Temporal Grounding Graphs for Language Understanding with Accrued Visual-Linguistic Context

Rohan Paul and **Andrei Barbu** and **Sue Felshin** and **Boris Katz** and **Nicholas Roy***
Massachusetts Institute of Technology, Cambridge, MA

Abstract

A robot's ability to understand or *ground* natural language instructions is fundamentally tied to its knowledge about the surrounding world. We present an approach to grounding natural language utterances in the context of factual information gathered through natural-language interactions and past visual observations. A probabilistic model estimates, from a natural language utterance, the objects, relations, and actions that the utterance refers to, the objectives for future robotic actions it implies, and generates a plan to execute those actions while updating a state representation to include newly acquired knowledge from the visual-linguistic context. Grounding a command necessitates a representation for past observations and interactions; however, maintaining the full context consisting of all possible observed objects, attributes, spatial relations, actions, etc., over time is intractable. Instead, our model, *Temporal Grounding Graphs*, maintains a learned state representation for a belief over factual groundings, those derived from natural-language interactions, and lazily infers new groundings from visual observations using the context implied by the utterance. This work significantly expands the range of language that a robot can understand by incorporating factual knowledge and observations of its workspace in its inference about the meaning and grounding of natural-language utterances.

1 Introduction

Effective human-robot interaction in homes or other complex dynamic workspaces requires a linguistic interface; robots should understand what owners want them to do. This is only possible with a shared representation of the environment, both past and present, as well as a mutual understanding of the knowledge exchanged in prior linguistic interactions. At present, humans and robots possess disparate world representations which do not lend themselves to enabling natural human-robot interactions. Humans possess a continuously-expanding rich understanding of the environment consisting of semantic entities and higher-order relationships which includes knowledge about previous interactions and events. In contrast, robots estimate a metric picture of the world from their sensors which determine which action sequences to take. Our goal is to bridge this *semantic gap* and enable robots, like humans, to acquire higher-order semantic knowledge about the environment through experience and use that knowledge to reason and act in the world. To this end, we develop an approach to building robotic systems which follow commands in the context of accrued visual observations of their workspace and prior natural-language interactions. Natural-language interactions can include both commands to perform an action or important factual information about the environment. The robot understands and retains facts for later use and is able to execute the intended tasks by reasoning over acquired knowledge from the past.

In recent years, probabilistic models have emerged that interpret or *ground* natural language in the context of a robot's world model. Approaches such as those of Tellex *et al.* [2011], Howard *et al.* [2014], Chung *et al.* [2015], and Paul *et al.* [2016], relate input language to entities in the world and actions to be performed by the robot by structuring their inference according to the parse or semantic structure embedded in language. A related set of approaches such as those of Zettlemoyer and Collins [2007], Chen *et al.* [2010], Kim and Mooney [2012], and Artzi and Zettlemoyer [2013] employ semantic parsing to convert a command to an intermediate representation, either λ-calculus or a closely related logic formalism, which can then be executed by a robot. A key limitation of current models is their inability to reason about or reference past observations. In essence, many models assume that the world is static, facts do not change, and that perception is entirely reliable. Matuszek *et al.* [2014] present a formulation that allows uncertainty in the knowledge about object in the scene and present an approach that learns object attributes and spatial relations through joint reasoning with perceptual features and language information. In a similar vein, Chen and Mooney [2011], Guadarrama *et al.* [2013], Walter *et al.* [2014], Hemachandra *et al.* [2015], and Andreas and Klein [2015], demonstrate systems more resilient in uncertainty in the input given the need for the robot to carry out actions. Even a simple statement such as "The fruit I placed on the table is my snack" followed by the command "Pack up my snack" requires reasoning about previous observations of

*R. Paul and A. Barbu contributed equally to this work. R. Paul and N. Roy are with the Robust Robotics Group and A. Barbu, S. Felshin and B. Katz are with the Center for Brains, Minds, and Machines and the Computer Science and AI Lab (CSAIL) at MIT. Contact: {rohanp, abarbu, sfelshin, boris, nickroy}@csail.mit.edu

the actions of an agent. While work such as that of Misra *et al.* [2016], and Liu *et al.* [2016] incorporates context it does so by learning the static relationships between properties of the environment and actions to be performed; a different task from learning facts online and remembering the actions of agents.

Complementary efforts in many communities have focused on acquiring rich semantic representations from observations or knowledge-based systems. These include efforts in the vision community towards compositional activity recognition [Barrett *et al.*, 2016; Yu *et al.*, 2015] and work such as that by Berzak *et al.* [2016] which combines activity recognition with language disambiguation. Further, approaches for associating language with semantic constructs have been explored in the semantic parsing community [Berant *et al.*, 2013]. Cantrell *et al.* [2010] investigate a closely-related problem, grounding robotic commands which contain disfluencies to objects in images. We consider a broader notion of grounding than Cantrell *et al.* [2010] in which we include the actions of agents as well as knowledge from prior linguistic interactions but note that their approach is complementary and could be used to increase the robustness of the work presented here.

We seek the ability to reason about the future actions of a robot using knowledge from past visual observations and linguistic interactions. Enabling such reasoning entails determining *which* symbols grounded by past observations should be retained to enable future inferences. One approach is to estimate and propagate all symbols from past percepts. This is intractable as the space of all static or dynamic semantic relations between all observed objects is exponentially large and continues to grow exponentially as new relations are learned. Alternatively, one can forgo symbol propagation and retain raw observations alone. This approach incurs a linear storage cost, but requires jointly interpreting the current utterance with all past utterances, combinatorially increasing the inference cost with each utterance.

In this work, we present *Temporal Grounding Graphs*, a probabilistic model that enables incremental grounding of natural language utterances using learned knowledge from accrued visual observations and language utterances. The model allows efficient inference over the constraints for future actions in the context of a rich space of perceptual concepts such as object class, shape, color, relative position, mutual actions, etc. as well as factual concepts (does an object belong to an agent such as a human) that grow over time. Crucially, the approach attempts to balance the computational cost of incremental inference versus the space complexity of knowledge persistence. Our model maintains a learned representation as a belief over factual relations from past language. The model accrues visual observations but delays estimation of grounding of perceptual concepts. Online, the model estimates the necessary past visual context required for interpreting the instruction. The context guides the construction of a constrained grounding model that performs focused time-series inference over past visual observations. This approach incurs an additional inference cost online but reduces the need for exhaustively estimating all perceptual groundings from past observations. Factors in the model are trained in a data-driven manner using an aligned vision-language corpus. We demonstrate the approach on a Baxter Research Robot following and executing complex natural language instructions in a manipulation domain using a standardized object data set.

