Interpreting Multimodal Referring Expressions in Real Time

Miles Eldon, David Whitney, Stefanie Tellex
Brown University

***Abstract*— Humans communicate about objects using language and gesture, fusing information from multiple modalities over time, and robots need to interpret their communication in order to collaborate with them on shared tasks. Existing work has addressed this problem in single modalities, such as natural language or gesture, or fused modalities in non-realtime systems, but a gap remains in creating systems that simultaneously fuse information from language and gesture incrementally, over time. Processing communicative input incrementally has the potential to speed up the robot's reaction time, as well as incorporate the timing of words and gesture into the understanding process, increasing the robot's speed and accuracy at interpreting the person's mental state. To address this problem, we define a multimodal Bayes' filter for interpreting a person's referring expressions to objects. Our approach outputs a distribution over the referent object at 14Hz, updating dynamically as it receives new observations of the person's spoken words and gestures. We collected a new dataset of people referring to objects in a tabletop setting and demonstrate that our approach is able to infer the correct object with 90% accuracy. Additionally, we demonstrate that our approach enables a Baxter robot to provide back-channel responses in real-time, pointing toward a mathematical framework for human-robot communication as a *joint activity* [Clark, 1996].**

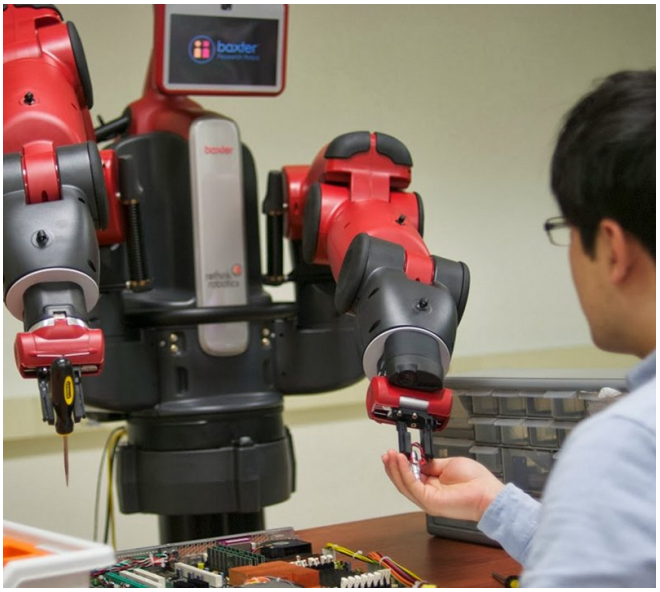


Fig. 1. Robots that collaborate with people need to understand their references to objects in the environment. For example, if a person asks for a tool using language and gesture, the robot needs to interpret the person's reference in order to pick up the correct tool.

I. INTRODUCTION

In order for humans and robots to collaborate in complex tasks, robots must be able to understand people's references to objects in the external world. For example, a robotic cooking assistant might fetch ingredients and tools, while a robotic factory assistant could deliver a part or a hospital robot could deliver water to a bedridden patient; Figure 1 shows a robot handing a tool to an engineer. To refer to objects, people use a combination of language, gesture, and body language such as eye gaze and pointing. People provide these signals continuously, and a person's reference can quickly change based on new information about the domain. Moreover, a human listener responds to these signals as they are given using *backchannels*, for example nodding their head when they understand and looking confused or interrupting to ask a question when they do not. Clark [1996] refers to this continuous dance as *joint activity* and compares language use to playing a duet because of its collaborative nature, where both parties act to establish common ground and reduce uncertainty.

Responding quickly and incorporating the relative timing of speech and gesture is critical for accurate understanding. Fast responses make interaction more fluid and enable a robot to provide backchannel feedback based on its ability to understand: when it is confident, it can indicate that it is confident, and when it is unsure, it can indicate that as well. This backchannel feedback could elicit appropriate responses from the person: they will move to the next task when the robot understands, or provide more information to disambiguate when the robot is confused.

