, the orange bottle is Eric"* |
| **Weak Accept**<br>*"Il think I see it, (pointing to the object)"* | **Weak Accept and Describe**<br>*"I see something there (pointing), it is an orange bottle"* |
| **Reject**<br>*"I don't see it"* | **Reject and Describe**<br>*"I don't see it, but I see an orange bottle there (pointing)"* |

*Figure 4. Example Responses Under Two Different Levels for the Collaborative Effort.*

In the experiments, human subjects were instructed to teach the robot names of different objects in the shared environment through natural language dialogue. At the end of the dialogue, two metrics were measured: (1) perceived grounding, which counts percentage of dialogues where the human partners believed that the robot had successfully acquired all the names; and (2) true grounding, which counts the number of correct names that are actually acquired by the robot through the dialogue.

Our empirical results have shown that a low collaborative effort leads to a higher perceived grounding performance, which means it can fool its human partners into believing a common ground is established. However, such beliefs do not reflect true common ground and are even more detrimental than failure in reaching a common ground. Our results further show that, under a low perceptual difference, different levels of collaborative effort do not make a difference in true grounding; however under a high perceptual difference, a high collaborative effort leads to significantly higher performance in true grounding. These results suggest that, to mediate perceptual differences and establish a common ground of the shared environment, the robot should make an extra


*Figure 5. Multimodal Collaboration in Referential Communication.*

effort to communicate with its partner and make him or her aware of its internal representation of the shared world. Detailed results of this experiment are described by Chai et al. (2014).

### Embodied Collaboration in Referential Communication

As described earlier, collaborative models for REG have shown promising results in language-based communication in a simulated environment of mismatched perceptual basis. In human-robot communication, a unique characteristic of physical world interaction is embodiment: robots and humans both have physical bodies and they can use nonverbal modalities (for example, gesture and eye gaze) to refer to the shared world and to provide immediate feedback (Chen et al. 2015, Kennington et al. 2015). Thus one of our efforts was to integrate the installment model with embodiment to facilitate human-robot referential communication. We were particularly interested in two variables representing embodiment — robot pointing gesture and human eye gaze feedback.

Deictic gestures play an important role in human-robot referential communication (Sauppe and Mutlu 2014). In our work, we directly model the pointing gesture as an additional action in the value function (described earlier). The cost of pointing gestures is incorporated as a feature in the linear regression model, which is calculated based on the distance from the robot to the target object, the size of the target object, adjacency of other objects to the target object, and so on.

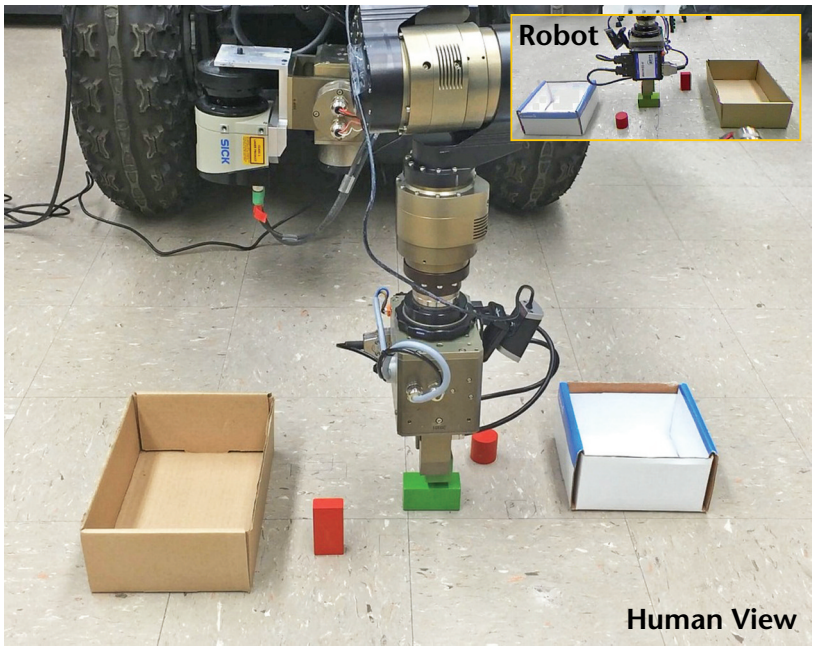
Psycholinguistic studies have shown that human eye gaze is directly linked with language comprehension (Tanenhaus et al. 1995). Immediately after hearing a referring expression, the listener's eyes move to the objects being referred to. Motivated by these findings we incorporated human's real-time gaze as intermediate feedback in the installment model. We used an ASL eye tracker to measure whether the fixated object coincided with the agent's intended object. If yes, the human feedback was considered positive; otherwise, it was considered negative.

