Annual Review of Control, Robotics, and Autonomous Systems

Robots That Use Language

Stefanie Tellex[1], Nakul Gopalan[2], Hadas Kress-Gazit[3], and Cynthia Matuszek[4]

[1]Department of Computer Science, Brown University, Providence, Rhode Island 02906, USA; email: stefie10@cs.brown.edu

[2]School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia 30332, USA; email: ngopalan3@gatech.edu

[3]Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, New York 14853, USA; email: hadaskg@cornell.edu

[4]Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore, Maryland 21250, USA; email: cmat@umbc.edu

Annu. Rev. Control Robot. Auton. Syst. 2020. 3:17.1–17.31

The *Annual Review of Control, Robotics, and Autonomous Systems* is online at control.annualreviews.org

https://doi.org/10.1146/annurev-control-101119-071628

Keywords

robots, language, grounding, learning, logic, dialogue

Abstract

This article surveys the use of natural language in robotics from a robotics point of view. To use human language, robots must map words to aspects of the physical world, mediated by the robot's sensors and actuators. This problem differs from other natural language processing domains due to the need to ground the language to noisy percepts and physical actions. Here, we describe central aspects of language use by robots, including understanding natural language requests, using language to drive learning about the physical world, and engaging in collaborative dialogue with a human partner. We describe common approaches, roughly divided into learning methods, logic-based methods, and methods that focus on questions of human–robot interaction. Finally, we describe several application domains for language-using robots.

1. INTRODUCTION

Grounded language understanding: interpreting a natural language utterance in terms of the physical state of the robot and the environment

Natural language processing (NLP): computational techniques for transforming human languages such as English into machine-usable structures

As robots become more capable, they are moving into environments where they are surrounded by people who are not robotics experts. Such robots are appearing in the home, in nondedicated manufacturing spaces, and in the logistics industry (1, 2), among other places. Since most users will not be experts, it is becoming essential to provide natural, simple ways for people to interact with and control robots. However, traditional keyboard-and-mouse and touch-screen interfaces require training and must be complex in order to enable a person to command complex robotic behavior (3). Higher-level abstractions, such as automata (4), programming abstractions (5), and structured language (6), offer a great degree of power and flexibility but also require a great deal of training to use.

By contrast, people use language every day to direct behavior, ask and answer questions, provid information, and ask for help. Language-based interfaces require minimal user training and allow the expression of a variety of complex tasks. This article reviews the current state of the art in natural language communication with robots, compares different approaches, and discusses the challenges of creating robust language-based human–robot interaction (HRI). The fundamental question for grounded language understanding is, How can words and language structures be grounded in the noisy, perceptual world in which a robot operates (7)?

We distinguish between two dual problems: language understanding, where the robot must interpret and ground the language, usually producing a behavior in response, and language generation, in which the robot produces communicative language, for example, to ask for explanations or answer questions. In the latter problem, the robot may need to reason about information-gathering actions (such as when to ask clarification questions) or incorporate other communication modalities (such as gestures). Systems that address both problems enable robots to engage in collaborative dialogue.

There is a long history of systems that try to understand natural language in physical domains, beginning with Reference 8. Generally, language is most effective as an interface when users are untrained, are under high cognitive load, and have their hands and eyes busy with other tasks. For example, in search-and-rescue tasks, robots might interact with human victims who are untrained and under great stress (9). The context in which language is situated can take many forms; examples include sportscasts of simulated soccer games (10), linguistic descriptions of spatial elements in video clips (11), GUI interactions (12), descriptions of objects in the world (13), spatial relationships (14), and the meaning of instructions (15). Language has also been used with a diverse group of robot platforms, ranging from manipulators to mobile robots to aerial robots. **Figure 1** shows some examples.

Language for robotics is currently an area of significant research interest, as evidenced by the papers covered in this article and the many recent workshops on this subject (e.g., the Grounding Language for Physical Systems workshop at the 2012 Conference on Artificial Intelligence, the Model Learning for Human-Robot Interaction workshop at the 2016 Robotics: Science and Systems conference, the Language Grounding for Robotics workshop at the 2017 Annual Meeting of the Association for Computational Linguistics, the Models and Representations for Natural Human-Robot Communication workshop at the 2018 Robotics: Science and Systems conference, and the Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics at the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies). Other survey papers have reviewed related topics; for example, Fong et al. (16) surveyed socially interactive robots, and Goodrich & Schultz (17) and Thomaz et al. (18) provided broad surveys of HRI, although neither focused on language specifically. This survey is intended for robotics researchers who wish to understand the current state of the art in natural language processing (NLP) as it pertains to robotics.

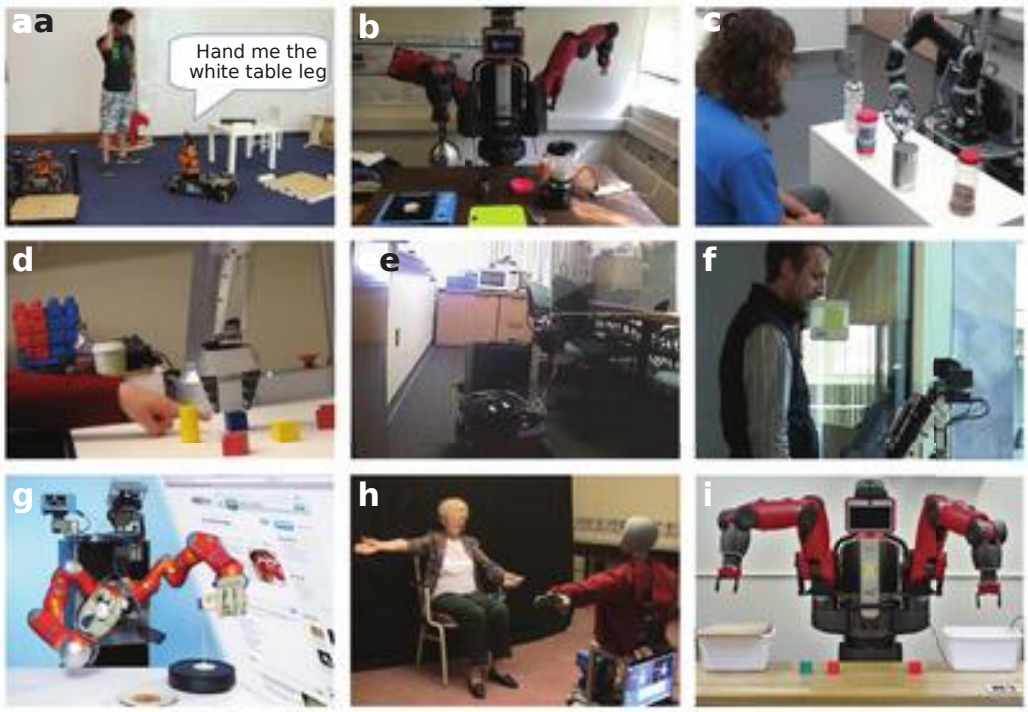

Figure 1

Robots used for language-based interactions. (*a*) Using language to ask for help with a shared task. Panel adapted from Reference 9 (*b*) A Baxter robot learning via dialogue, demonstrations, and performing actions in the world. Panel adapted from Reference 187 wi permission from IJCAI (**https://ijcai.org**). (*c*) A Jaco arm identifying objects from attributes (here "silver, round, and empty"). Panel adapted from Reference 174 with permission from IJCAI (**https://ijcai.org**). (*d*) A Gambit manipulator following multimodal pick-and-place instructions (32). (*e*) A Pioneer AT robot achieving goals specified as "go to the break room and report the location of the blue box." © 2009 IEEE. Reprinted, with permission, from Reference 31. (*f*) A CoBot learning to follow commands such as "take me to the meeting room." © 2013 IEEE. Reprinted, with permission, from Reference 188. (*g*) TUM-Rosie making pancakes by downloading recipes from wikiHow. © 2012 IEEE. Reprinted, with permission, from Reference 63. (*h*) A socially assistive robot helping elderly users in performing physical exercises. © 2012 IEEE. Reprinted, with permission, from Reference 146. (*i*) A Baxter robot performing a sorting task synthesized from natural language (73).

Figure 2 shows a system flow diagram for a language-using robot. First, natural language input is collected via a microphone or text. Words are converted to a semantic representation via language processing; possible representations range from a finite set of actions to an expression in a formal representation language, such as predicate calculus. For example, the words "red block" might be converted to a formal expression such as $\lambda x : \text{block}(x) \wedge \text{red}(x)$. Next, symbols in the semantic representation are connected or grounded to aspects of the physical world. For example, the system might use inference to search for objects in its world model that satisfy the predicates block and red. The results inform decision-making; the robot might perform a physical action (such as retrieval) or a communicative action (such as asking, "This red block?"). Many approaches to language for robotics fit into this framework; they vary in the behaviors they include, the problems they solve, and the underlying mathematics of the modules.