Despite the importance of real-time response to multimodal input, existing unimodal models do not integrate information from language and gesture [Matuszek et al., 2014, Tellex et al., 2011, Kollar et al., 2010]. Approaches that fuse information from language and gesture [Matuszek et al., 2014] do not take into account that information appears to the system over a period of time. These approaches make it impossible for a robot to provide back-channel feedback, because of the length of time required to interpret the communication and because of the inability to interpret partial utterances.

To provide a foundation for these capabilities, we propose a Bayes' filtering approach for interpreting multimodal information from language and gesture [Thrun et al., 2008]. Our framework relies on a factored observation probability that fuses information from language and gesture in real time to continuously estimate the object a person is referring to in the real world. We demonstrate our model in simulation, as well as providing quantitative results on a real-world RGB-D corpus of people referring to objects with language

and gesture. These results show that our approach quickly and accurately fuses multimodal information in real time to continuously estimate the object a person is referencing. Additionally, we demonstrate a robot that uses our model to provide backchannel responses in real-time to a person's language and gesture input.

II. RELATED WORK

Clark [1996] proposed that conversation is a *joint activity*, a coordinated, collaborative processes akin to playing a duet or performing a waltz. The two participants must establish *common ground*. Common ground refers to the process of two conversational participants establishing joint understanding about the beliefs of the others[1]. To establish common ground, people use backchannel feedback, such as head nods, looks of confusion, as well as explicit request for clarification such as asking a question. These mechanisms enable the participants in a conversation to engage in a feedback loop to iteratively establish common ground as the conversation progresses. Our approach for interpreting language and gesture in real time provides a foundation for producing backchannel feedback with a robot, pointing toward increased robustness as a person and robot iteratively establish common ground and actively communicate to reduce errors.

A large body of work focuses on language understanding for robots [MacMahon et al., 2006, Dzifcak et al., 2009, Kollar et al., 2010, Matuszek et al., 2012]. This work does not take into account the continuous nature of natural language input, and requires sentences or at least chunks of words before understanding can take place. Our approach, in contrast, incorporates information from each word as it is processed by the speech recognition system, integrating word information over time and fusing it with gesture information. Guadarrama et al. [2014] presents a framework for interpreting open-domain references to objects but focuses on interpreting language rather than language combined with gesture. Cantrell et al. [2010] presented a framework for understanding language incrementally in real time dialog but did not use gesture and did not use a corpus-based evaluation. Our approach is related to Holladay et al. [2014] but focuses on interpreting a person's gestures rather than enabling a robot to generate pointing gestures.

Many existing approaches for interpreting gesture rely on fixed vocabularies of gesture, such as "stop" or "follow" [Waldherr et al., 2000, Marge et al., 2011] without a principled way for fusing information from language and gesture. Our work unifies language and gesture interpretation into a single mathematical framework, and focuses on parameterized gestures such as pointing.

Matuszek et al. [2014] presented a multimodal framework for interpreting unscripted references to tabletop objects using language and gesture. Our approach similarly focuses on tabletop objects but integrates language and gesture continuously over time using a Bayes' filtering framework. This approach enables the robot to continuously process new information and produce an estimate that converges over time to the correct object as new information is observed from the person.

POMDP-approaches to dialog [Young et al., 2013, Young, 2010] have been extended to incorporate gesture and noise models, and our approach is related to these types of belief updates. We are eager to explore enabling a robot to adaptively respond to its estimate of which object a person is referring to, leading to backchannels. Dragan and Srinivasa [2013] created a framework enabling a robot to produce gesture. Similarly, Tellex et al. [2014] described an approach for enabling a robot to generate language by inverting a semantics framework. Our long-term aim is that by combining these types of generation approaches with real-time understanding, the robot will produce back-channel feedback that closes the loop of dialog and enable it to participate in dialog as a joint activity.

[1]Note that common ground is in dialog is distinct from *symbol grounding* proposed by Harnad [1990], which is the problem of mapping from language to aspects of the external world.