Based on these two factors, we conducted an experiment to examine embodied collaboration in referential communication (Fang, Doering, Chai 2015). In figure 5, the left picture shows the experimental setting where the human wears an ASL mobile eye tracker and the right picture shows an example of a gaze fixation captured by the ASL tracker during communication. As we were particularly interested in the situation where humans and robots have mismatched perceptual basis, we intentionally applied a very simple CV algorithm to generate a high misrecognition rate (73 percent).

Our empirical results have shown that models that incorporate the robot pointing gesture consistently outperformed the models that don't incorporate the pointing gesture. However, the incorporation of human eye gaze as intermediate feedback did not perform well. Human verbal feedback performed better than the gaze feedback by itself or together. Our results indicate that human gaze can be very noisy, especially in situated interaction. It is very difficult to reliably capture gaze fixations and thus reflect the corresponding feedback. More in-depth investigations on human gaze feedback are needed in the future.

## Grounding Language to Action

For robots to follow human language instructions and perform actions in the physical world, grounding language to perception alone is not sufficient. This section gives a brief introduction to grounding language to robot actions (She et al. 2014a, b; Misra et al. 2014). Most robotic systems are programmed with primitive operations such as move to, open




Human: Stack the green block on the left to the green block on the right
Robot: I don't know how to perform this stack, please give step-by- step instructions.
Human: Move the green block on the left to the top of the green block on the right.
Robot: Alright then (performing the move action).
Human: Open gripper.
Robot: OK.
Human: Now you achieved the stack action.
Robot: Understood.
Human: Sort the blocks by color
Robot: What do you mean by this sort?
Human: Put the red blocks in the brown box.
Robot: ….

*Figure 6. An Example of a Robotic Arm Learning New Actions through Step-by-Step Natural Language Instructions.*

gripper, and close gripper. For a robotic arm to achieve a high-level action (for example, stack an object on top of another), a discrete controller is often first applied to find a sequence of primitive operations. These primitive operations are then passed to a continuous planner and translated into trajectories of arm motors. Thus a critical question is how to connect language commands (for example, "stack the blue cylinder on the red block") with the corresponding sequence of primitive robotic operations. To address this question, our effort has focused on the representation of grounded verb frames and the acquisition and learning of grounded representations through collaborative step-by-step instructions.

### Learning Grounded Verb Semantics Through Collaborative Step-by-Step Instructions

Suppose a human operator asks a robotic arm (shown in figure 6) to "stack the green block on the left to the green block on the right," and suppose the robot's language interpretation can perfectly identify the verb frame *stack* with two arguments (origin and destination) and successfully ground the arguments (that is, "the green block on the left" and "the green block on the right") to the respective objects in world. The arm may still not be able to perform the action if it does not have knowledge of how to connect the "stack" action with the low-level primitive operations. However, the human operator can teach the robot high-level actions (for example, stack) in a

step-by-step manner as shown in figure 6.[2] In this case, a stack action involves two primitive operations: move to and open gripper.