2 Problem Formulation

We consider a robot manipulator capable of executing a control sequence μ_t composed of an end-effector trajectory $\{\mu_t(t'_i), \ldots, \mu_t(t'_f)\}$ initiated at time t. The robot's workspace consists of objects O that may be manipulated by the robot or other agents such as the human operator. We assume that each object possesses a unique identifier (a symbol) and an initial metric pose known *a-priori* to the robot. The robot observes its workspace through a visual sensor that collects an image $\mathcal{I}_t$ at time t, each associated with metric depth information for localization within the environment. Further, we assume the presence of an object recognition system for a known set of object classes which yields potential detections (each a sub-image) with an associated class likelihood. Let Z_t denote the set of detections $\{z_0^t, \ldots, z_{n_t}^t\}$ from I_t where n_t is the total number of detections. The human operator communicates with the robot through a natural language interface (via speech or text input) either instructing the robot to perform actions or providing factual information about the workspace. Let Λ_t denote the input language utterance from the human received at time t which can be decomposed into an ordered tree-structured set of phrases $\{\lambda_1, \ldots, \lambda_n\}$ using a parsing formalism. We consider the problem of enabling the robot to understand the language utterance from the human in the context of learned workspace knowledge from past visual observations or factual information provided by the human. Next, we define the space of concepts that represent *meaning* conveyed in a language utterance and subsequently formalize the grounding problem.

2.1 Grounding Symbols

We define a space of concepts that characterize the semantic knowledge about the robot's workspace. The robot's workspace can be expressed as a set of symbols associated with semantic entities present in the environment. These include object instances, the human operator and the robot itself, and form the symbolic world model Υ. Concepts convey knowledge about the properties or attributes associated with entities or relationships between sets of entities. It is common to employ a symbolic predicate-role representation for concepts [Russell and Norvig, 1995]. A predicate expresses a relation $\sigma \in \Omega$ defined over a set of symbols $\mathcal{R}$, each associated with an entity in the world model Υ. The space of possible predicates that may be true for a world representation forms the space of grounding symbols Γ:

$$\Gamma = \left\{ \gamma^{\sigma}_{\mathcal{R}} \mid \sigma \in \Omega, \mathcal{R} \subseteq \Upsilon \right\}. \tag{1}$$

For example, if the workspace model consists of a human and a box denoted as symbols o_1 and o_2, the grounding for the event that involves a person lifting the box object is represented by the grounding symbol $\text{PickUp}(o_1, o_2)$.

Grounding symbols can be categorized in terms of the type of concepts they convey about the workspace. A set of *declarative* grounding symbols Γ^{δ} convey knowledge about the robot's workspace. These include perceptual concepts $\Gamma^{\mathcal{P}}$ consisting of entities such as objects and agents, static relations like spatial regions (e.g, on, left, front) and event relations that denote interactions (e.g., place, approach, slide) between agents and objects present in the scene. Further, declarative groundings include *factual* concepts $\Gamma^{\mathcal{F}}$ consisting of arbitrary relations representing abstract (non-perceptual) knowledge. As an example, the phrase "the cup on the tray is mine and the fruit

Table 1: The space of grounding symbols and their arities. Object classes are provided and learned by object detectors. Regions, modifiers, human actions and planner constraints are learned as described in Section 3.4. The space of factual concepts is open and expands through natural-language interaction.

Agents	Robot_1, Human_1
Objects	Block_1, Can_1, Box_1, Fruit_1, Cup_1, ...
Regions	LeftOf_2, InFrontOf_2, OnTopOf_2, ...
Modifiers	Quickly_1, Slowly_1, Big_2, Red_2, ...
Human actions	PickUp_2, PutDown_2, Approach_2, ...
Factual concepts	Mine_1, Favourite_1, Forbidden_1, ...
Planner constraints	Intersect_2, Contact_2, SpatialRelation_2, ...

on the right is fresh" conveys notions of *possession* and *being fresh* that are factual. The space of facts grows through linguistic interactions where any previously-unknown property of an object is considered a fact. Facts can be true of multiple entities simultaneously. For example, in the statement, "the fruit and the box on the table are my snack", the notion of *snack* includes the two indicated objects in the scene.

A set of *imperative* groundings Γ^{π} describe the objectives or goals that can be provided to a robot motion planner to create a set of robot motions. Imperative groundings are characterized by a type of motion (e.g., picking, placing or pointing) and a set of spatial constraints (e.g., proximity, intersection or contact) that must be satisfied by the executed action. Further, we include aggregative constraints conveyed as conjunctive references (e.g., "the block and the can") or associations (e.g., "a set of blocks"). Finally, the set of imperative groundings also includes a symbol that conveys the assertion of an inferred factual grounding[1]. The space of grounding symbols can be cumulatively represented as $\Gamma = \Gamma^{\pi} \cup \Gamma^{\mathcal{P}} \cup \Gamma^{\mathcal{F}}$. Table 1 lists the representative set of grounding predicates used in this work.

2.2 Language Grounding with Context

The problem of understanding or *grounding* a natural language utterance involves relating the input language with semantic concepts expressed in the workspace. This process involves estimating the probable set of groundings that best convey the intended meaning of the perceived language utterance. For example, the interpretation for the utterance, "Pick up the block on the table" can be represented by the grounding set $\textsc{Block}(o_1) \wedge \textsc{EndEffector}(o_2) \wedge \textsc{Table}(o_3) \wedge \textsc{On}(o_1,o_3) \wedge \textsc{Contact}(o_2,o_1)$ where object symbols o_1, o_2 and o_3 are derived from the symbolic world model. The estimated association between the input language utterance and the set of grounding symbols expresses the intended meaning of the sentence and can be considered as determining a *conceptual* grounding for the utterance. Further, the set of object symbols that parameterize the estimated grounding symbols must be associated with the set of visual percepts arising with the geometric objects populating the workspace. This process can be viewed as estimating *existential* groundings. In the above example, this would entail associating object symbols o_1 and o_2 with the set of detections $Z_{0:t}$ that correspond to the physical block and table. In essence, the existential grounding process accounts for uncertainty in the robot's perception of semantic entities present in the environment.

[1] As we discuss later in Section 3, an imperative grounding associated with an asserted fact is used to propagate factual knowledge.

Following [Tellex *et al.*, 2011], the grounding process is mediated by a binary correspondence variable $\phi_{ij} \in \Phi$ that expresses the degree to which the phrase $\lambda_i \in \Lambda$ corresponds to a possible grounding $\gamma_j \in \Gamma$. This allows groundings to be expressed as probabilistic predicates modeling the uncertainty in the degree of association of a concept with a phrase. Further, limiting the domain of correspondence variables to a *true* or a *false* association, allows the factors relating phrases and candidate groundings to be locally normalized, which reduces the learning complexity. Similarly, we introduce correspondence variables Φ^{O} that convey probable associations between object symbols in the world model Υ and visual detections $Z_{0:t}$.