III. TECHNICAL APPROACH

Our aim is to estimate a distribution over the object that a person is referring to given language and gesture inputs. We frame the problem as a Bayes' filter [Thrun et al., 2008], where the hidden state, $x \in \mathcal{X}$, is the the object in the scene that the person is currently referencing. The robot observes the person's actions and speech, $\mathcal{Z}$, and at each time step estimates a distribution over the current state, x_t:

$$p(x_t|z_0 \dots z_{0:t}) \tag{1}$$

To estimate this distribution, we alternate performing a time update and a measurement update. The time update updates the belief that the user is referring to a specific subset of objects given previous information:

$$p(x_t|z_{0:t-1}) = \int p(x_t|x_{t-1}) \times p(x_{t-1}|z_{0:t-1}) \mathrm{d}x_{t-1} \tag{2}$$

The measurement update combines the previous belief with the newest observation to update each belief state:

$$p(x_t|z_{0:t}) = \frac{p(z_t|x_t) \times p(x_t|z_{0:t-1})}{p(z_t|z_{0:t-1})} \tag{3}$$

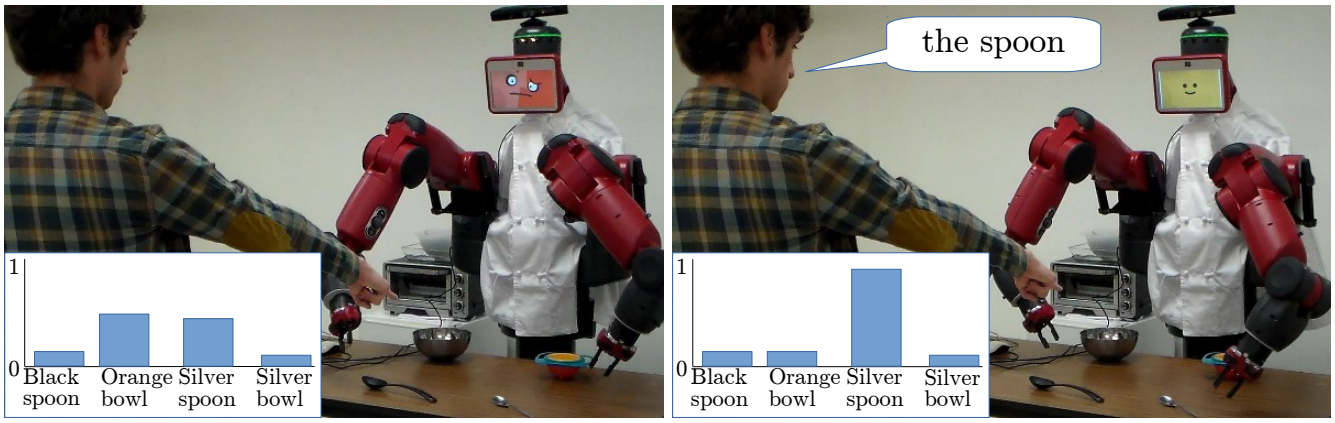
$$\propto p(z_t|x_t) \times p(x_t|z_{0:t-1}) \tag{4}$$

A. Prediction Model

We assume that a person is likely to continue referring to the same object, but at each timestep has a small probability, c, of transitioning to a different object:

$$p(x_t|x_{t-1}) = \begin{cases} 1-c & \text{if } x_t = x_{t-1} \\ c & \text{otherwise} \end{cases} \tag{5}$$

This assumption means that the robot's certainty slowly decays over time, in the absence of corroborating information, converging to a uniform distribution. It enables our framework to integrate past language and gesture information but also quickly adapt to new, conflicting information because it assumes the person has changed objects.



(a) Ambiguous gesture. (b) Clarification with language.

Fig. 2. After an ambiguous gesture, the model has a uniform distribution between two objects (a). The robot responds by indicating confusion. Clarification with language causes a probabilistic update leaving the model highly confident it has inferred the correct object (b). The robot responds by smiling and pointing to the correct object.