Given this teaching and learning instance, how should the robot internally represent knowledge or grounded semantics for the verb frame *stack?* If it's only associated with "move to" followed by "open gripper," the acquired knowledge will not be very useful in a new situation where it may involve several "move to" and "open gripper" operations to accomplish the stack action. Thus, a more desirable representation for grounded semantics of the verb frame *stack*(*A, B*) should capture the desired goal state of the physical world caused by this action. The goal state, which is represented by a conjunction of logical predicates, for example, "on(A, B) ∧ G_open," can be acquired by the robot after performing the low level operations. Representing verb semantic frames with final state of the physical world allows the planner to automatically identify a sequence of low-level primitive operations for any situation. For example, in our experiment with a SCHUNK robotic arm (She et al. 2014a), a teacher used 4 steps to teach the stack action. When applying the acquired model for "stack" in 20 novel situations, the robot was able to complete the action with an average of approximately 7 steps. In one particular situation, the robot was able to take 12 steps to successfully complete the stack action.

In our experiments, we have also examined the role of collaboration in teaching new actions (She et al. 2014b). We controlled two settings. In the collaborative step-by-step instruction setting, the human teacher provides one step at a time and watches the robot's corresponding action. If the action is successfully performed, the teacher will move to the next step; otherwise, the teacher will change its course of instruction to cope with the incorrect response. In the noncollaborative one-shot instruction setting, the teacher provides the instructions all at the beginning without watching and waiting for feedback from the robot. Our experimental results have shown that, although one-shot instructions take less time to teach and learn, collaborative step-by-step instructions allow the robot to acquire better representations of verb frames and thus lead to more action completion in novel situations.

### Learning Grounded Verb Hypothesis Space

While representing grounded verb semantics with the intended goal state has shown promising in a simplified block world (She et al. 2014b), the acquired representation can be overfitting to the particular learning instances. To address this problem, our recent work extends a single goal state representation to a hypothesis space of goal states for verb representations (She and Chai 2016). For example, suppose a human teaches the robot how to "fill the cup with water." After experiencing the change of state of the physical world by performing the action taught by the human, the robot is able to ground verb frame *fill*(*x, y*) to the desired goal state "Has(x, y) ∧ Grasping(x) ∧In(x, o1) ∧¬ In(x, o2)." Based on this goal state, a hypothesis space using a specific-to-general hierarchy can be built as shown in figure 7. In this hypothesis space, any hypothesis of goal state other than the shaded ones allows the planner to come up with exactly the same sequence of primitive operations as the original goal state (at the bottom of the hierarchy). The hypotheses higher on the hierarchy have fewer number of predicates and thus have higher chances to be satisfied in novel situations. Therefore during learning, the robot automatically acquires and updates a hypothesis space for each verb frame. Given a new situation, when a verb command is issued by the human, the robot will identify the most relevant hypothesis from the hypothesis space to calculate a sequence of primitive operations. Using data made available by Misra et al. (2015), our empirical results have shown that the hypothesis space representation significantly outperforms the representation with single hypothesis of goal state. Details on this approach and empirical evaluations are described by She and Chai (2016).

## Conclusions

Enabling situated human-robot communication faces many challenges and opportunities (Bohus and Horvitz 2010). One of the significant challenges is the capability of grounding human language to a robot's internal representations of perception and action. This involves multiple aspects of complexities. Even the robot has existing knowledge about how a word (for example, an adjective or a noun) is connected with the underlying visual features; during real-time communication, the robot may still not be able to ground human language to its own representation of the perceived world due to subtle change of the environment. Computer vision algorithms have improved tremendously in recent years, especially given the advances in deep learning. However, in a new environment, when there is not sufficient training data, machine perceptual systems are still fragile. The perceptual differences between humans and robots in situated communication remain a practical problem. As shown in this article, an effective solution to this problem is to incorporate collaborative behaviors into grounded language processing and enable collaboration from the robot to ground communication.