A note on convention: in the remainder of this paper, we use the phrase "grounding natural language" to imply the *conceptual* grounding of an input language utterance to the space of grounding symbols that characterize workspace knowledge acquired by the robot. We address the issue of estimating correspondences between object symbols in the world model with the visual percepts in Section 3.3. Further, in our formulation the grounding for an utterance is obtained by determining the likely true correspondences. For brevity at times we refer to "estimating true correspondences between phrases in the utterance and grounding symbols" as simply "estimating groundings for language". Finally, we use capital symbols in equations to refer to sets of variables.

We now discuss the temporally extended scenario where estimating the grounding for an input language utterance involves reasoning over an accrued context of past visual observations and language utterances. For example, consider the scenario where a robot observes a human place a can in the scene. This is followed by the human uttering "the can that I put down is my snack" and commanding the robot to "pack up my snack". Interpreting this sequence of language utterances requires inferring and reasoning with factual information, such as which objects constitute the "snack", recognizing that the "put down" action is associated with the "can", and inferring that the robot is instructed to execute a "lift and place" action sequence to satisfy the command. Given the context of visual observations $Z_{0:t}$ and utterances $\Lambda_{0:t-1}$ leading up to time t, the problem of estimating the set of true correspondences Φ_t for the utterance Λ_t and control actions μ_t can be posed as:

$$P(\mu_t, \Phi_t | \Lambda_{0:t}, Z_{0:t}, \Gamma_t). \tag{2}$$

In the following section, we develop a probabilistic model for estimating this distribution.

3 Probabilistic Model

In this section, we present the *Temporal Grounding Graphs* model for interpreting natural language utterances with past visual-linguistic context of observations of the world coupled with descriptive language. We pose the problem of estimating Equation 2 as "filtering" on a dynamic Bayesian network and subsequently detail the model for estimating groundings for an utterance at each time step. Further, we discuss model training and present an analysis for runtime and space complexity for the model.

3.1 Temporal Model

We pose the problem of computing Equation 2 as incremental estimation on a temporal model. The interpretation of a current utterance Λ_t can rely on the background knowledge

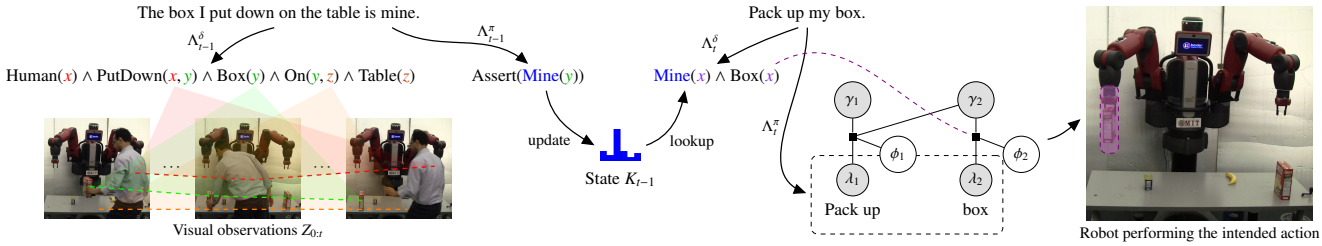


Figure 1: An overview of grounding commands from the initial utterances to a robotic action. Utterances are jointly grounded into declarative and imperative grounding symbols. Perceptual groundings are computed jointly with object tracks, shown with corresponding colors. Factual knowledge is aggregated and used to disambiguate later commands. A grounding graph is illustrated for a command showing how the structure of the model mirrors that of the sentence. Probable imperative correspondences are estimated conditioned on the declarative groundings and prior state. They determine the constraints used by a motion planner to generate an action sequence and direct the robot's behaviour.

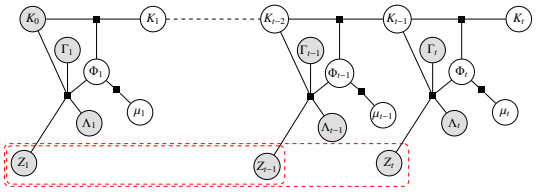


Figure 2: Grounding an instruction Λ_t with past visual-linguistic observations $\{Z_{0:t}, \Lambda_{0:t-1}\}$ is posed as filtering on a temporal model. A state variable K_t provides minimal state persistence over factual knowledge derived from past language inputs $\Lambda_{0:t-1}$. Estimating perceptual groundings from past visual observations $Z_{0:t}$ is delayed, variables encircled with red plate. The declarative context within the instruction Λ_t focuses inference over past visual observations.

derived from the past language descriptions $\Lambda_{0:t-1}$ or semantic information derived from accrued visual observations $Z_{0:t}$. As a result, conditional dependencies exist between the interpretation for current utterance and past observations. The direct estimation of this distribution is challenging since the interpretation of the current command Λ_t may involve reference to past knowledge that must be estimated jointly from past observations, creating conditional dependencies that grow with time. One approach is to introduce a state variable incorporating inferred past groundings that is propagated over time. An explicit state variable has the advantage of making the past observations conditionally independent of future observations given the current state, but such a variable must maintain all perceptual groundings derived from the history of the visual input. The space of perceptual groundings $\Gamma_t^{\mathcal{P}}$ is exponentially large as it must include all possible relationships between entities in the scene that may be referenced by a language utterance and is infeasible to maintain for complex scenes over time. Alternatively, the model can accrue observations and directly estimate Equation 2 at each time step. Although this approach reduces the space complexity for maintaining context, the cost of incremental inference is high since the current and past language utterances must be interpreted jointly.