This article is organized as follows. Section 2 gives preliminary material common to all methods. Section 3 covers technical approaches, organized around the method used to achieve language-using robots. Section 4 provides an orthogonal view that organizes the state of the art

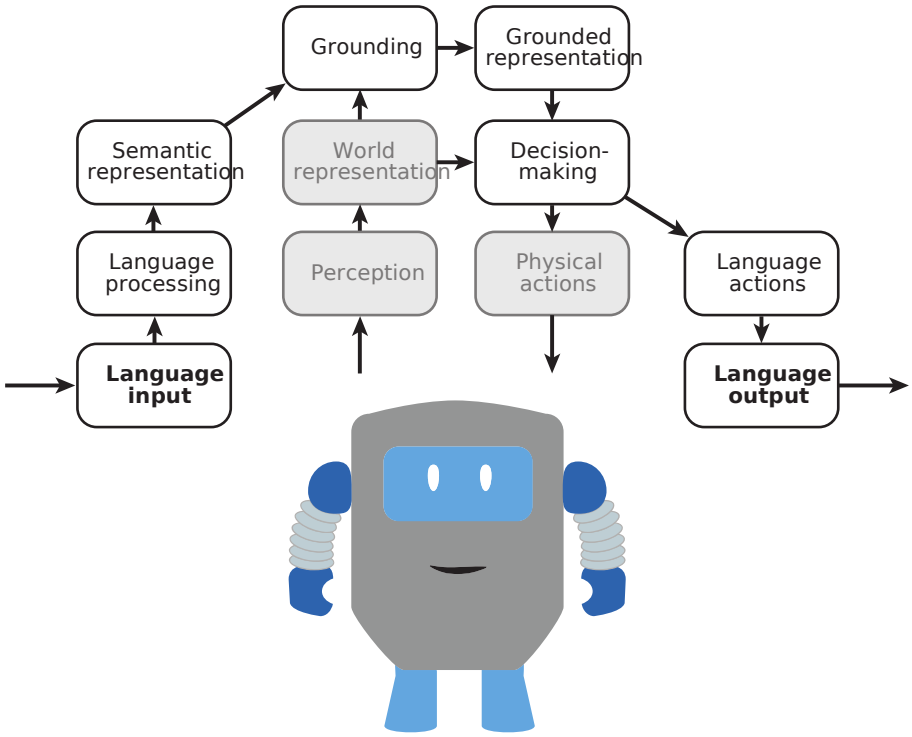

Figure 2

System diagram showing language input and output integrated into a robotic system. Many approaches include only a subset of the modules. Grayed-out modules are relevant to language interpretation but are not reviewed in this article.

around the problem being addressed: human-to-robot communication, robot-to-human communication, and two-way communication. Section 5 concludes with a summary and a discussion of current open questions.

2. PRELIMINARIES

In this section, we define common terminology used in this field and provide technical background needed to understand many of the approaches described in subsequent sections. We review the concept of grounded language, the syntactic and semantic structure of language, and statistical language processing.

2.1. Grounded Language

Grounded language: language that refers to or is interpreted in reference to the physical world

Grounded language (also called situated language or physically situated language) has meaning in the context of the physical world—for example, by describing the environment, physical actions, or relationships between things (7, 19). Possible groundings range from low-level motor commands to perceptual inputs to complex sequences of actions. Grounded language acquisition is the proces of learning these connections between percepts and actions. For example, if a person instructs a robot to pick up a cup, the robot must map the word “cup” to a particular set of percepts in its high-dimensional sensor space—for example, by recognizing that a particular pattern in its camera sensor corresponds to this word. Then, to follow the command, it must produce a plan or policy

Table 1 Examples of natural language and possible groundings

Natural language[a]	Possible sensor/actuator[b]	Category[c]	Grounding/interpretation[d]
"Turn left"	Wheels, legs	Command understanding	Contra-rotate the steering actuators
"Red"	Camera	World sensing	Output label red from color classifier
"This is a laptop"	Camera, RGB-D sensor	Object recognition	Output label laptop from multiclass classifier
"Above you"	Range sensor	Understanding spatial relationships	Location in positive z-space with respect to the robot
"Hand me the orange mug on the left"	Manipulator plus all sensors above	Combined	All of the above

Abbreviation: RGB-D, red, green, and blue plus depth.
[a]Natural language that might occur when instructing or informing a robot.
[b]Possible sensors or actuators providing the physical context.
[c]The underlying task or reasoning problem implicitly encoded in the language.
[d]The physically situated, or grounded, meaning of the language.

that causes its end effector to create a stable grasp of the cup and lift it. Many aspects of this plan are implied by the language but not explicitly stated; for example, if the cup has water in it, the robot should lift it in a way that does not cause the water to spill. This mapping between language and objects, places, paths, and events or action sequences in the world is a key challenge for language and robotics and represents the grounding problem. For robots, language is used primarily as a mechanism for describing objects or desired actions in the physical world; much of the work described in this survey is in the domain of grounded language. A key research question is how to represent this mapping between words and symbols and high-dimensional data streaming in from sensors and high-dimensional outputs that are available from actuators. **Table 1** shows examples of language and possible groundings. Note that in some cases, the grounding is a discrete output from a classifier, while in other cases it is a high-dimensional controller command, such as "contra-rotate the steering actuators." These are examples of possible groundings that have been used in the literature; two key research questions are what the grounding process should look like and how this mapping should be carried out.

2.2. Syntactic Representations and Analysis

Natural language has a hierarchical, compositional syntax (20) that is studied in linguistics and cognitive semantics. This structure enables people to understand novel sentences by combining individual words in new ways (21). This syntactic structure can be used to help extract semantic representations of the words' meaning. A variety of formalisms have been created to express this structure, of which the best known is context-free grammars (CFGs), developed in the 1950s by Noam Chomsky (22). CFGs and their many variants are used to describe the syntactic structure of natural language. Sipser (23) provided a formal definition of CFGs, and **Figure 3** gives an example CFG for a small subset of English along with an associated parse tree. Many variants of CFGs exist. Pretrained parsers are a common tool, many of them (24, 25) trained using the Penn Treebank (26), a corpus of text manually annotated with parse trees. Other parsers are trained on corpora of text, such as newspaper articles. These data are often not a good fit for robotics tasks, which typically contain imperative commands and spatial language, leading to reduced performance on robotics tasks by off-the-shelf tools.

Parse tree: a structure that represents the syntactic decomposition of an utterance in the form of a rooted tree

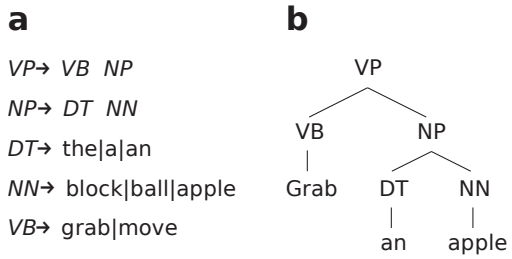

Figure 3

Grammar and parse tree for the English sentence "Grab an apple." (*a*) Context-free grammar for a small subset of English. (*b*) The structure defining compositional relations among word meanings. Abbreviations: DT, determiner; NN, noun; NP, noun phrase; VB, verb; VP, verb phrase.

Many robotics applications use combinatory categorial grammars (CCGs) (27). CCGs are a grammar formalism created to handle linguistic constructions such as coordination (e.g., "John and Mary like apples and oranges") that cannot be expressed by CFGs. CCGs are useful because they model both the syntax and the semantics of language—an approach that is useful for real-world language learning. These learned elements take the form of lexical entries, which combine natural language, syntax, and semantics. Extensive work has been done on automatically creating parsers (28–30), typically learning from pairs of natural language sentences and sentence meaning CCGs have been applied to robotic language understanding in many contexts (29, 31–33), which are reviewed in the following sections.

2.3. Formal Semantic Representations of Language

Semantic representations, which capture the meanings of words and sentences in a formalism that can be operated on by computers, can be extracted with (or from) syntactic structures, such as the example in **Figure 3**. A possible semantic interpretation can be captured by the first-order logic formula $\exists x(\text{apple}(x) \wedge \text{grab}(x))$, which states, "There exists an x that is an apple and that is being grabbed." Given a consistent formal meaning for, e.g., grab(), this expression can be interpreted and used for understanding actions in the world. Extensive work has been done on symbolic representations of semantics (e.g., 20, 34–36). CFG productions can be combined using λ-calculus rules to automatically construct semantic representation from a syntax tree. In this section, we briefly mention the main semantic building blocks that are used by many approaches.

λ-Calculus: a formalism for expressing computation in terms of function arguments and application

First-order predicate logic extends propositional (Boolean) logic with predicates, functions, and quantification. Semantic meaning can be extracted using compositional operators associated with each branch of the syntactic tree. To perform language grounding in the context of robotics, these operators must be grounded in the physical world, i.e., through sensors and actuators; for example, grab() could be grounded to a manipulation action. Additional information regarding the formal syntax and semantics of first-order predicate logic can be found in logic texts, such as the textbook by Huth & Ryan (37).

Temporal logic: logic that includes temporal operators; roughly speaking, the truth value of a formula is evaluated over sequences of states labeled with the truth values of the propositions

Temporal logics are modal logics that contain temporal operators (38), allowing for the representation of time-dependent truths. (For example, the phrase "grab an apple" implies that the apple should be grabbed at some future point in time, an operation referred to as "eventually," written $\Diamond GrabApple$, where *GrabApple* is a Boolean proposition that becomes *True* when the apple is grabbed.) There are different temporal logics that vary in several important dimensions, including whether time is considered to be discrete or continuous, whether time is linear (formulas are defined over single executions) or branching (formulas are defined over trees of executions), and whether the logics are deterministic or include probabilistic operators and reasoning. In a

recent review, Kress-Gazit et al. (39) described the use of several temporal logics in the context of robot control.

2.4. Statistical Natural Language Processing and Deep Learning

Substantial progress in NLP has been made by eschewing the explicit modeling of linguistics structures. For example, *n*-gram models that focus on counting words (40) robustly capture aspects of language use without requiring a full understanding of syntax or meaning, by leveraging the statistics of word co-occurrence. Shallow parsing or chunking is useful to capture aspects of syntax and semantics without performing a complete analysis of the sentence (41). Many approaches rely on less linguistically plausible but more robust structures to achieve learnability and tractability. Modern approaches use word vectors to capture or learn structure, such as long short-term memory units (LSTMs) (42) combined with Word2Vec (43) or Paragraph Vector (44). These approaches learn a vector representation associated with either words or longer documents and then compute over an entire sentence to perform tasks such as language modeling, parsing, or machine translation. Many robotics applications leverage these techniques to learn a statistical or deep model that maps between a human language and one of the formal representations mentioned above.