B. Observation Model

We assume access to an observation model of the form:

$$p(z_t|x_t) \tag{6}$$

Observations consist of a tuple consisting of a person's actions, $\langle l, r, h, s\rangle$ where:

- l represents the observed origin (l_o) and vector (l_v) for the left arm.
- r represents the observed origin (r_o) and vector (r_v) for the right arm.
- h represents the observed origin (h_o) and vector (h_v) for head.
- s represents the observed speech from the user, consisting of a list of words.

Formally, we have:

$$p(z_t|x_t) = p(l, r, h, s|x_t) \tag{7}$$

We factor assuming that each modality is independent of the others given the state (the true object that the person is referencing):

$$p(z_t|x_t) = p(l|x_t) \times p(r|x_t) \times p(h|x_t) \times p(s|x_t) \tag{8}$$

The following sections describe how we model each type of input from the person. We track body pose using the NITE tracker [ope, 2014] and define our models for the arm and head gestures in terms of the person's body pose.

Gesture. We model pointing gestures as a vector through three dimensional space. First, we calculate a gesture vector using the skeleton pose returned by NITE. For arms, we compute a vector from the elbow to the wrist, then project this vector so that the origin is at the write. For head pose, we compute a vector based on the body orientation. **ST: Is the previous sentence correct?** Next, we calculate the angle between the gesture vector and the vector from the gesture origin to the mean of each cluster, and then use the PDF of a Gaussian ($\mathcal{N}$) with variance (σ) to determine the weight that should be assigned to that object. We define a function $\mathrm{A}(o, p_1, p_2)$ as the angle between the two points, p_1 and p_2 with the given origin, o. Then

$$p(l|x_t) \propto \mathcal{N}(\mu_l = 0, \sigma_l, A(l_o, l_v, x_t)) \tag{9}$$

$$p(r|x_t) \propto \mathcal{N}(\mu_r = 0, \sigma_r, A(r_o, r_v, x_t)) \tag{10}$$

If the person's arm is more than a certain angle away from the table, we assume they are referring to none of the objects, and perform an update. As a result, these gestures do not effect the robot's estimate of the objects being referenced.

Head Pose. Head pose is modeled in the same manner as arm gestures.

$$p(h|x_t) \propto \mathcal{N}(\mu_h = 0, \sigma_h, A(h_o, h_v, x_t)) \tag{11}$$

Speech. We model speech as a bag of words. We take the words in a given speech input and count how many words in this text match descriptors (denoted x_d) of specific objects.

$$p(s|x_t) = \prod_{w \in s} p(w|x_t) \tag{12}$$

C. Null Words and Gesture

To account for continuous gesture and non-continuous speech input, we have both null poses and speech. When no words are spoken, we assume a null word which has a uniform distribution over the objects. This effect means that spoken words cause a discrete bump in probability according to the language model, which then decays over time. While gesture remains a continuous input throughout the entire interaction, many gestures have little or no meaning. To allow for these without overloading the model with noise, we also calculate the angle between each arm vector and each foot. If the angle between the arms and a foot is smaller than the angle between the arms and any object, we assume that the user is in a resting pose, and treat that gesture as indicating uniform probability over all states.

D. Model Parameters

We tuned model parameters by hand. We considered collecting and annotating a data set to train the model parameters, but we found our initial process to be quite accurate. We generated the language model by hand, adding to it based on results of our pilot studies. After our initial tuning, we fixed model parameters, and results reported in the paper all use the same fixed set of parameters. We expect that as we add larger sets of objects, a language model trained using data from Amazon Mechanical Turk or other corpora will be necessary to increase robustness over a larger set of objects.

In our experiments, we had the following parameters: the transition probability, c was 0.0005. We set this parameter to give an object that has 100% confidence an approximately 10% drop in confidence per second with all null observations. Standard deviation for the Gaussian used to model probability of gesture, σ_l, σ_r, and σ_h was 1.0 radians. We found that this standard deviation allowed for accurate pointing, without skewing the probabilities during an arm swing. The language model consisted of 16 unique words, containing common descriptors for the objects such as "bowl," "spoon," "metal," "shiny," etc. It also included words that were commonly misinterpreted by the speech recognition system, such as "bull" when the user was requesting a bowl.