We seek to balance the complexity of context maintenance with efficiency of incremental estimation and introduce a state variable denoted by K_t that expresses the belief over factual grounding symbols $\Gamma_t^{\mathcal{F}}$. We assume that factual groundings are uncorrelated. The likelihood over the state can be factored as a product of individual grounding likelihoods as:

$$K_t = \left\{ \gamma^{\sigma}_{\mathcal{R}_t} \mid \forall \sigma \in \Omega^{\mathcal{F}}, \forall \mathcal{R}_t \subseteq \Upsilon_t \right\} P(K_t) = \prod_{i=1}^{|\Gamma_t^{\mathcal{F}}|} P\left(\gamma^{\sigma}_{\mathcal{R}_t i}\right). \quad (3)$$

Note that the likelihood of the state variable expresses the likelihood *over* the factual grounding variables. This is in contrast to the grounding likelihood that explicitly models the degree of association *between* an utterance and a factual concept. The inclusion of the state variable in Equation 2 at each time step enables the following factorization:

$$P(\mu_t, \Phi_t, K_t | \Lambda_{0:t}, Z_{0:t}, \Gamma_t) =$$
$$\sum_{K_{t\text{-}1}} \overbrace{P(\mu_t, \Phi_t, K_t | \Lambda_t, Z_{0:t}, K_{t\text{-}1}, \Gamma_t)}^{\text{Current inference}} \overbrace{P(K_{t\text{-}1} | \Lambda_{0:t\text{-}1}, Z_{0:t\text{-}1}, \Gamma_t)}^{\text{Previous state estimate}}. \quad (4)$$

The state variable K_{t-1} represents the cumulative belief over factual groundings till the previous time step $t-1$. The state variable K_{t-1} decouples the estimation of the grounding Φ_t for the current command Λ_t from the past language inputs $\Lambda_{0:t-1}$. The factorization introduced in Equation 4 poses the inference over the full time history as filtering at each time step over groundings for the command and associated actions. Further, the online estimation process also propagates or updates the state variable informed by the visual and language observations. Figure 2 illustrates the unrolled temporal model. Crucially, we only propagate factual groundings $\Gamma_t^{\mathcal{F}}$. The set of visual detections $Z_{0:t}$ are stored directly as part of context. The inference over perceptual groundings $\Gamma_t^{\mathcal{P}}$ is delayed till a future utterance is received, a process we outline next.

3.2 Grounding Network

We now detail the model for inference instantiated at each time step within the dynamic Bayesian network outlined in the previous section. The grounding model estimates the likely set of correspondences Φ_t, the control sequence μ_t and the updated knowledge state K_t from the utterance Λ_t at time t using the prior state K_{t-1} and visual observations $Z_{0:t}$. We assume that the estimated groundings of an input command are sufficient for determining the output robot actions. The inference over the control sequence μ_t is decoupled from other variables given the knowledge of expressed correspondences Φ_t. Further, we assume that given the prior state K_{t-1} and the observed correspondences Φ_t, the updated state variable K_t is conditionally independent of other variables in the model. The above assumptions factorize the joint likelihood as:

$$P(\mu_t, \Phi_t, K_t | \Lambda_t, Z_{0:t}, K_{t-1}, \Gamma_t) =$$
$$\overbrace{P(\mu_t | \Phi_t)}^{\text{Planner}} \overbrace{P(K_t | \Phi_t, K_{t-1})}^{\text{State update}} \overbrace{P(\Phi_t | \Lambda_t, Z_{0:t}, K_{t-1}, \Gamma_t)}^{\text{Grounding model}}. \quad (5)$$

In the above, the grounding model $P(\Phi_t | \Lambda_t, Z_{0:t}, K_{t-1}, \Gamma_t)$ estimates the correspondences for the input instruction Λ_t given

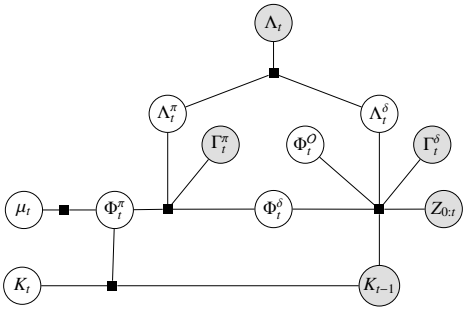


Figure 3: The grounding network for inference at each time instant t. An utterance Λ_t is grounded through imperative correspondences, Φ_t^π, related to future robot actions, and declarative correspondences, Φ_t^δ, relating perception and factual knowledge. A parsing mechanism facilitates this by partitioning Λ_t into $\{\Lambda_t^\pi, \Lambda_t^\pi\}$. The declarative grounding factor incorporates visual observations $Z_{0:t}$ and prior state estimate K_{t-1}. Estimation of imperative groundings is conditioned on declarative correspondences and yields constraints for planning robot actions μ_t. Estimated groundings update and propagate the state estimate K_t.

context accrued till the present time step $\{K_{t-1}, Z_{0:t}\}$. Note that an input language utterance may instruct the robot to perform a task, refer to past or current knowledge about the workspace or provide factual knowledge for future inferences. Estimating the conditional likelihood $P(\Phi_t|\Lambda_t, Z_{0:t}, K_{t-1}, \Gamma_t)$ involves determining the set of true imperative correspondences Φ_t^π, the declarative correspondences Φ_t^δ and the set of correspondences Φ_t^O that object symbols in the world model Υ with visual percepts $Z_{0:t}$. Partitioning the set of grounding variables as imperative and declarative sets allows Equation 5 as:

$$P(\mu_t, \Phi_t, K_t|\Lambda_t, Z_{0:t}, K_{t-1}, \Gamma_t) = \sum_{\Lambda_t^\pi, \Lambda_t^\delta} \overbrace{P(\mu_t|\Phi_t^\pi)}^{\text{Planner}} \overbrace{P(K_t|\Phi_t^\pi, K_{t\text{-}1})}^{\text{State update}}$$
$$\underbrace{P(\Phi_t^\pi|\Lambda_t^\pi, \Phi_t^\delta, \Gamma_t^\pi)}_{\text{Imperative grounding}} \underbrace{P(\Phi_t^\delta, \Phi_t^O|\Lambda_t^\delta, Z_{0:t}, K_{t\text{-}1}, \Gamma_t^\delta)}_{\text{Declarative grounding}} \underbrace{P(\Lambda_t^\pi, \Lambda_t^\delta|\Lambda_t)}_{\text{Parsing}}. \quad (6)$$

Figure 3 illustrates the resulting factor graph. The declarative grounding factor $P(\Phi_t^\delta, \Phi_t^O|\Lambda_t^\delta, Z_{0:t}, K_{t\text{-}1}, \Gamma_t^\delta)$ involves estimating perceptual and factual groundings conditioned on the context of visual observations and the propagated knowledge state. Further, this factor jointly estimates the likely correspondences that connect the symbols in the world model with percepts obtained by the visual sensor. The estimation of imperative groundings is conditioned on the declarative grounding variables described by the likelihood $P(\Phi_t^\pi|\Lambda_t^\pi, \Phi_t^\delta, \Gamma_t^\pi)$. The declarative grounding variables decouple the estimation of the imperative groundings from the past visual observations and the current knowledge state. The inclusion of declarative grounding variables in the factor estimating the imperative grounding enables disambiguation of the future actions that the robot may perform. For example, consider the utterance, "lift the box that was placed by the person on the table" where the intended manipulation of the box is constrained by the knowledge of a past interaction between the person and the specific box. The induced factorization in Equation 6 relies on partitioning a set of linguistic constituents as imperative (future actions) or declarative (knowledge from the past). Parsing of the instruction and estimating a partitioning of the constituents set as imperative or declarative is modeled by the factor $P(\Lambda_t^\pi, \Lambda_t^\delta|\Lambda_t)$.