It is often the case that during communication a robot will encounter new words, new objects, and new actions it does not have existing knowledge about. As shown in this article and other recent work (Cantrell et al. 2012; Mohan, Kirk, and Laird 2013; Mohseni-Kabir et al. 2015; Thomason et al. 2016), language and collaborative dialogue play an impor-

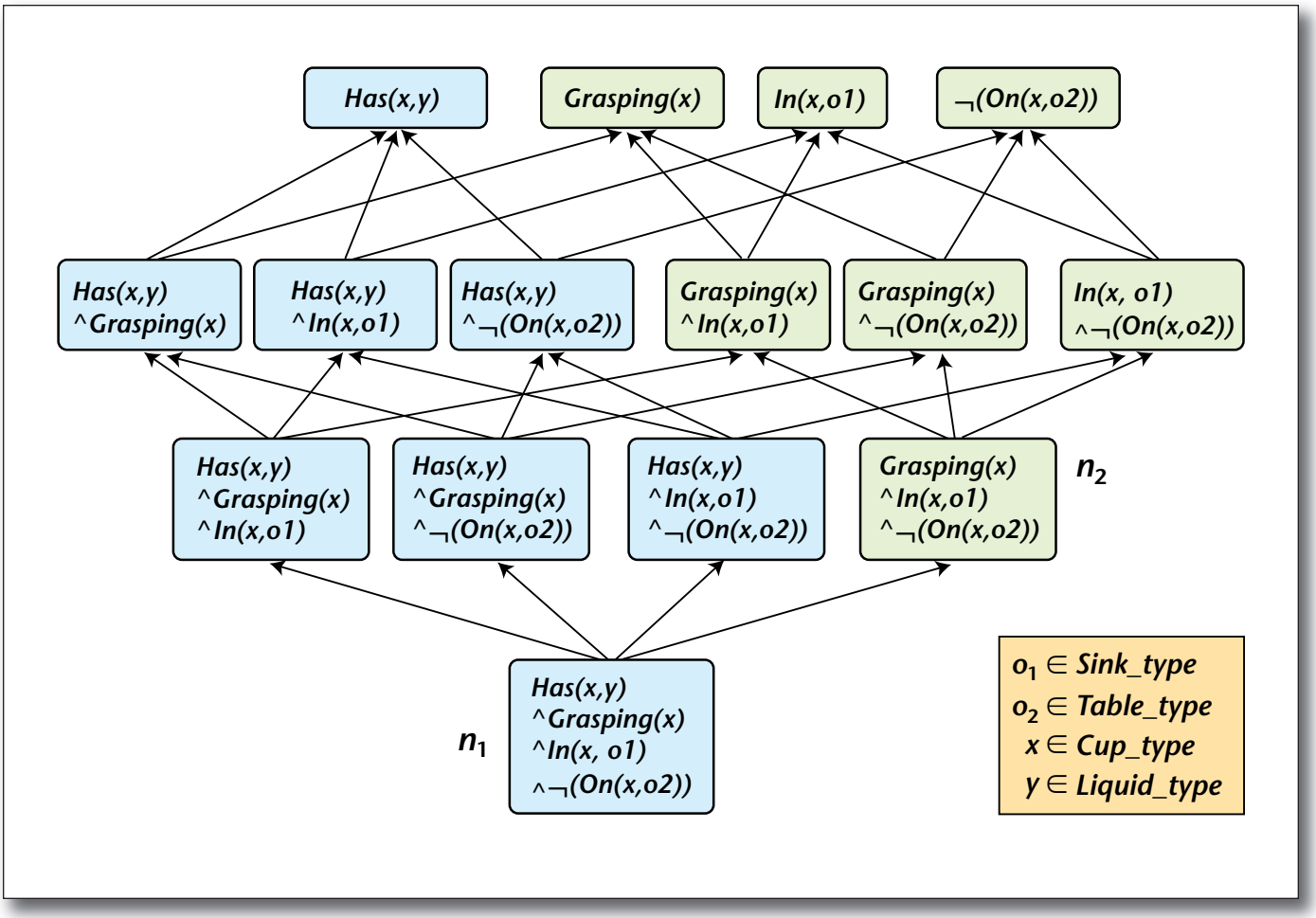



*Figure 7. An Example of Grounded Hypothesis Space for Verb Frame fill(x, y).*

tant role in enabling the robot to continuously learn grounded meanings, the environment, and tasks from its human partners. To further support interactive robot learning through natural language dialogue, our current work is to develop approaches to ground language to participants of actions in more complex visual scenes (for example, a kitchen environment) (Yang et al. 2016, Gao et al. 2016). In addition, we are exploring acquisition of rich task structures from human language instructions and visual demonstrations (Liu et al. 2016a, 2016b). The ultimate goal is to enable robots to continuously learn from human partners through their life-long interactions.