Algorithm 1: Interactive Bayes Filtering Algorithm

Input: $bel(x_{t-1}), z_t$

Output: $bel(x_t)$

for x_t **do**

$\bar{bel}(x_t) = \prod_{x_{t-1}} p(x_t|x_{t-1}) * bel(x_{t-1})$

if not is_null_gesture(l)

$\bar{bel}(x_t) = p(l|x_t) * \bar{bel}(x_t)$

if not is_null_gesture(r)

$\bar{bel}(x_t) = p(r|x_t) * \bar{bel}(x_t)$

if not is_null_gesture(h)

$\bar{bel}(x_t) = p(h|x_t) * \bar{bel}(x_t)$

for $w \in s$ **do**

$\bar{bel}(x_t) = p(w|x_t) * \bar{bel}(x_t)$

end

$bel(x_t) = \bar{bel}(x_t)$

end

Algorithm 1 shows pseudocode for our approach, while Figure 2 shows an example of the system's execution. The person's gesture is ambiguous, and the system initially infers an approximately bimodal distribution between the orange bowl and silver spoon. The robot indicates it has not understood by showing a confused face. This reaction elicits a disambiguating response from the person, who says, "the spoon." The model incorporates information from language and infers the person is referring to the silver spoon. The robot indicates it has understood with a facial expression and by pointing to the correct object. Note that the person's linguistic response was itself ambiguous, but provided complementary information to the person's gesture, leading to overall success at interpreting their intent.

Although in this example we are demonstrating the approach at two specific timesteps, the system is updating its distribution continuously, enabling it to fuse language and gesture as it occurs and quickly updating in response to new input from the person, verbal or nonverbal. Our approach runs at 14Hz, including a 30Hz sleep cycle, on an Asus machine with 8 2.4 GHz Intel Cores that is also performing all perceptual and network processing.

TABLE I
SIMULATION RESULTS

	$\sigma^2 = 0.5$	$\sigma^2 = 1.0$
Language only	36.5%	37%
Head only	50.9%	35.8%
Arms only	62.4%	42.1%
Multimodal (all)	65.7%	54.1%

IV. EVALUATION

We evaluated our model in simulation, comparing the full model to versions without multimodal information. Additionally we assessed its performance on an RGB-D audio and video corpus of people referring to objects. Finally we created an end-to-end robotic demonstration, demonstrated in the **video** attachment to our paper, that is available online[2].

A. Simulation Results

We evaluated our approach in simulation by generating data from the model and assessing its accuracy at estimating the object being referenced. We generated simulated pointing data for each arm and the head, as well as spoken utterances at each timestep according to the model parameters. We then used these parameters to update the system's estimate of the object being referred to. Table I shows the results. Our accuracy metric is the fraction of time that the robot is pointing at the correct object. We report performance using language alone, gesture alone, and language combined with gesture. These results demonstrate that the system is successfully able to fuse multi-modal information to achieve higher accuracy than each modality alone, but do not show performance in the real world. To assess robustness to signal noise, we used two different levels of variance, demonstrating that even when the signal is highly noisy, our approach is able to fuse information across modalities to improve performance.

B. Real-World Corpus-Based Results

Our real-world experiments measured our algorithm's performance when a person referred to an object visually and with gesture. The subject stood in front of a table with four

[2]https://vimeo.com/107725829

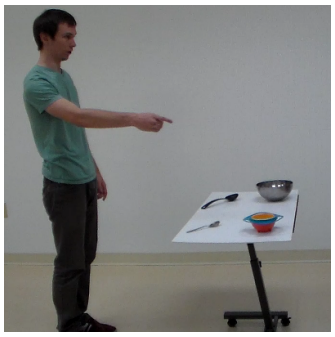


Fig. 3. Scene from our data collection environment.