3. CLASSIFICATIONS BY TECHNICAL APPROACH

In this section, we cluster approaches based on three broad categories: lexically grounded methods (Section 3.1), learning methods (Section 3.2), and HRI-centered approaches (Section 3.3). The first category, lexically grounded methods, focuses on defining word meanings in a symbol system, typically through a manual or knowledge base grounding process, and using logics, grammars, and other linguistic structures to understand and generate sentences. The second category of approaches covers learning word and utterance meanings from large data sets, with inspirations drawn from machine learning and computational linguistics. Finally, HRI-centered approaches focus on the language experience for people interacting with robots. While we use these broad categories to discuss approaches, in practice much of the work in this field belongs to more than one category. The categories are not intended to be mutually exclusive; rather, they provide a possible framework for considering the overall research space.

3.1. Lexically Grounded Methods

This section describes work that uses a priori grounded tokens such as objects and actions, with formal symbolic representations for the underlying semantics. Many of these approaches are based on formal logics; temporal logics are frequently used, as there are algorithms to transform the resulting formulas into behaviors that provide guarantees on performance and correctness (39). These approaches are often less robust to unexpected inputs produced by untrained users and can be difficult to implement at scale due to the manually grounded lexicon. However, they enable grounding rich linguistic phenomena such as anaphora (for example, the “it” in “grab the apple, I want to eat it”) and reasoning about incomplete information.

3.1.1. Grounding tokens. Common to the formal approaches described in this section is the grounding of linguistic tokens, such as nouns and verbs, to perceptual information and robot actions. For example, the token “cup” can be grounded to the output of an object detector, or the action “open door” can be grounded to a motion planner that controls a manipulator. These groundings can be either learned or manually prescribed, but in contrast to learning approaches (Section 3.2), analysis of utterances and groundings is performed using syntactic and formal

semantic structures. Because manually grounding words in a lexicon is a time-consuming process, existing knowledge bases and cognitive architectures are often used to automatically enrich the lexicon using a base set of manual groundings.

Referring expressions: natural language expressions that uniquely denote objects, areas, or people to which the speaker is referring

3.1.1.1 Knowledge bases and ontologies. Many existing knowledge bases provide real-world, common-sense knowledge that can be used to create language-using robots. WordNet (45) provides a lexicon of word meanings in English along with relations to other words in a hierarchy. These relations map symbols to other symbols and can be used to initialize or enrich groundings, especially nouns. VerbNet (46) is a large lexicon of verbs, including frames, argument structures, and parameterized actions. Given a grounding of an action, many verbs can be used in associated natural language utterances (47). Similarly, FrameNet (48) created a data set of verb meanings with parameterized actions. ImageNet (49) is an image database organized using nouns in the WordNet hierarchy. This data set has been used extensively in computer vision and provides information that could enable a robot to detect objects and ground noun phrases. Data sets that are specific to a particular type of grounding task also exist, such as RefCOCO (50) for referring expressions (41).

3.1.1.2 Cognitive architectures. Similar to knowledge bases, cognitive architectures encode relationships between symbols; however, these architectures typically encode complex relations between concepts in cognitive models designed to support reasoning mechanisms that enable completion of inferential tasks. In the context of language and robotics, work has been done with Soar (51; **https://soar.eecs.umich.edu**), ACT-R (Adaptive Character of Thought–Rational) (52), and DIARC (Distributed Integrated Affect, Reflection, and Cognition) (53, 54), among others.

Soar (51; **https://soar.eecs.umich.edu**) is a theoretical framework and software tool designed to model human cognition. It includes knowledge, hierarchical reasoning, planning, execution, and learning, with the intent of creating general-purpose intelligent agents able to accomplish many different tasks. Researchers have proposed NL-Soar (55), a system that enables language understanding and generation that is interleaved with task execution. From the language side, tree based syntactic models, semantic models, and discourse models are constructed that enable the system to create a dialogue with a person. Building on this work, Huffman & Laird (56) introduced Instructo-Soar, enabling new instructions to be grounded to procedures in Soar. Instructo-Soar assumes simple imperative sentences that are straightforward to parse and instantiate as a new operator template. Language groundings can also be learned from mixed-initiative HRIs that include language, gestures, and perceptual information (57). The language to be grounded is first syntactically parsed based on a given grammar and dictionary, and then the noun phrases are mapped to objects in the perceptual field, the verbs to actions in the Soar database, and spatial relations to a set of known primitives.

ACT-R and ACT-R/E (Adaptive Character of Thought–Rational/Embodied), introduced by Trafton et al. (52), are frameworks in which cognition is implemented in an embodied agent that must move in space. ACT-R/E has as a goal the ability to model and understand human cognition in order to reproduce and imitate human cognitive capabilities. It has some language capabilities in order to accept commands such as “go hide” to play hide-and-seek.

The DIARC architecture (53, 54), which has been under development for more than 15 years, adopts a distributed architecture that does not attempt to model human cognition. Instead, different instantiations that correspond to different cognitive abilities with varying levels of complexity can be created, determined by the intended use. Using the DIARC architecture (54, 58–60), researchers created a system that incrementally processes natural language utterances, creates goal

for a planner and executes the instructions, as shown in **Figure 1e** (31). In that work, the lexicon is labeled with both syntactic annotations from a CCG (27, 29) and semantic annotations in the form of λ-expressions related to the temporal logic CTL* (38) and first-order dynamic logic. When an utterance is provided, it is incrementally parsed—i.e., a parse is available after every token, the parse is updated as new tokens are received, and the semantics are incrementally produced. Later work employed pragmatic inference to enable more complex language interaction where the meaning of the utterances may be implicit and where context and semantics are combined (61, 62).

PRAC (Probabilistic Action Cores) (63), while not a cognitive architecture per se, generalizes the notion of a knowledge base by creating a system that enables inferring over, disambiguating, and completing vague or underspecified natural language instructions by using information from existing lexical databases and drawing on background knowledge from WordNet and FrameNet, among other sources. From this information, the robot can infer a motor action that causes a source object to end up in a goal location. **Figure 1g** shows an image from this work.

All of these architectures rely on hand-coded atomic knowledge that a human designer imparts to the robot, plus composition operators that enable the creation of more complex knowledge. These frameworks are carefully designed based on theories of cognition, leading to rich, evocative demonstrations. However, it is difficult for these systems to scale to large data sets of language or situations produced by untrained users. This sort of scaling and robustness is a key future challenge.

3.1.2. Formal reasoning. In addition to grounding tokens such as objects and places into detectors, approaches that utilize formal reasoning typically attach semantic structures to lexical items, such as verbs, and to the production rules of the grammar. These semantic structures are used to understand the semantics of utterances and define new lexical items, such as objects and actions. The semantics are typically fed into either a dialogue manager or a planner that executes situated robot actions. Broadly speaking, the following approaches to language interactions follow a similar pipeline: Natural language utterances in the form of text are syntactically parsed and then semantically resolved (and, in some work, pragmatically analyzed) to produce formal representations of the language's meaning.

Early examples of end-to-end systems that used formal representations for natural language interactions were GRACE (Graduate Robot Attending Conference) and GEORGE (Graduate Robot Attending Conference), robots that competed in the Association for the Advancement of Artificial Intelligence (AAAI) robot challenges. At the 2004 AAAI National Conference on Artificial Intelligence, GRACE acted as an information kiosk, providing information about the conference and giving directions, while GEORGE physically escorted people to their destinations (64). Both robots utilized the Nautilus parser (65), which uses a CFG to produce an intermediate syntactic representation that can be pattern matched to a semantic structure available to the interpreter. Building on the Nautilus parser and the GRACE system, the MARCO agent (66) was created to interpret route instructions given in natural language, combining syntactic and semantic structures with information from the perception system regarding the environment.

The process of grounding and executing natural language instructions from websites such as wikiHow was explored by Tenorth et al. (67). The system uses the Stanford parser (68), which uses a probabilistic CFG to syntactically parse instructions. These instructions are grounded using WordNet (45) and CYC (69) and are captured as a set of instructions in a knowledge base. Later work (70) discussed controlled natural language as a way to repair missing information through explicit clarification. Nyga et al. (71) used a similar probabilistic model to utilize relational knowledge to fill in gaps for aspects of the language missing from the workspace.

Raman et al. (72) and Lignos et al. (47) grounded high-level natural language tasks to linear temporal logic (LTL) (38) formulas by using part-of-speech tagging and parsing to create the syntactic structure. VerbNet (46) is then used to find the sense of the verb and assign a set of LTL formulas as the semantics. In that work, the mapping of verb senses to LTL is done manually; in other work (73, 74), semantic mappings are learned using the distributed correspondence graph framework (75); **Figure 1*i*** shows an image from this work.

Siskind (76) presented another framework for formally reasoning about time and state changes with manually defined verb meanings. The approach allowed a robot to identify objects and generate actions by defining a formal framework for objects and contact. The work was based on force dynamics and event logic, a set of logical operators about time.

3.2. Learning Methods

This section covers work on learning models of language meanings from large data sets. The task is to learn a mapping between natural language and symbols in a formal language. In some approaches, the symbols are given. In others, symbols are created as these groundings are learned these methods are robust to a wide variety of language produced by untrained users but offer few guarantees on performance and correctness.

3.2.1. Data and domains for learning methods.

Learning-based approaches use a wide variety of data sets, tasks, and formats for training. Data sets typically consist of natural language paired with some form of sensor-based context information about the physical environment. An annotated symbolic representation is often also provided. The form of sensor data varies; raw perceptual input, such as joint angles, is often too low level, but higher-level representations depend on the specific approach. **Table 2** lists some of the common data sets currently used in language grounding and robotics along with the type of sensor, language, and annotation data.