We assume that the estimated imperative groundings derived from a language command fully characterize the goals and objectives for future robot actions. The estimation of the control sequence generation is conditioned on the estimated imperative groundings alone expressed in the factor $P(\mu_t|\Phi_t^\pi)$. If the language utterance conveys factual information about entities in the world, then the estimated factual groundings for the utterance serve as a "measurement" for knowledge state. The estimated imperative grounding Φ_t^π involves an action to update the prior knowledge K_{t-1} using the estimated factual groundings to form the propagated state K_t, modeled in the likelihood $P(K_t|\Phi_t^\pi, K_{t\text{-}1})$.

3.3 Factors

In this section, we detail the individual models that constitute the ground network represented by Equation 6.

Dependency Parse and Syntactic Analysis

The parsing model $P(\Lambda_t^\pi, \Lambda_t^\delta|\Lambda_t)$ is realized using a dependency parser, START [Katz, 1988], that factors the input language Λ_t into a set of constituents related to each other using syntactic dependency relationships. The structure of the computed linguistic relations informs the structure of the grounding models such that conditional dependencies between grounding variables follow the parse structure. The set of constituents and syntactic dependencies, as well as the morphological and lexical features provided by START, determine if a constituent plays an imperative Λ_t^π and/or declarative Λ_t^δ role in the input utterance. This partitioning informs the grounding factors $P(\Phi_t^\pi|\Lambda_t^\pi, \Phi_t^\delta, \Gamma_t^\pi)$ and $P(\Phi_t^\delta, \Phi_t^O|\Lambda_t^\delta, Z_{0:t}, K_{t\text{-}1}, \Gamma_t^\delta)$. Any linguistic constituents which are unknown are ignored aside from those associated with a single entity. These are assumed to be new facts about that entity which enables the corpus of factual knowledge to expand online.

Declarative Grounding Model

The declarative grounding model estimates the set of correspondences for declarative groundings Φ_t^δ. These include facts, static spatial relations, and dynamic relations such as events. Further, the factor also estimates the correspondences Φ_t^O between objects symbols and percepts. The estimation is conditioned on the set of declarative linguistic constituents Λ_t^δ derived from the parse of an utterance, the set of visual observations $Z_{0:t}$ and the prior state K_{t-1}. Since factual relations express abstract or non-perceptual knowledge, the set of factual groundings is decoupled from visual observations conditioned on the utterance and the prior state. This enables the following factorization where we assume an undirected representation:

$$\Psi(\Phi_t^\delta, \Phi_t^O, \Lambda_t^\delta, Z_{0:t}, K_{t-1}, \Gamma_t^\delta) = \overbrace{\Psi(\Phi_t^\mathcal{F}, \Lambda_t^\delta, K_{t-1}, \Gamma_t^\mathcal{F})}^{\text{Factual grounding}} \overbrace{\Psi(\Phi_t^\mathcal{P}, \Phi_t^O, \Lambda_t^\delta, Z_{0:t}, \Gamma_t^\mathcal{P})}^{\text{Perceptual grounding}}. \quad (7)$$

The factor $\Psi(\Phi_t^\mathcal{F}, \Lambda_t^\delta, K_{t-1}, \Gamma_t^\mathcal{F})$ models the likelihood of a factual grounding. The conditional likelihood is inferred from language $\Lambda_t^\mathcal{F}$ and the state variable K_{t-1} which includes the estimated belief over factual knowledge propagated from a past context.

The factor $\Psi(\Phi_t^\mathcal{P}, \Phi_t^O, \Lambda_t^\delta, Z_{0:t}, \Gamma_t^\mathcal{P})$ models the joint likelihood over perceptual groundings $\Phi_t^\mathcal{P}$ and correspondences

between world model symbols and percepts $\Phi_t^{\mathcal{O}}$ given language Λ_t^{δ} and accrued visual observations $Z_{0:t}$. This factor is realized using a vision-language framework developed by Barrett *et al.* [2016], the *Sentence Tracker*. It models the correspondences between visual observations and an event described by a sentence as a factorial hidden Markov model [Ghahramani *et al.*, 1997]. Given a parse of an input sentence it first determines the number of participants in the event described by that sentence. Next, it instantiates one tracker HMM for each participant. Tracker HMMs model sequences of object detections that both have high likelihood and are temporally coherent and observe the visual input $Z_{0:t}$ to estimate object tracks from raw detections. Tracker HMMs are in turn observed by declarative grounding HMMs. An HMM for a declarative symbol encodes the semantics of that symbol as a model for one or more sequences of detections; for example *approach* can be modeled as an HMM with three states: first, the objects are far apart, then they are closing, and finally they are close together. Grounding models observe trackers according to the parse structure of the sentence thereby encoding subtle differences in meaning, like the difference between being given a block and giving away a block. For example, given the sentence "The person put down the block on the table" and a parse of that sentence as $\textsc{Human}(x) \wedge \textsc{Block}(y) \wedge \textsc{PutDown}(x, y) \wedge \textsc{Table}(z) \wedge \textsc{On}(y, z)$ it instantiates three trackers, one for each variable, and five declarative grounding models, one for every predicate, and connects them according to the predicate argument structure of the parse. Inference simultaneously tracks and recognizes the sentence finding the globally-optimal object tracks for a given sentence and set of object detections. The formulation described above expresses Equation 7 as:

$$\Psi(\Phi_t^{\delta}, \Phi_t^{\mathcal{O}}, \Lambda_t^{\delta}, Z_{0:t}, K_{t-1}, \Gamma_t^{\delta}) = \overbrace{\Psi(\Phi_t^{\mathcal{F}}, \Lambda_t^{\delta}, K_{t-1}, \Gamma_t^{\mathcal{F}})}^{\text{Factual grounding}} \overbrace{\Psi(\Phi_t^{\mathcal{P}}, \Lambda_t^{\delta}, \Phi_t^{\mathcal{O}}, \Gamma_t^{\mathcal{P}})}^{\text{Event grounding}} \overbrace{\Psi(\Phi_t^{\mathcal{O}}, Z_{0:t})}^{\text{Object tracking}}. \quad (8)$$

It is exactly this inference over correspondences $\{\Phi_t^{\delta}, \Phi_t^{\mathcal{O}}\}$ to groundings that allows us to forget the raw observations when grounding future instructions that refer to events in the past.