## Acknowledgement

We would like to thank Malcolm Doering, Kenneth Hanson, Cody Littley, Spencer Ottarson, and Shaohua Yang for their contributions to this research. The work on the SCHUNK robotic arm was conducted in collaboration with Yunyi Jia, Yu Cheng, and Ning Xi from the Robotics and Automation Lab at Michigan State University. The research at the LAIR lab was supported by CNS-0957039, IIS-1050004, IIS-1208390, IIS-1617682 from the National Science Foundation, N00014-11-1-0410 from the Office of Naval Research, and a DARPA SIMPLEX grant N66001-15-C-4035 through the University of California, Los Angeles.

## Notes

1. A system demo can be found at www.youtube.com/watch?v=vPA2AUJq6cI

2. A system demo can be found at www.youtube.com/watch?v=MGA6aqKGM0w

## References

Austin, J. L. 1962. *How to Do Things with Words*. Oxford University Press.

Bohus, D., and Horvitz, E. 2010. On the Challenges and Opportunities of Physically Situated Dialog. In *Dialog With Robots: Papers from the AAAI Fall Symposium*. Technical Report FS-10-05. Palo Alto, CA: AAAI Press.

Cantrell, R.; Talamadupula, K.; Schermerhorn, P.; Benton, J.; Kambhampati, S.; and Scheutz, M. 2012. Tell Me When and

Why to Do It! Run-Time Planner Model Updates Via Natural Language Instruction. In *Proceedings of the International Conference on Human-Robot Interaction* (HRI'12), 71–478. New York: Association for Computing Machinery.

Chai, J. Y.; Hong, P.; and Zhou, M. 2004. A Probabilistic Approach for Reference Resolution in Multimodal User Interfaces. In *Proceedings of the 9th ACM International Conference on Intelligent User Interfaces* (IUI), 70–77. New York: Association for Computing Machinery. dx.doi.org/10.1145/964442.964457

Chai, J. Y.; She, L.; Fang, R.; Ottarson, S.; Littley, C.; Liu, C.; and Hanson, K. 2014. Collaborative Effort Towards Common Ground in Situated Human-Robot Dialogue. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (HRI'14), 33-40. New York: Association for Computing Machinery. dx.doi.org/10.1145/2559636.2559677

Chen, L.; Javaid, M.; Di Eugenio, B.; Efran, M. 2015. The Roles and Recognition of Haptic-Ostensive Actions in Collaborative Multimodal Human-Human Dialogues. *Computer Speech and Language* 32(1): 201–231. dx.doi.org/10.1016/j.csl.2015.03.010

Clark, H. H. 1996. *Using Language*. Cambridge University Press. dx.doi.org/10.1017/CBO9780511620539

Clark, H. H., and Wilkes-Gibbs, D. 1986. Referring as a Collaborative Process. *Cognition* 22(1): 1–39. dx.doi.org/10.1016/0010-0277(86)90010-7

Dale, R. 1995. Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science* 19(2) 233–263. dx.doi.org/10.1207/s15516709cog1902_3

Fang, R.; Doering, M.; and Chai, J. Y. 2014. Collaborative Models for Referring Expression Generation Towards Situated Dialogue. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence,* 1544–1550. Palo Alto, CA: AAAI Press.

Fang, R.; Doering, M.; and Chai, J. Y. 2015. Embodied Collaborative Referring Expression Generation in Situated Human-Robot Dialogue. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction,* HRI 2015, 271–278. New York: Association for Computing Machinery.

Fang, R.; Liu, C.; and Chai, J. Y. 2012. Integrating Word Acquisition and Referential Grounding Towards Physical World Interaction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (ICMI), 109–116. New York: Association for Computing Machinery. dx.doi.org/10.1145/2388676.2388703

Fang, R.; Liu, C.; She, L.; and Chai, J. Y. 2013. Towards Situated Dialogue: Revisiting Referring Expression Generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (EMNLP), 392–402. Stroudsburg, PA: Association for Computational Linguistics.