We accompany **Table 2** with a brief example of applying a data set for a robotic task. The MARCO data set (66) of navigation instructions is the most widely used of the existing data sets (14, 29, 66, 77, 78). Beyond being one of the earliest available data sets in this space, its wide upta is partly because it contains not only route directions but a complete simulation environment in which to navigate. Thus, potential users of the data set do not need to provide their own robot or handle potentially different sensing or actuation capabilities. Instead, language-learning approaches can be directly compared with previous approaches to the same problem by using the natural language instructions in MARCO, then testing in the same simulated environment.

For example, 10 years after the original work used a handcrafted grammar to explicitly model language (66), Mei et al. (79) used a long short-term memory recurrent neural network (LSTM-RNN) to learn to follow directions. This work estimated action sequences from natural language directions, performing end-to-end learning directly from raw data consisting of tuples of natural language instructions, perceived world state, and actions. The LSTM-RNN encodes the navigational instruction sequence and decodes to action sequences, incorporating the observed context (world state) as an extra connection in the decoder step.

The challenge in using any of these data sets is the mismatch between the data provided and th actual data that will be encountered in a real robotic task. The robot in a task may have different sensors, actuators, and representations than the one used in the task. For example, the MARCO data set uses butterflies as a landmark object; most real environments do not have these butterflie but do have other landmarks that may not appear in MARCO. Learning more general concepts such as “landmarks” is an important open question for future work.

A key question for data-based methods is determining a space of possible meanings for words: Into what domain might language be grounded? Domains may consist of specific objects or areas in

Table 2 Data sets used in language grounding and robotics

Data set	Type of data	URL
MARCO (66)	Navigation instructions given to a robot to navigate a map, and the route followed	**http://www.cs.utexas.edu/users/ml/clamp/navigation**
Scene (33)	Images and descriptions of objects in the image	**http://rtw.ml.cmu.edu/tacl2013_lsp**
Cornell NLVR (189)	Pairs of images and logical statements about them that are true or false	**http://lic.nlp.cornell.edu/nlvr**
CLEVR (190)	Images and pairs of questions and answers	**http://cs.stanford.edu/people/jcjohns/clevr**
EQA (104)	Pairs of questions and answers in simulated 3-D environments (the agent needs to search the environment to find the answer)	**http://embodiedqa.org**
IQA (191)	Pairs of questions and answers in different simulated 3-D environments	**http://github.com/danielgordon10/thor-iqa-cvpr-2018**
R2R navigation (192)	Panoramic views in real buildings paired with instructions to be followed	**http://bringmeaspoon.org**
H2R Laboratory language grounding (91, 102)	Predicate-based subgoal conditions paired with natural language instructions	**http://github.com/h2r/language_datasets**
CIFF (106, 193)	Data for three separate navigation domains in 3-D environments, containing instructions paired with trajectories	**http://github.com/clic-lab/ciff**
SLU (14, 83)	Pairs of language command and trajectories for navigation and mobile manipulation	**http://people.csail.mit.edu/stefie10/slu**

Abbreviations: CIFF, Cornell Instruction Following Framework; CLEVR, Compositional Language and Elementary Visual Reasoning; EQA, Embodied Question Answering; H2R, Humans to Robots; IQA, Interactive Question Answering; NLVR, Natural Language for Visual Reasoning; R2R, Room-to-Room; SLU, Spatial Language Understanding.

the environment, perceptual characteristics, robot actions, or combinations thereof. The meaning of language is often grounded into predefined formalisms, which maps well to existing work in formal semantics (20). However, in work more oriented toward more machine learning, there is a trend toward systems that learn the representation space itself from data, leading to systems that do not need a designer to prespecify a fully populated set of symbols and allowing robots to adapt to unexpected input. For example, Matuszek et al. (13) and Pillai & Matuszek (80) showed that symbols for shape, color, and object type can be learned from perceptual data, enabling the robot to create new symbols based on its perceptual experience, while Richards & Matuszek (81) extended that work to creating symbols that are not category limited.

We divide the following approaches into those that use primarily predefined languages (Section 3.2.2), those that are more concerned with discovering the domain (Section 3.2.3), and recent work on using deep neural networks for language understanding (Section 3.2.4). In practice, work in this area falls along a spectrum, ranging from formal-methods approaches that use completely manually defined word meanings (66), to learning mappings between words and a prespecified formal language (10, 73, 82), to learning new symbols from data while specifying perceptually motivated features (83), to learning new features from data as well as a mapping between word meanings and those features (13).

3.2.2. Learning to map to predefined symbolic spaces.

Mapping to predefined symbolic structures has a natural analog in machine translation research. In machine translation, the goal is to translate a sentence from one language to another language (for example, “pick up the block” in

English to "podnieść blok" in Polish). Many approaches take as input a parallel corpus of sentences in the two natural languages and then learn a mapping between the languages. When applied to robotics, the input language is a natural language, and the output is a formal representation language that the robot can act on. The challenge is then to specify an appropriate formal robotic language and acquire a data set or parallel corpus with which to train the model.

This approach has been applied to a variety of domains, such as enabling a robot to learn to interpret natural language directions from pairs of directions and programs that follow those directions (10, 66, 77). The same approach can be used for the inverse problem of generating natural language descriptions of formally represented events, such as RoboCup soccer games (84). MacGlashan et al. (85) showed that a robot can learn to map to a predefined space of symbolic reward functions using the classic IBM Model 2 machine translation approach (86); once the reward function has been inferred, the robot finds a plan that maximizes the reward, even in environments with unexpected obstacles. Misra et al. (87) learned to map between words and a predefined symbolic planning space using a graphical modeling approach, interpreting commands such as "turn off the stove."

Other approaches use semantic parsing to automatically extract a formal representation of word meanings in some formal robot language. These systems vary in terms of the formal language used. For example, Matuszek et al. (82) created a system that learns to parse natural language directions into Robot Control Language (RCL), a control language for movement. This work could learn programmatic structures in language such as loops (e.g., "drive until you reach the end of the hallway"). Alternatively, Artzi & Zettlemoyer (29) created a system for learning semantic parses for mapping instructions to actions in order to follow natural language route instructions, while Thomason et al. (88) used an approach that learned semantic parse information and grounded word meanings from dialogue interactions with users. Fasola & Matarić (89) used a probabilistic approach to learn mappings between commands and a space of actions of service robots, including models for spatial prepositions. Boteanu et al. (73, 74), Brooks et al. (90), and Arumugam et al. (91) grounded language to objects and specifications expressed in LTL. A key difference in all of these approaches is the formal language chosen to represent the meaning of the human language; in many cases, the formal language can represent only a subset of the meanings possible in natural language.

3.2.3. Learning to map to undefined spaces. We draw a distinction between learning to map between predefined symbol spaces and approaches that extend the space of symbols that natural language may be grounded into. We emphasize that this is a spectrum; all learning approaches rely to a greater or lesser extent on some predefined structure. Less prespecification means the system is more general and can be extended to unexpected tasks and environments but also increases the difficulty of the learning problem. Substantial current effort is focused on learning from very little prespecified data.

The Generalized Grounding Graph (G^3) framework (83) was introduced to interpret natural language commands given to a robotic forklift, as well as to interpret route instructions for a wheelchair (14) and a micro air vehicle (92). It uses a graphical model framework to represent the compositional structure of language, so that the framework can map between words in language and specific groundings (objects, places, paths, and events) in the physical world. It learns feature weights in a prespecified feature space to approximate a function for mapping between words in language and aspects of the world. This work has been extended to enable robots to ask natural language questions that clarify ambiguous commands (78, 93), to enable robots to ask for help (94). It has also been extended to create an efficient interface for interpreting grounded

language by mapping to planning formalisms, an approach that dramatically increases the speed with which words can be interpreted by the robot. Building on this framework, Paul et al. (95) created a system that learns to interpret subsets of objects, such as "the middle block in the row of five blocks."

Other approaches do not require features to be prespecified but do encode a space of possible features as well as data sources from which features are derived. Roy & Pentland (96) created a system for learning nouns and adjectives from video of objects paired with infant-directed speech. It learned to segment audio and map phonemes to perceptual features without a predefined symbol system. Matuszek et al. (13) created a system for learning word meanings for words by automatically creating new features for visual object attributes, while Pillai & Matuszek (80) learned to select negative examples for grounded language learning. Guadarrama et al. (97) created a system for interpreting open-vocabulary object descriptions and mapping them to bounding boxes in images, leveraging large online data sets combined with a model to learn how to use information from each data set. Blukis et al. (98) developed a method that learns to create a semantic map of the environment by projecting between the camera frame and a global reference frame. These approaches represent emerging steps toward an end-to-end learning framework from language to low-level robot control.

3.2.4. Grounding language using deep learning.

Modern deep learning–based approaches of convolutional neural networks, recurrent neural networks, and deep Q-networks led to successes in computer vision, machine translation, and reinforcement learning. Using neural networks or a connectionist architecture is not novel. Older neural network–based approaches (e.g., 99, 100) learned robot behavior from demonstrations and mapped language to these behaviors. Roy & Pentland (96) used recurrent neural networks to learn word boundaries by phoneme detection directly from speech signals. However, the amount of data being used and represented in modern deep learning methods is much larger in scale and allows for end-to-end learning. These novel deep approaches were applied to solve problems of language grounding (e.g., 79, 101). In this article, we do not survey these methods in great detail, but we do provide a short introduction to the types of problems that have been tackled with deep learning–based approaches. We split this discussion based on the problems addressed by these methods.

3.2.4.1. *Instruction following with sequence-to-sequence approaches.*

Some of the earliest progress was made in the area of instruction following (91, 79). This is a supervised problem where an agent performs a sequence of actions in response to a natural language command. In this problem setup, a common theme is to treat a language command and a sequence of actions performed by the agent as a machine translation problem using recurrent neural network–based sequence-to-sequence approaches (79). Others have abstracted the problem to learn the grounding from natural language to subgoals or goals (102, 103). These methods have been implemented in robots only when the abstract fixed grounding symbols have been provided (91).