Imperative Grounding Model

We now discuss the factor $P(\Phi_t^{\pi}|\Lambda_t^{\pi}, \Phi_t^{\delta}, \Gamma_t^{\pi})$ that estimates the correspondences Φ_t^{π} for the imperative instruction given the imperative part of the utterance Λ_t^{π} and the set of declarative correspondence Φ_t^{δ} described previously. These sets of imperative groundings include the goals and constraints that are provided to a robot planner to generate the motion plan to satisfy the human's intent. This factor is realized by extending the *Distributed Correspondence Graph* (DCG) formulation [Howard *et al.*, 2014; Paul *et al.*, 2016] which efficiently determines the goal objectives (the object(s) to act upon) and motion constraints (contact, proximity, visibility etc.) from natural-language instructions. For example, an utterance like "lift the farthest block on the right" results in contact constraints with one object, a block, which displays the spatial properties, "it's on the right", and relations, "it's the furthest one", implied by the sentence. Further, the model allows estimation of aggregate constraints implied in sentences like "pick up the can and the box".

Formally, the imperative grounding likelihood can be expressed in the undirected form as:

$$\Psi(\Phi_t^{\pi}, \Lambda_t^{\pi}, \Phi_t^{\delta}, \Gamma_t^{\pi}) = \prod_{i=1}^{|\Lambda_t^{\pi}|} \prod_{j=1}^{|\Gamma^{\pi}|} \overbrace{\Psi(\phi_{ij}^{\pi}, \lambda_i^{\pi}, \gamma_{ij}^{\pi}, \Phi_{c_{ij}}^{\pi}, \Phi_t^{\delta})}^{\text{Imperative groundings}}. \quad (9)$$

The joint distribution factors hierarchically over the set of linguistic constituents $\lambda_i \in \Lambda_t^{\pi}$ as determined by a syntactic parser. The set of linguistic constituents are arranged in a topographical order implied by the syntactic relations in the utterance. The structure informs the factorization of the joint distribution over individual factors where grounding for a linguistic constituent is conditioned on the estimated correspondences $\Phi_{c_{ij}}^{\pi}$ for "child" constituents that appear earlier in the topological ordering. Importantly, the conditioning on true declarative correspondences Φ_t^{δ} from earlier constituents couples the estimation of imperative and declarative groundings for an input instruction. This probabilistic linkage allows disambiguation of action objectives based on stated declarative knowledge in the instruction. For example, given the utterance "pick up the block that the human put down" the observed actions of the human disambiguate which object should be manipulated.

Factors in Equation 9 are expressed as log-linear models with feature functions that exploit lexical cues, spatial characteristics and the context of child groundings. Inference is posed as a search over binary correspondences and is executed by beam search. The ordered sequence of inferred groundings serve as an input to a planner that generates a robot-specific motion plan μ_t to satisfy the inferred objective Φ_t^{π}.

State Propagation

The state variable K_t maintains a belief over factual groundings; see Equation 3. It expresses the degree to which a factual attribute is true of an object. The support for factual grounding variables ranges over workspace entities. For example, the grounding for an utterance such as "the block on the table is mine" informs the degree to which the fact, the possessive Mine, is true for the entities which "block" is grounded to. The grounding likelihood can be viewed as an observation from language, informative of the latent belief over the factual grounding contained in the knowledge state, and is used to update the propagated state. The factor $P(K_t|\Phi_t^{\pi}, K_{t-1})$ models the updated state variable linked with the grounding obtained for the current utterance with the previous state estimate forming the prior. Since factual groundings are assumed to be uncorrelated[2], the posterior distribution over each grounding variable in the state is updated independently using a Bayes filter initialized with a uniform prior. This permits the belief over stored facts to evolve over time providing resilience to errors and ambiguities as well as accounting for changes in the environment.

3.4 Model Training

The imperative and declarative factors that constitute the model are trained[3] through a data-driven process while the parsing factor used an existing rule-based model.

[2] In general one would want knowledge to be structured and to include an inference mechanism to deduce consequences and ensure consistency. This remains part of future work.

[3] Each factor is trained independently using labeled ground truth data. An EM-style approach for jointly training all factors remains future work.

The imperative grounding factor, realized using the *Distributed Correspondence Graph* (DCG) model, was trained using an aligned corpus of language instructions paired with scenes where the robot performs a manipulation task. A data set consisting of 51 language instructions paired with randomized world configurations generating a total of 4160 examples of individual constituent-grounding factors. Ground truth was assigned by hand. A total of 1860 features were used for training. Parameters were trained using a quasi-Newton optimization procedure. The declarative grounding factor, realized using the *Sentence Tracker*, was trained using captioned videos without any annotation about what the captions or the words that comprise those actions referred to in the video. An EM-like algorithm acquired the parameters of the declarative grounding factor using a corpus of 15 short videos, 4 seconds long, of agents performing actions in the workspace. The parsing factor was realized using *START*, a natural language processing system which is primarily used for question answering from semi-structured sources such as the World Factbook and Wikipedia. START was unchanged and we used its standard APIs for language parsing and generation and for executing actions in response to user requests.

3.5 Complexity Analysis

Incremental estimation in temporal grounding graphs relies on propagating a state K_t while retaining visual observations $Z_{0:t}$ trading off the runtime of grounding a single utterance with the space and time complexity of estimating perceptual groundings from visual observations. Table 2 provides a complexity analysis for the proposed approach compared to two common alternatives: the first column of the table presents the analysis corresponding to a model that relies entirely on the observation history without any state maintenance, i.e., estimates $P(\Phi_t|\Lambda_{0:t}, Z_{0:t}, \Gamma_t)$. The last column corresponds to a model that maintains a *full* symbolic state without retaining either the visual or linguistic observations. We introduce the following notation for the analysis: $\mathbb{\Lambda}$ and $\mathbb{Z}$ are the worst case longest sentence and video requiring the most number of detections, C_w is the declarative symbol with the largest state space, C_Λ is the number of participants in the event described by the worst case sentence, $\mathbb{\Gamma}^{\mathcal{F}}$ and $\mathbb{\Gamma}^{\mathcal{P}}$ are the number of the factual and perceptual grounding predicates, and $\mathbb{o}$ is the number of possible object instances.