objects placed approximately one foot apart, forming four corners of a square, as shown in Figure 3. We instructed subjects to ask for the indicated object in the most natural way possible, using whatever combination of gesture and language they felt was appropriate. We indicted the object to refer to using a laser pointer, and we periodically shifted to a different object on a predetermined schedule. They wore a microphone to pick up high-quality audio, and . We used the HTML5 Speech Recognition package in conjunction with Google Chrome to recognize speech. This package reports incremental output as recognition proceeds, and we perform a model update each time a new word is perceived. We used 13 subjects, and each subject participated in five trials, for a total of 65 trials.

Results showing the percent of the time the estimated most likely object was the true object appear in Table II with 95% confidence intervals. During a typical trial, the model starts out approximately uniform or unimodal on the previous object (we did not reset the model between trials.) As the subject points and talks, the model quickly converges to the correct object. Our first set of results give a sense of how quickly the model converges.

To assess overall accuracy, we report the system's accuracy at the end of a trial in III. Multimodal accuracy with language and gesture is more than 90%, demonstrating that our approach is able to quickly and accurately interpret unscripted language and gesture produced by a person. We found that our head pose estimator was quite inaccurate, performing at worse than change unimodally. Thus overall results that include head pose perform worse than language and gesture. We believe this effect is because people's head pose quickly changes, so the signal itself is noisy, and the OpenNI head pose estimator is also inaccurate.

The difference in accuracy between gesture alone and the multimodal output is not as large as one might expect. This is in part caused by the small delay in speech recognition software as opposed to the instantaneous gesture input. Additionally, many subjects, when told they could use gestures, leaned towards relying almost entirely on gesticulation. There were some users, however, who relied on an equal mix of both, and showed large leaps in accuracy between arms and multimodal. The most extreme example is of a user who, over their five trials, achieved only 45.5% accuracy

TABLE II

REAL-WORLD RESULTS

Random	25%
Language only	32.4% +/- 10%
Gesture only	73.12% +/- 9%
Head only	21.67% +/- 10%
Multimodal (Language and Gesture)	**81.99% +/- 5.5%**
Multimodal (All)	64.84% +/- 8%

TABLE III

REAL-WORLD RESULTS (END OF INTERACTION)

Random	25%
Language only	46.15%
Gesture only	80.0%
Head only	18.46%
Multimodal (Language and Gesture)	**90.77%**
Multimodal (All)	61.54%

with arms alone and 42.2% with speech alone, yet managed to achieve 85.7% multimodal accuracy, only 2 percentage points away from the sum of the two probabilities, showing the ease at which alternating speech and gesture can give incredibly accurate results overall. While a combination of ambiguous speech and gesture such as "that spoon" followed by a gesture would be more accurate than just a gesture, we found that most test subjects either spoke with complete ambiguity or none, using phrases either of the form "hand me that thing" or "hand me the silver spoon". Therefore we were unable to fully test this hypothesis.

C. Robotic Demonstration

Because our approach enables a robot to quickly monitor a person's references to an object, the robot can respond to these estimates in real time, producing backchannel feedback that can trigger a person to produce disambiguating language and gesture when the person does not understand, and move on to the next task when the robot does understand. We demonstrate this behavior by enabling Baxter to demonstrate its certainty about what object is being referenced, this eliciting more feedback from the person. When the robot is very unsure, it indicates its confusion by displaying a confused face, shown in Figure 2(a). The look of confusion triggers a disambiguating response from the person; this response causes the model to update and the robot responds by looking confident and moving its arm to point to the correct object. This **video** shows this behavior.

V. CONCLUSION

We have demonstrated a Bayes' filtering approach to interpreting a person's multimodal language and gesture references to objects continuously in real time. Our approach enables a robot to understand a person's references to objects in the real world. This paper represents steps toward continuous language understanding and the vision presented by Clark [1996] of language as joint activity.