Some approaches try to reduce the amount of supervision by converting this instruction-following problem into a reinforcement learning problem. This was first done with classical policy gradient methods by Branavan et al. (12); more recently, it has been applied to richer environments with visual inputs (104–106). A common strategy is to model the agent and its environment as a Markov decision process and encode the instruction given to the agent as the state of the environment. Such agents have been able to answer questions about the properties of objects or navigate to objects in simulation. This approach is difficult to implement in a physical given the number of episodes required to learn behaviors.

3.2.4.2 Grounding objects in images. Grounding or captioning objects within images to their names is an active area of research within deep learning. Initially, this work used classifiers to recognize an object class within an image (107). It then progressed to captioning images densely, that is, recognizing all objects within an image (108, 109). A general approach, first described in Karpathy & Fei-Fei (109), is to align vectorized object representations within the image with the vectorized representations of sentences used to describe the objects in the image. These approaches are capable of labeling activities being performed by the objects of interest and also allow retrieval of images described by natural language (108). They have been implemented in physical robots in an object retrieval setting by training the robot on simulated images (110, 111).

3.2.4.3 Grounding control from robot perception. Bhukis et al. (98) developed a system that learned to map between navigation instructions and low-level control actions, mediated by the robot's sensor input and control actions. This work aimed to perform end-to-end learning from language to control actions and has since been demonstrated in physical robots.

3.3. Approaches Centered on Human–Robot Interaction

The final broad category of work we consider is that which lies primarily in the area of HRI. While work in the previous sections is grouped by learning and representation models, here we describe how NLP research supports and is supported by robots that interact directly with people. It is often these approaches that create the most robust behaviors and end-to-end systems, drawing on insights from learning and logic-based methods.

We discuss language-based HRI efforts divided broadly by tasks, considered on a spectrum (see **Figure 4**). On one end, language provides a natural supporting mechanism for robot learning (Section 3.3.1). In this area, language is used as a tool to help robots learn other tasks. On the other end, robots provide an ideal testbed for learning to understand physically situated language; here, the robot is a platform for learning grounded language. This subtopic is substantial and has been covered in Section 3.2. Tied to both areas are efforts whose primary goal is the development of systems that use language in order to support robust HRI (Section 3.3.2).

3.3.1. Language-based interactions to improve robot learning.

Robots that learn have the potential to automatically adapt to their environment and achieve more robust behavior. In this section, we describe how language technology can enable more efficient and effective robot

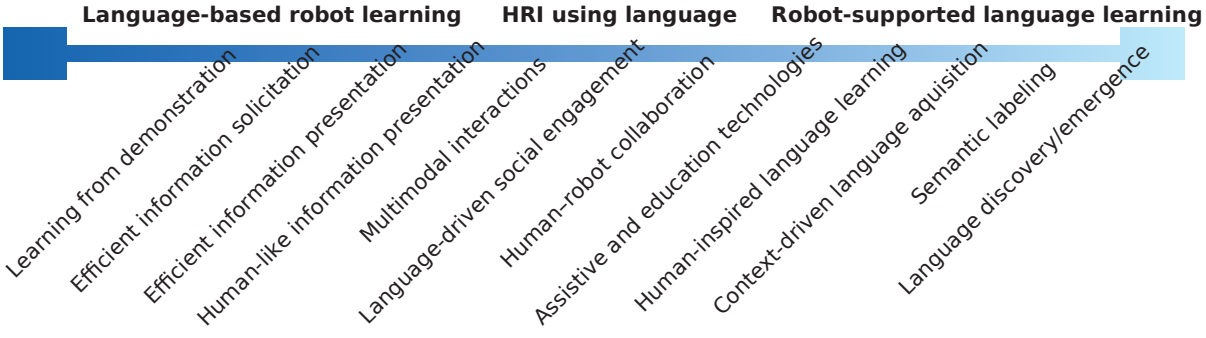

Figure 4

A categorization of work using language for human–robot interaction (HRI). This visualization spans efforts that use language to support efficient robot learning, efforts to use language in order to maximize the effectiveness of HRIs, and the use of robots as physically situated agents to support language learning.

learning, especially from human teachers. Natural language provides a rich, accessible mechanism for teaching robots while still being grounded in the physical world. The vast body of literature on human learning provides questions about learning modalities, information presentation, reward functions, and interaction-based learning. We describe current work on developing robot systems that learn about the world from natural language inputs, including efforts on learning from demonstration (LfD), learning reward functions from language, active learning, and learning how to elicit instructional language.

When learning physical concepts like object characteristics or actions, the physical referent must be linked to linguistic structures. This is seen both explicitly, as in referring expressions (e.g., "this is a yellow block"), and implicitly, as when connections are learned from the coexistence of words and percepts during training. Exploring this connection between linguistic references and their grounded referents is the basis of substantial work on LfD, in which demonstrations connect the learning concepts and the language used.

In LfD, language is used as a learning signal to improve robot learning and capabilities. Steels & Kaplan (112) used language and camera percepts to learn taxonomy-based object instance and their associations with words. Billard et al. (99) used LfD to ground language with a constrained vocabulary to sequences of actions demonstrated by the teacher. Chao et al. (113) used LfD to ground concepts for goal learning, where the concepts are discrete, grounded percepts based in shared sensory elements with human explanations. Concepts are denoted in words to human participants, but language is not part of the learning problem; word meanings are provided to the system by the designer. Krening et al. (114) used object-focused advice provided by people to improve the learning speed of an agent. Language can also be used to describe actions rather than perceived objects, as in programming by demonstration, in which demonstrations of actions are paired with natural language commands (115). Programming by demonstration can also rely on more complex semantic parsing, as in the approach developed by Artzi et al. (116), in which language is interpreted in the context provided by robot state. In all of these papers, humans use language to provide information, advice, or warnings to the robot to improve task performance.

Language can be used to provide explicit feedback to a learning system. The mechanism for learning from that feedback can be treated as a learning problem itself. In this framework, language is learned jointly with policies rather than jointly with direct observations, allowing learning that is less situation specific (85). This approach can allow a nonspecialist to give an agent explicit reward signals (117) or can model implicit feedback strategies inherent in human teaching (118, 119).

Robots asking questions about their environment is a form of active learning in which the learning agent partially or fully selects data points to label. Asking questions that correspond to a person's natural teaching behavior (120) is balanced with selecting data that optimize learning, as queries to a user are a sharply limited resource (121). In general, incorporating active learning makes learning more efficient and makes it possible to learn from fewer data points (122, 123). This form of learning can be implemented in a domain-independent way, as done by Knox et al. (124), and can improve efficiency on learning tasks, including both explicit language grounding (125) and more general robotics problems, such as learning conceptual symbols (126), spatial concepts (127), or task constraints (128).

Another topic in learning from language provided by nonspecialists is how to correctly elicit information and demonstrations from people. Chao & Thomaz (129) explored conducting dialogue correctly, with appropriate multimodal timing, turn-taking, and responsiveness behavior (130). Learning from nonspecialists also means figuring out what questions to ask; Cakmak & Thomaz (131) studied how humans ask questions and designed an approach to asking appropriately targeted questions for LfD, while Pillai & Matuszek (80) demonstrated a method for automatically selecting negative examples in order to train classifiers for positively labeled grounded terms.

3.3.2. Human–robot interaction using language. HRI is one of the most active areas for grounded language research. Language provides a natural mechanism for interacting with physical agents in order to direct their actions, learn about the environment, and improve interactions. At the same time, interacting with people provides a rich source of information and training data that robots can learn from in order to improve their capabilities. Language-based HRI is a broad, active field of study. In this section, we provide an overview of some of the categories of current research on HRI and language.

Childhood education is a significant area of research for HRI studies (132), both because there is a chronic shortage of personnel in education and child care and because increasing the role of technology in childhood education is a practical way of attracting a larger and more diverse population into STEM fields. Research in this area focuses largely on the role of interactive play in child development. This play can take the form of acting out stories between children and robots (133), assisting with language development (134–137), or serving as intelligent tutoring systems (138, 139).

Language in HRI is often paired with other interaction modalities. Modalities such as gesture and gaze direction affect everything from deictic (referential or attention-drawing) interactions to what role a robot may play in a setting (140). There is a growing body of work in which language is incorporated into multimodal HRIs (141). Matuszek et al. (32) used a combination of language and unconstrained, natural human gestures to drive deictic interactions when using language to teach a robot about objects, while Huang et al. (92) used modeling to evaluate robots' use of gesture. In the inverse direction, Pejsa et al. (142) used people's speech, gaze, and gestures to learn a multimodal interaction model, which was then used to generate natural behaviors for a narrating robot.

Another key area of HRI research is work on assistive robotics, in which robots perform tasks designed to support persons with physical or cognitive disabilities. This support can take many forms; with respect to language, social and cognitive support is most common. Socially assistive robot systems have been used to engage elderly users in physical exercise (143, 144), incorporating language pragmatics and anaphor resolution (145, 146) as well as verbal feedback. Verbal robots have also been explored in the context of autism support (147) and tutoring for deaf infants (148).

4. CLASSIFICATIONS BY PROBLEM ADDRESSED

Most of the above approaches can be applied to more than one communication task. Here we review those tasks, divided into three sections: understanding communications from a human to a robot (the largest body of work), generating linguistic communication from a robot to a human, and two-way systems that endeavor to both understand and generate language.

4.1. Human-to-Robot Communication

Human-to-robot communication is the problem of enabling robots to interpret natural language directives given by people. Understanding a person's language requires mapping between words and actions or referents in the physical world. Two specific subproblems include understanding commands and information given to the robot by a person.