Gao, Q.; Doering, M.; Yang, S.; and Chai, J. Y. 2016. Physical Causality of Action Verbs In Grounded Language Understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL) Volume 1: Long Papers. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.18653/v1/p16-1171

Grice, H. P. 1975. Logic and Conversation. In *Syntax and Semantics 3: Speech Acts*, ed. P. Cole and J. L. Morgan, 41–58. New York: Seminar Press.

Kennington, C.; Iida, R.; Tokunaga, T.; and Schlangen, D. 2015. Incrementally Tracking Reference in Human/Human Dialogue Using Linguistic and Extra-Linguistic Information. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.3115/v1/n15-1031

Krahmer, E., and Deemter, K. V. 2012. Computational Generation of Referring Expressions: A Survey. *Computational Linguistics* 38(1): 173–218. dx.doi.org/10.1162/COLI_a_00088

Krahmer, E.; Van Erk, S.; and Verleg, A. 2003. Graph-Based Generation of Referring Expressions. *Computational Linguistics* 29(1): 53–72. dx.doi.org/10.1162/089120103321337430

Liu, C., and Chai, J. Y. 2015. Learning to Mediate Perceptual Differences in Situated Human-Robot Dialogue. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Liu, C.; Chai, J. Y.; Shukla, N.; and Zhu, S. 2016a. Task Learning Through Visual Demonstration and Situated Dialogue. *Symbiotic Cognitive Systems: The Workshops of the 30th AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Liu, C.; Fang, R.; and Chai, J. Y. 2012. Towards Mediating Shared Perceptual Basis in Situated Dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue,* 140–149. Stroudsburg, PA: Association for Computational Linguistics.

Liu, C.; Fang, R.; She, L.; and Chai, J. Y. 2013. Modeling Collaborative Referring for Situated Referential Grounding. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Sigdial), 78–86. Stroudsburg, PA: Association for Computational Linguistics..

Liu, C.; She, L.; Fang, R.; and Chai, J. Y. 2014. Probabilistic Labeling for Efficient Referential Grounding Based on Collaborative Discourse. In *Proceedings of The 52nd Annual Meeting of Association of Computational Linguistics* (ACL) Volume 2: Short Papers, 13–18. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.3115/v1/p14-2003

Liu, C.; Yang, S.; Sadiya, S.; Shukla, N.; He, Y.; Zhu, S.; and Chai, J. Y. 2016b. Jointly Learning Grounded Task Structures from Language Instruction and Visual Demonstration. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (EMNLP). Stroudsburg, PA: Association for Computational Linguistics.

Matuszek, C.; Fitzgerald, N.; Bo, L.; Zettlemoyer, L.; Fox, D. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proceedings of the 29th International Conference on Machine Learning*. Madison, WI: Omnipress.

Misra, D.; Sung, J.; Lee, K.; and Saxena, A. 2014. Tell Me Dave: Context Sensitive Grounding of Natural Language to Mobile Manipulation Instructions. In *Robotics: Science and Systems X*. Berlin: Robotics Science and Systems Foundation.

Misra, D.; Tao, K.; Liang, P.; and Saxena, A. 2015. Environment-Driven Lexicon Induction for High-Level Instructions. In *Proceedings of The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Volume 1: Long Papers, 992–1002. Stroudsburg, PA: Association for Com-

putational Linguistics. dx.doi.org/10.3115/v1/p15-1096

Mitchell, M.; Van Deemter, K.; and Reiter, E. 2013. Generating Expressions That Refer to Visible Objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1174–1184. Stroudsburg, PA: Association for Computational Linguistics

Mohan, S.; Kirk, J.; and Laird, J. 2013. A Computational Model for Situated Task Learning with Interactive Instruction. Paper presented at the International Conference on Cognitive Modeling, 11–14 July, Ottawa, Canada.