In the future we plan to expand our language model to incorporate models of compositional semantics and lower-level visual features so that the robot is not limited to

prespecified object models. Additionally we aim to enable the robot to generate back-channel feedback based on its model using a POMDP framework, ultimately aiming to demonstrate that by providing appropriate backchannels, the robot elicits disambiguating responses from the person, increasing overall speed and accuracy of the interaction. Dragan and Srinivasa [2013] created a framework for generating legible gesture, and we anticipate that enabling a robot to respond by pointing as in Holladay et al. [2014] when it is sure and reflecting its confusion when it is unsure will enable the human-robot dyad to increase efficiency by naturally eliciting more information when the robot is confused and indicate that the person can move on when the robot has understood. We also plan to extend our approach beyond a unigram language model so that it supports models of compositional semantics by embedding a parsing chart into the state [Jurafsky et al., 1995, Earley, 1970]. These methods will enable the robot to understand nested referring expressions such as "the bowl on the table" incrementally. Finally, we aim to extend our approach beyond just object references; a similar modeling approach could be used to understand references to locations in the environment, and ultimately general command interpretation.

References

Openni tracker. http://wiki.ros.org/openni_tracker, 2014.

R. Cantrell, M. Scheutz, P. Schermerhorn, and X. Wu. Robust spoken instruction understanding for hri. In *Proceedings of the 5th ACM/IEEE international conference on human-robot interaction*, HRI '10, pages 275–282, New York, NY, USA, 2010. ACM. ISBN 978-1-4244-4893-7.

H. H. Clark. *Using Language*. Cambridge University Press, May 1996. ISBN 0521567459.

A. Dragan and S. Srinivasa. Generating legible motion. In *Robotics: Science and Systems*, June 2013.

J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn. What to do and how to do it: translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, pages 3768–3773, Piscataway, NJ, USA, 2009. IEEE Press. ISBN 978-1-4244-2788-8.

J. Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102, 1970.

S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell. Open-vocabulary object retrieval. RSS, 2014.

S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335 – 346, 1990. ISSN 0167-2789.

R. M. Holladay, A. D. Dragan, and S. S. Srinivasa. Legible robot pointing. RO-MAN, 2014.

D. Jurafsky, C. Wooters, J. Segal, A. Stolcke, E. Fosler, G. Tajchaman, and N. Morgan. Using a stochastic context-free grammar as a language model for speech recognition. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 189–192. IEEE, 1995.

T. Kollar, S. Tellex, D. Roy, and N. Roy. Toward understanding natural language directions. In *Proceedings of HRI-2010*, 2010.

M. MacMahon, B. Stankiewicz, and B. Kuipers. Walk the talk: connecting language, knowledge, and action in route instructions. In *Proceedings of the 21st national conference on artificial intelligence*, volume 2 of *AAAI '06*, pages 1475–1482. AAAI Press, 2006. ISBN 978-1-57735-281-5.

M. Marge, A. Powers, J. Brookshire, T. Jay, O. C. Jenkins, and C. Geyer. Comparing heads-up, hands-free operation of ground robots to teleoperation. *Robotics: Science and Systems VII*, 2011.

C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox. Learning to parse natural language commands to a robot control system. In *Proc. of the 13th Intl Symposium on Experimental Robotics (ISER)*, 2012.

C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox. Learning from unscripted deictic gesture and language for human-robot interactions. 2014.

S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. AAAI*, 2011.

S. Tellex, R. Knepper, A. Li, D. Rus, and N. Roy. Asking for help using inverse semantics. In *Robotics: Science and Systems (RSS)*, 2014.

S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. MIT Press, 2008.

S. Waldherr, R. Romero, and S. Thrun. A gesture based interface for human-robot interaction. *Autonomous Robots*, 9(2):151–173, 2000.

S. Young. Cognitive user interfaces. *Signal Processing Magazine, IEEE*, 27(3):128–140, 2010.

S. Young, M. Gašić, B. Thomson, and J. D. Williams. Pomdp-based statistical spoken dialog systems: A review. 2013.