4.1.1. Giving robots commands. Command understanding is the problem of mapping between language and physical actions on the part of the robot. One early and widely considered domain is route direction following, where a mobile robot must interpret instructions on how to move through an environment. MacMahon (66) created a large data set of route directions in simulation, which has been used in a number of papers (10, 29). Kollar et al. (14) used a statistical approach to interpret instructions for a robotic wheelchair. Shimizu & Haas (149) used a

conditional random field approach to learn word meanings, and Matuszek et al. (77) used a machine translation approach to learn to follow instructions in real-world environments, including counting and procedural language such as "the third door" or "until the end of the hall." Robotic platforms used for this problem include a robotic wheelchair (14, 66), robotic unmanned aerial vehicles (92), and mobile robots (150). Understanding navigational commands remains a significant and ongoing area of research (151).

A second class of problems is interpreting natural language commands for manipulator robots. This problem has been studied in the subdomains of interpreting textual recipes (152, 153), following instructions for a robotic forklift (83), and interpreting instructions to a tabletop arm (32, 67) and in Baxter robots (73, 74). Such language may refer only to the robot's motion; for example, Correa et al. (154) created a robotic forklift with a multimodal user interface that interpreted shouted commands such as "stop!" However, since manipulators manipulate things in the world at least some of the time, this class of commands is frequently blended with understanding language about objects.

Another frequently studied task is understanding instructions in cooking, particularly focusing on following the semiconstrained language of recipes. Beetz et al. (153) used a reasoning system to interpret recipes and cook pancakes. Tasse & Smith (155) created a data set of recipes mapped to a formal symbolic representation, while Kiddon et al. (156) created an unsupervised hard expectation–maximization approach to automatically map recipes to sequenced action graphs; neither system used robots. Bollini et al. (152) created a system for interpreting recipes but did not ground ingredients into perception. Although the language of recipes is constrained, understanding them remains a challenging problem, in part because ingredients are combined into new things that do not exist at the time of original interpretation—for example, flour, eggs, water, and sugar are transformed into a batter, which is then transformed into a quick bread. Interpreting forward-looking language that maps to objects that do not yet exist is a difficult problem. Similarly, instructions often require the robot to detect certain perceptual properties, as in "cook until the cornbread is brown." Correctly detecting these properties requires advances in perception combined with language to create or select a visual detector to identify when this condition has been met.

4.1.2. Giving robots information about the world.

A second element of language interpretation is enabling robots to use language to improve their knowledge of the world. Compared with instruction following, this topic is a less studied area, but there is nonetheless a rich array of approaches. Cantrell et al. (157) created a system that updates its planning model based on human instructions, while the system of Walter et al. (158) incorporates information from language into a semantic map of the environment. Pronobis & Jensfelt (159) described a multimodal probabilistic framework that incorporates semantic information from a wide variety of modalities, including perceived objects and places as well as human input.

We briefly discuss two specific important subproblems in human-to-robot communication: how robots can resolve references to and understand descriptions of objects, and understanding descriptions involving spatial relationships. One of the major areas in which robots have the potential to help people is in interacting with objects in the environment, meaning it is critical to be able to learn about and understand physical references, both spatial (as in "the door near the elevators") and descriptive (as in "the yellow one between the two toys" or, more abstractly, "a nice view").

4.1.2.1. References to objects.

Robots may need to retrieve, manipulate, avoid, or otherwise be aware of objects being referred to in language. Language about objects and landmarks in the world

can be broken down by level of specificity; we roughly categorize language at these different levels of abstraction as (*a*) general language about object characteristics, such as color, shape, or size (32, 160, 161); (*b*) descriptions of objects at the type, or category membership, level, which encompass approaches that tie language into object recognition (80, 96, 162, 163); and (*c*) language about particular instances of objects, such as "my mug" (14, 62, 83, 112). These categories often overlap. For example, the first step for recognizing an instance is often finding all objects in that category, or object types might be further differentiated by their attributes, as in "the yellow block."

Another issue is interpreting complex descriptions. For example, one route direction corpus contains the instruction "you will see a nice view," referring to a view out of a set of windows the robot would pass. This expression requires the robot to make a subjective judgment about the world. A corpus of object descriptions contains the phrase "a small pyramid like the pharaohs lived in" (32), which requires differentiating direct physical descriptions from background knowledge. In addition, it is not always clear what defines an object. A bottle consists of a bottle and a cap, and a person referencing "the bottle" may mean both, or they may say, "Grab the bottle and then turn the cap," referring to them separately. For assembly tasks, a part such as a screw and a table leg m combine to form a completed assembly, the table (94, 164). Grounding these sorts of expressions is an open problem.

4.1.2.2 Referring-expression resolution. Understanding natural language expressions that denote particular things in the robot's environment is another key subproblem. Referring expressions may occur in commands (e.g., "go through the door near the elevators," in which the robot must identify the referenced door) as well as manipulation instructions (e.g., "pick up the green pepper") (62, 83). Chai et al. (165) created a system that interprets multimodal referring expressions using a graph-based approach. Matuszek et al. (32) and Whitney et al. (166) merged information from language and gesture to interpret multimodal referring expressions in real time using a filtering approach and a joint classification approach, respectively; an image from Matuszek et al. (32) is shown in **Figure 1*d***. Golland et al. (167) generated spatial descriptions using game theory to generate human-interpretable referring expressions in a virtual environment.

4.1.2.3 Spatial relationships. Interpreting spatial relationships is a well-known, complex problem in NLP. For route instructions, the language may take the form of "the door near the elevators" or "past the kitchen"; for object descriptions, it may take the form of "at the top left corner." Understanding these instructions frequently requires not only referring-expression resolution to understand phrases referring to landmarks but also pragmatic disambiguation of possible meanings. Spatial prepositions are frequently used to refer to objects, places, or paths in the physical world. Spatial prepositions are a closed-class part of speech; a typical language has only a few, and new ones are rarely added. Cognitive semantics has focused on the structure of spatial language and how humans use it, especially the argument structure as well as semantic features that allow it to be interpreted (36, 168). Some work has focused specifically on spatial prepositions (11, 127, 169, 170). This problem also arises in the context of referring-expression resolution, since expressions such as "near" or "between" require identifying a place or an object from distractors.

4.2. Robot-to-Human Communication

In the context of natural language user interfaces, people frequently expect spoken responses when they speak to a system such as a robot. Language is an obvious way to engage in active disambiguation, convey information, and provide context. Researchers have studied the problem of enabling a robot to produce natural language to answer questions, ask for help, or provide instructions. This

problem is the inverse of language understanding: The robot desires to communicate something to the person and must find words that convey its ideas. Subproblems include robots instructing people, robots asking questions, and robots informing people.

4.2.1. Robots instructing people and asking questions. Often a robot might use language to try to get a person to do something, typically by asking for help or asking the person to carry out an action. The most basic approach to language generation is template-based or scripted approaches, in which a designer encodes the words the robot will say. For example, Fasola & Matarić (166) used templates to generate language to motivate physical exercise for older adults (as shown in **Figure 1*h***). This approach is straightforward and can result in sophisticated sentences but is limited in its adaptability to novel environments and situations. Other approaches focus on enabling a robot to adaptively generate sentences based on the context. Knepper et al. (164) generated natural language requests for help in assembling Ikea furniture from untrained, distracted users. CoBots navigate an office environment delivering objects and ask for navigation help using a human-centered planner to determine whom to ask for assistance (171).

A second sort of instruction is actively using language to induce a person to provide additional information, for example, by asking a question. Deits et al. (93) presented an algorithm to generate targeted questions based on information theory to reduce confusion. Rosenthal & Veloso (172) modeled humans as information providers, using a partially observable Markov decision process to ask questions when the robot encountered problems. Thomason et al. (173) created a system for opportunistically collecting information from someone about objects in its environment (in which a robot asks about objects near a person, including questions irrelevant to the immediate task) and learning about objects from attributes (174) (as shown in **Figure 1*c***). Pillai et al. (125), Cakmak & Thomaz (131), and others have used active learning to select focused questions that allow the robot to efficiently collect information. All of these approaches use statistical frameworks to generate instructions or queries given the robot's current physical context.

4.2.2. Robots informing people. In addition to trying to instruct people with language, a robot may also need to inform people about aspects of the world. For example, Chen et al. (84) created a system that learns to generate natural language descriptions of RoboCup soccer games by probabilistically mapping between word meanings and game events. Mutlu et al. (175) created a storytelling robot that uses language as well as gaze to engage a human listener. Cascianelli et al. (176) created a system that enables a robot to learn to describe events in a video stream and released a data set for service robotics applications. All of these applications require the robot to communicate with a person about aspects of the environment.

4.2.3. Generating references to objects. For the same reasons that a robot may need to understand references to things in its environment (see Section 4.1.2.1), a robot may need to generate referring expressions about objects, landmarks, or people. Dale & Reiter (177) carried out seminal work on generating referring expressions for definite noun phrases referring to physical objects, such as "the red cup," following Gricean maxims of quantity and quality of the communication (178) and focusing on computational cost. This approach assumes a symbolic representation of context, rather than grounding to perception. Golland et al. (167) generated spatial descriptions using game theory to produce referring expressions in a virtual environment that are interpretable by a human partner. Mitchell et al. (179) generated expressions that refer to visible objects that a robot might observe with its camera. Tellex et al. (94) provided an inverse-semantics algorithm for generating requests for help, including expressions such as "the black leg on the white table" (shown in **Figure 1*a***). Fang et al. (180) created a system for collaborative referring-expression

generation using a graph-based approach that changes the generated language based on human feedback, while Zender et al. (181) created a system for enabling a mobile robot to generate natural language referring expressions for objects in the environment and to resolve expressions, using context to determine how specific or general to make the resolution. From a robotics perspective, these examples represent different contexts in which a physical agent may use language productio to improve its ability to accomplish real-world tasks or goals.