Mohseni-Kabir, A.; Rich, C.; Chernova, S.; Sidner, C. L.; and Miller, D. 2015. Interactive Hierarchical Task Learning from a Single Demonstration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction,* 205–212. New York: Association for Computing Machinery.

Naim, I.; Song, Y. C.; Liu, Q.; Huang, L.; Kautz, H.; Luo, J.; and Gildea, D. 2015. Discriminative Unsupervised Alignment of Natural Language Instructions with Corresponding Video Segments. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.3115/v1/N15-1017

Prasov, Z., and Chai, J. Y. 2010. Fusing Eye Gaze with Speech Recognition Hypotheses to Resolve Exophoric References in Situated Dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* 471–481. Stroudsburg, PA: Association for Computational Linguistics.

Sauppe, A.; and Mutlu, B. 2014. Robot Deictics: How Gesture and Context Shape Referential Communication. 2014. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. New York: Association for Computing Machinery. dx.doi.org/10.1145/2559636.2559657

She, L., and Chai, J. Y. 2016. Incremental Acquisition of Verb Hypothesis Space Towards Physical World Interaction. In *Proceedings of the 54th Annual Meeting of The Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.18653/v1/p16-1011

She, L.; Cheng, Y.; Chai, J. Y.; Jia, Y.; Yang, S.; and Xi, N. 2014a. Teaching Robots New Actions Through Natural Language Instructions. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 868–873. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/ROMAN.2014.6926362

She, L.; Yang, S.; Cheng, Y.; Jia, Y.; Chai, J. Y.; and Xi, N. 2014b. Back to the Blocks World: Learning New Actions Through Situated Human-Robot Dialogue. In *Proceedings of the 15th Annual Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Sigdial) 89–97, Philadelphia, June 18-20.

Sutton, R. S.; and Barto, A. G. 1998. *Introduction to Reinforcement Learning*. Cambridge, MA: The MIT Press.

Tanenhaus; M. K.; Spivey-Knowlton, M. J.; Eberhard, K. M.; Sedivy, J. C. 1995. Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science* 268: 1632. dx.doi.org/10.1126/science.7777863

Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M. R.; Banerjee, A. G.; Teller, S. J.; and Roy, N. 2011. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Thomason, J.; Sinapov, J.; Svetlik, M.; Stone, P.; and Mooney, R. J. 2016. Learning Multimodal Grounded Linguistic Semantics by Playing "I Spy." In *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (IJCAI-16). Palo Alto, CA: AAAI Press.

Yang, S.; Gao, Q.; Liu, C.; Xiong, C.; Zhu, S.; and Chai, J. Y. 2016. Grounded Semantic Role Labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.18653/v1/n16-1019

Yu, H., and Siskind, J. M. 2013. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL 2013) Volume 1: Long Papers, 53–63. Stroudsburg, PA: Association for Computational Linguistics.

**Joyce Chai** is a professor in the Department of Computer Science and Engineering at Michigan State University. She received a Ph.D. in computer science from Duke University in 1998. Her research interests include natural language processing, situated dialogue agents, artificial intelligence, human-robot communication, and intelligent user interfaces. Her recent work has focused on grounded language processing to facilitate situated communication with artificial agents (for example, robots). The objective is to enable natural interaction between humans and agents and allow agents to continuously learn from humans about the shared environment and joint tasks.

**Rui Fang** is a research scientist at Thomson Reuters. He received a Ph.D in computer science from Michigan State University in 2014. His research interests include natural language processing, multimodal dialogue systems, and human-machine interaction.

**Changsong Liu** is a post-doc researcher in the LAIR lab at Michigan State University. He received a Ph.D in computer science from Michigan State University in 2015. His research interests include natural language processing, human-machine dialogue, and educational uses of technology. His current work focuses on task learning through language instruction and visual demonstration.

**Lanbo She** is a Ph.D. student in the LAIR lab at Michigan State University. He received a B.E. in electronic engineering from the University of Science and Technology of China. His current work focuses on incremental learning for task acquisition through human-robot dialogue.