Not keeping any state results in low space complexity $O(\mathbb{\Lambda}t + \mathbb{Z}t)$, while having high grounding time complexity $O(t{C_w}^{t\mathbb{\Lambda}}\mathbb{Z}^{C_\Lambda})$. Note the exponential dependency on t, the number of time steps. Reasoning about any new sentence requires re-reasoning about all previously seen sentences and any correlations between those sentences resulting in an exponentially increasing joint distribution. Even for short exchanges this runtime is infeasible. Conversely, keeping all perceptual state in a symbolic manner results in extremely efficient inference of groundings, $O({C_w}^{\mathbb{\Lambda}}\mathbb{Z}^{C_\Lambda})$. The likelihoods of any declarative groundings are already recorded and only groundings which are relevant to the given command must be updated from time $t-1$ to t. Yet this process must record all possible inferences for any grounding in any previous observation so they can be available for grounding when the stimulus is discarded. Doing so is prohibitively expensive, $O(\mathbb{\Gamma}^{\mathcal{P}}\mathbb{o}^{C_\Lambda C_w}t + \mathbb{\Gamma}^{\mathcal{F}}\mathbb{o}t)$, due to the arity of groundings; even a binary grounding requires storing a fact about every pair of possible objects.

We strike a middle ground between these alternatives by taking advantage of two facts. First, any one sentence requires a small number of declarative groundings which can be estimated using prior visual observations in time linear in the size of those observations. Second, storing factual groundings removes the need to re-run inference over prior utterances dramatically speeding up grounding time to $O(t{C_w}^{\mathbb{\Lambda}}\mathbb{Z}^{C_\Lambda})$. Compared to not having any stored state, we see significant speedup due to the the lack of t as an exponent while adding only a small amount of storage, $O(\mathbb{\Gamma}^{\mathcal{F}}\mathbb{o}t)$ which increases by at most a small constant factor with each new sentence, Note that the propagated state contains only factual groundings. We revert to using the visual observations to save on exponential increase in storage complexity and record facts to lower the exponent of inferring groundings.

4 Evaluation

The system was deployed on the Baxter Research Robot operating on a tabletop workspace. The robot observed the workspace using images captured using a cross-calibrated Kinect version 2 RGB-D sensor at ~20Hz with 1080x760 resolution. Objects were localised using a multi-scale sub-window search in image space with filtering using depth information. A binary SVM with colour histogram features was used for object recognition. The robot engaged in several interactions with human agents that were speaking or typing natural language sentences providing information, narrating the events, or requesting the execution of a command. Spoken commands from the human operator were converted to text using an Amazon Echo Dot. We demonstrate the space of capabilities of the model through a qualitative evaluation and create a corpus of interactions to demonstrate its robustness.[4] The model was evaluated qualitatively and quantitatively, both of which we discuss next.

4.1 Qualitative Results

Figure 4 presents representative examples of grounding a sequence of instructions from the human operator using the proposed model. A robot performs five tasks requiring a combination of state keeping, disambiguating partial information, and observations of its environment. At times, such as in example (d), all such capabilities are needed as part of a joint inference process in order to arrive at the correct grounding. As the scenario unfolds the robot becomes more certain about its groundings, eventually being able to perform its assigned task.

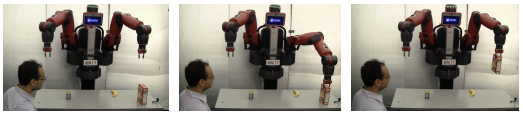
4.2 Quantitative Evaluation

To evaluate our approach quantitatively we collected a video corpus of humans performing actions while providing declarative facts and commands for the robot to execute. Our corpus consists of longer videos composed by combining 96 short, 3 second long, videos consisting of a person performing one action out of 5 (*pick up*, *put down*, *slide*, *move toward*, *move away from*) with one of eight objects from the YCB data set [Calli *et al.*, 2015] (3 fruit, 3 cups, and 2 boxes) where each object had one of three colors (red, green, yellow), and was either small or large, and could be on the table or on top of

[4] Video demonstrations and the corpus used for quantitative evaluation are available at: `http://toyota.csail.mit.edu/node/28`

Table 2: O-bounds on the asymptotic complexity of different approaches to state keeping with Temporal Grounding Graphs.

	No state keeping	Proposed approached	Full state keeping
Observations at each inference step	$\{\Lambda_{0:t}, Z_{0:t}\}$	$Z_{0:t}$	$\varnothing$
State maintained after each inference	$\varnothing$	$\Gamma^{\mathcal{F}}$	$\Gamma^{\mathcal{P}} \cup \Gamma^{\mathcal{F}}$
Space complexity	$\mathbb{\Lambda}t + \mathbb{Z}t$	$\mathbb{\Gamma}^{\mathcal{F}} \mathbb{o} t$	$\mathbb{\Gamma}^{\mathcal{P}} \mathbb{o}^{C_\Lambda C_w} t + \mathbb{\Gamma}^{\mathcal{F}} \mathbb{o} t$
Grounding time complexity	$t C_w{}^t \mathbb{\Lambda} \mathbb{Z}^{C_\Lambda}$	$t C_w{}^{\mathbb{\Lambda}} \mathbb{Z}^{C_\Lambda}$	$C_w{}^{\mathbb{\Lambda}} \mathbb{Z}^{C_\Lambda}$



(a) "The cracker box on the table is my snack." "Pick up my snack."
Inference disambiguates the object for the pick action by relying on the updated fact that the box is the agent's snack.

(b) "Lift the box that I put down."
The object to be lifted is disambiguated in the context of a video depicting an action.

(c) "The box and the can are my snack." "Pack up my snack."
The inferred grounding is an abstract aggregation (*snack*) composed of two tracks corresponding to the can and the box resulting in a multi-action task.

(d) "The box I will put down is my snack." "Pick it up."
A combination of syntactic and visual features are used to resolve the co-reference.

(e) "The fruit on the table is mine." "The green fruit is mine." "Point at my fruit."
Partial information from ambiguous statements is fused together to select an object.

Figure 4: Examples of grounding instructions given visual observations and factual information gathered through one or more natural-language interactions.

another object. The 96 short videos were collected by filming users, not the authors, interacting with objects on a tabletop. They were given generic directions such as "slide a cup on the table". The cues ensured all possible pairings of actions, objects and spatial layouts. These shorter videos were stitched together to form the final video corpus consisting of between one and three seed videos concatenated together, an optional declarative sentence associated with that smaller component video, and a final command for the robot to perform an action with one of the objects. Out of the possible 96^3 videos we created 255 video-sentence pairs. Of these videos, 180 depicted either one or two actions performed by a human followed by a command whose interpretation refers to one or both actions. For example, "pick up the red object the person put down" associated with a clip where one object was picked up and another was put down making the correct interpretation of this command depend on the actions observed. The remainder of the corpus, 75 videos, depicts either two or three actions performed by a human with optional associated declarative facts that may refer to those actions along with a final command. For example, a sequence of videos is paired with the sentences "the green object is the oldest", "the fruit on the table is the oldest", followed by "point at the oldest object". Sentential annotations for the quantitative evaluation were provided by the authors but in the future we intend to test on more diverse user-generated utterances.