4.3. Two-Way Communication

Two-way communication involves enabling a collaborative interaction between a human and a robot, either asynchronously or in dialogue. Such a robot must both interpret a person's communicative acts and generate communicative actions of its own. Two-way communication requires more than simply combining language understanding and generation; the robot must reason about uncertainty in its own percepts, retain conversational state, react quickly to a person's input, and work toward a communicative collaboration. Partly as a result of these challenges, much work has focused on issues that arise from building robotic systems that engage in dialogue with a user and the associated design questions that arise. A variety of end-to-end robotic systems have been created that use language. These systems typically involve the integration of many software and hardware components in order to create an end-to-end user interaction. The focus is often on multimodal communication, where language constitutes one communication mode in the overall interaction.

For example, Bohus & Horvitz (182) created a computational framework for turn-taking that allows an embodied conversational agent to take and release the conversational floor using gaze, gesture, and speech. Some of these systems communicate by understanding language, performing actions, and seeking help when problems are encountered. Matuszek et al. (32) created a system for learning from unscripted deictic gesture combined with language in order to perform manipulations. Okuno et al. (183) created a robot for giving route directions by integrating language utterances, gestures, and timing. Fasola & Matarić (146) created a socially assistive robot system designed to engage elderly users in physical exercise. Veloso et al. (184) created the CoBots, mobil robots that engage in tasks in an office environment, such as fetching objects. Marge et al. (185) created a heads-up, hands-free approach for controlling a pack-bot as it moved on the ground. Tse & Campbell (186) created a system that incorporates and communicates probabilistic information about the environment. A more direct approach is to learn the spatial semantics of actions directly from language (187) (shown in **Figure 1*b***). The CoBot systems learned to follow commands such as "take me to the meeting room," engaging in dialogue with humans in their environment to improve their abilities (188) (shown in **Figure 1*f***).

While these robots understand language, the robot-to-human side of communication can take a form other than, or in addition to, speech. This multimodality reflects the multimodal nature of interagent communication: Even when talking, humans expect to be able to use gesture, gaze, and body language, as well as utterance timing and even prosody (voice tone and inflection). Language using robots must therefore be aware of these expectations and work to address or mitigate them; failing to do so runs the risk of frustrating users when attempting to communicate.

5. CONCLUSION AND OPEN QUESTIONS

Language-using robots require models that span all areas of robotics, from perception to planning to action. Researchers from diverse communities have contributed to ongoing work in this exciting area, and much remains to be done. In this article, we have reviewed methods for robots that use

language. We covered technical approaches, ranging from formal methods to machine learning to HRI approaches. We discussed problems to solve for robotic language use, including learning from and receiving information from people, asking questions, and giving people instructions. We presented some of the most immediately relevant NLP problems, such as referring-expression resolution. Additionally, we briefly reviewed work in related areas, including linguistics, cognitive science, computational linguistics, vision and language, ontologies and formal representations, and nonverbal communication.

Research in formal methods has pointed toward mechanisms for capturing complex linguistic phenomena such as anaphora resolution, interpreting commands about ongoing action, and abstract objects. However, statistical methods often use simpler representations focused on concrete noun phrases and commands for ease of learning. As more sophisticated formal models mature, statistical methods will enable learning of formal-methods-based representations, combining benefits of robustness with more capable and complex language understanding. At the same time, advances in deep learning have enabled approaches that can learn from less data with end-to-end supervision. We expect that deep learning applied to robotic language use will build on existing work to learn with less and less supervision over time. We see opportunities for sophisticated semantic structures from formal methods combined with learning approaches from deep learning to create a new generation of language using robots capable of robustly interpreting sophisticated commands produced by untrained users.

The power and challenge of language lie in its ability to construct arbitrarily fine-grained and specific sentences that apply to parts of the robot and its environment. As a result, robust language-using robots must integrate language with all parts of a robotic system, a formidable task. As we move toward language-using collaborative robots, we need more robust models for the entire planning and perceptual stack of the robot in order to integrate with natural language requests, questions people might pose, learning from language, and the generation of appropriate language and dialogue by the robot. Similarly, the robot must combine verbal and nonverbal modalities in interactive systems in order to fully understand how people interact and to detect and recover from errors. Although daunting, the scale and complexity of the problems described in this survey are indicative of the potential power in bringing language into robotics and in building flexible, interactive, and robust systems by bringing the fields together.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Takayama L, Ju W, Nass C. 2008. Beyond dirty, dangerous and dull: what everyday people think robots should do. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, pp. 25–32. New York: ACM
2. Guizzo E, Ackerman E. 2012. The rise of the robot worker. *IEEE Spectr.* 49:34–41
3. Zucker M, Joo S, Grey MX, Rasmussen C, Huang E, et al. 2015. A general-purpose system for teleoperation of the DRC-HUBO humanoid robot. *J. Field Robot.* 32:336–51
4. Bohren J, Rusu RB, Jones EG, Marder-Eppstein E, Pantofaru C, et al. 2011. Towards autonomous robotic butlers: lessons learned with the PR2. *2011 IEEE International Conference on Robotics and Automation*, pp. 5568–75. Piscataway, NJ: IEEE
5. Blank D, Kumar D, Meeden L, Yanco H. 2006. The Pyro toolkit for AI and robotics. *AI Mag.* 27(1):39–50

6. Kress-Gazit H, Fainekos GE. 2008. Translating structured English to robot controllers. *Adv. Robot.* 22:1343–59
7. Harnad S. 1990. The symbol grounding problem. *Phys. D* 42:335–46
8. Winograd T. 1970. *Procedures as a representation for data in a computer program for understanding natural language*. PhD Thesis, Mass. Inst. Technol., Cambridge
9. Murphy RR, Tadokoro S, Nardi D, Jacoff A, Fiorini P, et al. 2008. Search and rescue robotics. In *Springer Handbook of Robotics*, ed. B Siciliano, O Khatib, pp. 1151–73. Berlin: Springer
10. Chen DL, Mooney RJ. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 859–65. Palo Alto, CA: AAAI Press
11. Tellex S, Kollar T, Shaw G, Roy N, Roy D. 2010. Grounding spatial language for video search. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pap. 31. New York: ACM
12. Branavan S, Chen H, Zettlemoyer LS, Barzilay R. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Vol. 1, pp. 82–90. Stroudsburg, PA: Assoc. Comput. Linguist.
13. Matuszek C, FitzGerald N, Zettlemoyer L, Bo L, Fox D. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 2012 International Conference on Machine Learning*, pp. 1435–42. Madison, WI: Omnipress
14. Kollar T, Tellex S, Roy D, Roy N. 2010. Toward understanding natural language directions. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 259–66. Piscataway, NJ: IEEE
15. MacMahon MT. 2007. *Following natural language route instructions*. PhD Thesis, Univ. Tex., Austin. Data available at **http://robotics.csres.utexas.edu/~adastra/papers/MacMahon-Route-Instruction-Corpus.tar.gz**
16. Fong T, Nourbakhsh I, Dautenhahn K. 2003. A survey of socially interactive robots. *Robot. Auton. Syst.* 42:143–66
17. Goodrich MA, Schultz AC. 2007. Human-robot interaction: a survey. *Found. Trends Hum.-Comput. Interact.* 1:203–75
18. Thomaz A, Hoffman G, Cakmak M. 2016. Computational human-robot interaction. *Found. Trends Robot.* 4:105–223
19. Mooney RJ. 2008. Learning to connect language and perception. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, ed. D Fox, CP Gomes, pp. 1598–601. Palo Alto, CA: AAAI Press
20. Heim I, Kratzer A. 1998. *Semantics in Generative Grammar*. Oxford, UK: Blackwell
21. Pinker S. 2003. *The Language Instinct: How the Mind Creates Language*. London: Penguin
22. Hopcroft JE, Motwani R, Ullman JD. 2001. *Introduction to Automata Theory, Languages, and Computation*. Boston: Addison-Wesley. 2nd ed.
23. Sipser M. 2006. *Introduction to the Theory of Computation*. Boston: Thomson Course Technol.
24. Charniak E. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pp. 132–39. Stroudsburg, PA: Assoc. Comput. Linguist.
25. Klein D, Manning CD. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Vol. 1, pp. 423–30. Stroudsburg, PA: Assoc. Comput. Linguist.
26. Marcus MP, Marcinkiewicz MA, Santorini B. 1993. Building a large annotated corpus of English: the Penn Treebank. *Comput. Linguist.* 19:313–30
27. Steedman M. 2000. *The Syntactic Process*. Cambridge, MA: MIT Press
28. Zettlemoyer L, Collins M. 2005. Learning to map sentences to logical form: structured classification with probabilistic categorial grammars. *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 658–66. Arlington, VA: AUAI Press
29. Artzi Y, Zettlemoyer L. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Trans. Assoc. Comput. Linguist.* 1:49–62