A human judge watched these video sequences captioned with the sentences they were paired with. That judge annotated the expected action the robot should perform and the target object with which it should be performed. An inferred robotic command was only considered correct when it performed the correct action on the correct objects in the intended location. These annotations were compared with those generated automatically resulting in an accuracy of 92.5% demonstrating the effectiveness of the state propagation in the model. Chance performance is $\frac{1}{27}$, corresponding to a $\frac{1}{9}$ chance of choosing the correct object and a $\frac{1}{3}$ chance of performing the correct action with that object. Performance on short videos was worse, with 90.2% of inferences being correct, than on long videos, where 94.7% of inferred actions and target objects were correct. While longer videos provide more opportunities for failure they also provide more context for inference. Failures occurred largely due to errors in perception. Actions which moved objects toward or away from the camera were difficult to perceive. Several objects were occluded while an action was performed which occasionally led to an incorrect interpretation of that action.

5 Conclusion

The model presented here significantly extends the space of commands that robots can understand. It incorporates factual knowledge from prior linguistic interactions and visual observations of a workspace including the actions of other agents in that workspace into a coherent approach that performs joint inference to understand a sentence providing a command or new factual knowledge. Knowledge accrues and is refined over time through further linguistic interactions and observations. We intend to extend this model to ground to sequences of actions and collections of objects, to engage in dialog while executing multi-step actions, to keep track of and infer the locations of partially observed objects, and to serve as a basis for a grounded and embodied model of language acquisition.

Acknowledgements

We acknowledge funding support in part by the Toyota Research Institute Award Number LP-C000765-SR; the Robotics Collaborative Technology Alliance (RCTA) of the US Army; the Center for Brains, Minds, and Machines (CBMM) funded by NSF STC award CCF-1231216; and AFRL contract No. FA8750-15-C-0010. We thank Naomi Schurr and Yen-Ling Kuo for assistance during system evaluation.

References

[Andreas and Klein, 2015] Jacob Andreas and Dan Klein. Alignment-based compositional semantics for instruction following. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[Artzi and Zettlemoyer, 2013] Yoav Artzi and Luke Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62, 2013.

[Barrett *et al.*, 2016] D. P. Barrett, A. Barbu, N. Siddharth, and J. M. Siskind. Saying what you're looking for: Linguistics meets video search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2069–2081, Oct 2016.

[Berant *et al.*, 2013] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on Freebase from question-answer pairs. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[Berzak *et al.*, 2016] Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. Do you see what I mean? Visual resolution of linguistic ambiguities. *arXiv preprint arXiv:1603.08079*, 2016.

[Calli *et al.*, 2015] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The YCB object and model set: Towards common benchmarks for manipulation research. In *International Conference on Advanced Robotics (ICAR)*, pages 510–517. IEEE, 2015.

[Cantrell *et al.*, 2010] Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and Xuan Wu. Robust spoken instruction understanding for HRI. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 275–282. IEEE Press, 2010.

[Chen and Mooney, 2011] David L Chen and Raymond J Mooney. Learning to interpret natural language navigation instructions from observations. In *AAAI*, volume 2, pages 1–2, 2011.

[Chen *et al.*, 2010] David L Chen, Joohyun Kim, and Raymond J Mooney. Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research*, pages 397–435, 2010.

[Chung *et al.*, 2015] Istvan Chung, Oron Propp, Matthew R Walter, and Thomas M Howard. On the performance of hierarchical distributed correspondence graphs for efficient symbol grounding of robot instructions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5247–5252. IEEE, 2015.

[Ghahramani *et al.*, 1997] Zoubin Ghahramani, Michael I Jordan, and Padhraic Smyth. Factorial hidden Markov models. *Machine learning*, 29(2-3):245–273, 1997.

[Guadarrama *et al.*, 2013] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Gouhring, Yangqing Jia, David Klein, Pieter Abbeel, and Trevor Darrell. Grounding spatial relations for human-robot interaction. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 1640–1647. IEEE, 2013.

[Hemachandra *et al.*, 2015] Sachithra Hemachandra, Felix Duvallet, Thomas M. Howard, Nicholas Roy, Anthony Stentz, and Matthew R Walter. Learning models for following natural language directions in unknown environments. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 5608–5615. IEEE, 2015.

[Howard *et al.*, 2014] Thomas M Howard, Istvan Chung, Oron Propp, Matthew R Walter, and Nicholas Roy. Efficient natural language interfaces for assistive robots. In *IEEE/RSJ Intl Conf. on Intelligent Robots and Systems (IROS) Work. on Rehabilitation and Assistive Robotics*, 2014.

[Katz, 1988] Boris Katz. Using English for indexing and retrieving. In *Recherche d'Information Assistée par Ordinateur (RIAO)*, pages 314–332, 1988.

[Kim and Mooney, 2012] Joohyun Kim and Raymond J Mooney. Unsupervised PCFG induction for grounded language learning with highly ambiguous supervision. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 433–444. Association for Computational Linguistics, 2012.

[Liu *et al.*, 2016] C. Liu, S. Yang1, S. Saba-Sadiya1, N. Shukla, Y. He, Song-Chun Zhu, , and J. Y. Chai. Jointly learning grounded task structures from language instruction and visual demonstration. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[Matuszek *et al.*, 2014] Cynthia Matuszek, Liefeng Bo, Luke Zetllemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Québec City, Quebec, Canada, March 2014.

[Misra *et al.*, 2016] Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300, 2016.

[Paul *et al.*, 2016] Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M. Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *Proceedings of Robotics: Science and Systems*, Ann Arbor, Michigan, June 2016.

[Russell and Norvig, 1995] Stuart Russell and Peter Norvig. Artificial intelligence: A modern approach. *Prentice-Hall, Egnlewood Cliffs*, 25:27, 1995.

[Tellex *et al.*, 2011] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2011.

[Walter *et al.*, 2014] Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. A framework for learning semantic maps from grounded natural language descriptions. *The International Journal of Robotics Research (IJRR)*, 33(9):1167–1190, August 2014.

[Yu *et al.*, 2015] Haonan Yu, N Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. A compositional framework for grounding language inference, generation, and acquisition in video. *J. Artif. Intell. Res. (JAIR)*, 52:601–713, 2015.

[Zettlemoyer and Collins, 2007] Luke S Zettlemoyer and Michael Collins. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of Empirical Methods in Natural Language Processing and Computational Nautal Language Learning (EMNLP-CoNLL)*, pages 678–687, Prague, 2007.