30. Hockenmaier J, Steedman M. 2002. Generative models for statistical parsing with combinatory categorial grammar. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 335–42. Stroudsburg, PA: Assoc. Comput. Linguist.
31. Dzifcak J, Scheutz M, Baral C, Schermerhorn P. 2009. 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 *2009 IEEE International Conference on Robotics and Automation*, pp. 4163–68. Piscataway, NJ: IEEE
32. Matuszek C, Bo L, Zettlemoyer L, Fox D. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 2556–63. Palo Alto, CA: AAAI Press
33. Krishnamurthy J, Kollar T. 2013. Jointly learning to parse and perceive: connecting natural language to the physical world. *Trans. Assoc. Comput. Linguist.* 1:193–206
34. Jackendoff RS. 1983. Semantics of spatial expressions. In *Semantics and Cognition*, pp. 161–87. Cambridge, MA: MIT Press
35. Wierzbicka A. 1996. *Semantics: Primes and Universals*. Oxford, UK: Oxford Univ. Press
36. Talmy L. 2005. The fundamental system of spatial schemas in language. In *From Perception to Meaning: Schemas in Cognitive Linguistics*, ed. B Hamp, pp. 199–232. Berlin: De Gruyter Mouton
37. Huth M, Ryan M. 2004. *Logic in Computer Science: Modelling and Reasoning About Systems*. New York: Cambridge Univ. Press
38. Emerson EA. 1990. Temporal and modal logic. In *Handbook of Theoretical Computer Science*, Vol. B, ed. J van Leeuwen, pp. 995–1072. Cambridge, MA: MIT Press
39. Kress-Gazit H, Lahijanian M, Raman V. 2018. Synthesis for robots: guarantees and feedback for robot behavior. *Annu. Rev. Control Robot. Auton. Syst.* 1:211–36
40. Manning CD, Schütze H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press
41. Jurafsky D, Martin J. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall. 2nd ed.
42. Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Comput.* 9:1735–80
43. Mikolov T, Chen K, Corrado G, Dean J. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL]
44. Le Q, Mikolov T. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, ed. EP Xing, T Jebara, pp. 1188–96. Proc. Mach. Learn. Res. Vol. 32. N.p.: PMLR
45. Fellbaum C, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press
46. Schuler KK. 2005. *VerbNet: a broad-coverage, comprehensive verb lexicon*. PhD Thesis, Univ. Pa., Philadelphia
47. Lignos C, Raman V, Finucane C, Marcus M, Kress-Gazit H. 2015. Provably correct reactive control from natural language. *Auton. Robots* 38:89–105
48. Baker CF, Fillmore CJ, Lowe JB. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, Vol. 1, pp. 86–90. Stroudsburg, PA: Assoc. Comput. Linguist.
49. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–55. Piscataway, NJ: IEEE
50. Mao J, Huang J, Toshev A, Camburu O, Yuille AL, Murphy K. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20. Piscataway, NJ: IEEE
51. Laird JE. 2012. *The Soar Cognitive Architecture*. Cambridge, MA: MIT Press
52. Trafton JG, Hiatt LM, Harrison AM, Tamborello FP II, Khemlani SS, Schultz AC. 2013. ACT-R/E: an embodied cognitive architecture for human-robot interaction. *J. Hum.-Robot Interact.* 2:30–55
53. Schermerhorn PW, Kramer JF, Middendorff C, Scheutz M. 2006. DIARC: a testbed for natural human-robot interaction. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Vol. 2, pp. 1972–73. Palo Alto, CA: AAAI Press

54. Scheutz M, Williams T, Krause E, Oosterveld B, Sarathy V, Frasca T. 2018. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*, ed. MIA Ferreira, JS Sequeira, R Ventura, pp. 165–93. Cham, Switz.: Springer
55. Rubinoff R, Lehman JF. 1994. Real-time natural language generation in NL-Soar. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pp. 199–206. Stroudsburg, PA: Assoc. Comput. Linguist.
56. Huffman SB, Laird JE. 1993. Learning procedures from interactive natural language instructions. In *Machine Learning: Proceedings of the Tenth International Conference*, ed. PE Utgoff, pp. 143–50. San Francisco: Morgan Kaufmann
57. Mohan S, Mininger A, Kirk J, Laird JE. 2012. *Learning grounded language through situated interactive instruction*. Tech. Rep. FS-12-07, Assoc. Adv. Artif. Intell., Palo Alto, CA
58. Cantrell R, Scheutz M, Schermerhorn P, Wu X. 2010. Robust spoken instruction understanding for HRI. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 275–82. Piscataway, NJ: IEEE
59. Cantrell R, Schermerhorn P, Scheutz M. 2011. Learning actions from human-robot dialogues. In *2011 IEEE International Workshop on Robot and Human Interactive Communication*, pp. 125–30. Piscataway, NJ: IEEE
60. Krause E, Zillich M, Williams T, Scheutz M. 2014. Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 2796–802. Palo Alto, CA: AAAI Press
61. Williams T, Briggs G, Oosterveld B, Scheutz M. 2015. Going beyond command-based instructions: extending robotic natural language interaction capabilities. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1387–93. Palo Alto, CA: AAAI Press
62. Williams T, Acharya S, Schreitter S, Scheutz M. 2016. Situated open world reference resolution for human-robot dialogue. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 311–18. Piscataway, NJ: IEEE
63. Nyga D, Beetz M. 2012. Everything robots always wanted to know about housework (but were afraid to ask). In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 243–50. Piscataway, NJ: IEEE
64. Simmons RG, Bruce A, Goldberg D, Goode A, Montemerlo M, et al. 2003. *GRACE and GEORGE: autonomous robots for the AAAI robot challenge*. Tech. Rep. WS-03-01, Assoc. Adv. Artif. Intell., Palo Alto, CA
65. Perzanowski D, Schultz AC, Adams W. 1998. Integrating natural language and gesture in a robotics domain. In *Proceedings of the 1998 IEEE International Symposium on Intelligent Control (ISIC) Held Jointly with IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA) Intelligent Systems and Semiotics (ISAS)*, pp. 247–52. Piscataway, NJ: IEEE
66. MacMahon MT. 2006. Walk the talk: connecting language, knowledge, and action in route instructions. In *Proceedings of the 21st National Conference on Artificial Intelligence*, Vol. 2, pp. 1475–82. Palo Alto, CA: AAAI Press
67. Tenorth M, Nyga D, Beetz M. 2010. Understanding and executing instructions for everyday manipulation tasks from the world wide web. In *2010 IEEE International Conference on Robotics and Automation*, pp. 1486–91. Piscataway, NJ: IEEE
68. de Marneffe MC, MacCartney B, Manning CD. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Vol. 6, pp. 449–54. Bern, Switz.: Eur. Lang. Res. Assoc.
69. Lenat DB. 1995. CYC: a large-scale investment in knowledge infrastructure. *Commun. ACM* 38:33–38
70. Kirk NH, Nyga D, Beetz M. 2014. Controlled natural languages for language generation in artificial cognition. In *2014 IEEE International Conference on Robotics and Automation*, pp. 6667–72. Piscataway, NJ: IEEE
71. Nyga D, Roy S, Paul R, Park D, Pomarlan M, et al. 2018. Grounding robot plans from natural language instructions with incomplete world knowledge. In *Proceedings of the 2nd Conference on Robot Learning*, ed. A Billard, A Dragan, J Peters, J Morimoto, pp. 714–23. Proc. Mach. Learn. Res. Vol. 87. N.p.: PMLR

72. Raman V, Lignos C, Finucane C, Lee K, Marcus M, Kress-Gazit H. 2013. Sorry Dave, I'm afraid I can't do that: explaining unachievable robot tasks using natural language. In *Robotics: Science and Systems IX*, ed. P Newman, D Fox, D Hsu, pap. 23. N.p.: Robot. Sci. Syst. Found.
73. Boteanu A, Howard T, Arkin J, Kress-Gazit H. 2016. A model for verifiable grounding and execution of complex natural language instructions. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2649–54. Piscataway, NJ: IEEE
74. Boteanu A, Arkin J, Patki S, Howard T, Kress-Gazit H. 2017. Robot-initiated specification repair through grounded language interaction. In *Proceedings of the 2017 AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration*. Palo Alto, CA: AAAI Press
75. Howard TM, Tellex S, Roy N. 2014. A natural language planner interface for mobile manipulators. In *2014 IEEE International Conference on Robotics and Automation*, pp. 6652–59. Piscataway, NJ: IEEE
76. Siskind JM. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.* 15:31–90
77. Matuszek C, Fox D, Koscher K. 2010. Following directions using statistical machine translation. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 251–58. Piscataway, NJ: IEEE
78. Tellex S, Thaker P, Deits R, Kollar T, Roy N. 2012. Toward information theoretic human-robot dialog. In *Robotics: Science and Systems VIII*, ed. N Roy, P Newman, S Srinivasa, pp. 409–16. Cambridge, MA: MIT Press
79. Mei H, Bansal M, Walter MR. 2016. Listen, attend, and walk: neural mapping of navigational instructions to action sequences. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2772–78. Palo Alto, CA: AAAI Press
80. Pillai N, Matuszek C. 2018. Unsupervised selection of negative examples for grounded language learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 6517–23. Palo Alto, CA: AAAI Press
81. Richards LE, Matuszek C. 2019. *Learning to understand noncategorical physical language for human-robot interactions*. Paper presented at the Workshop on AI and Its Alternatives in Assistive and Collaborative Robotics, Robotics: Science and Systems XV, Freiburg, Ger., June 22–26
82. Matuszek C, Herbst E, Zettlemoyer L, Fox D. 2012. Learning to parse natural language commands to a robot control system. In *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, ed. JP Desai, G Dudek, O Khatib, V Kumar, pp. 403–15. Cham, Switz.: Springer
83. Tellex S, Kollar T, Dickerson S, Walter M, Banerjee A, et al. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 1507–14. Palo Alto, CA: AAAI Press
84. Chen DL, Kim J, Mooney RJ. 2010. Training a multilingual sportscaster: using perceptual context to learn language. *J. Artif. Intell. Res.* 37:397–435
85. MacGlashan J, Babes-Vroman M, desJardins M, Littman M, Muresan S, et al. 2015. Grounding English commands to reward functions. In *Robotics: Science and Systems XI*, ed. LE Kavraki, D Hsu, J Buchli, pap. 18. N.p.: Robot. Sci. Syst. Found.
86. Brown PF, Pietra VJD, Pietra SAD, Mercer RL. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* 19:263–311
87. Misra DK, Sung J, Lee K, Saxena A. 2014. Tell me Dave: context-sensitive grounding of natural language to mobile manipulation instructions. In *Robotics: Science and Systems X*, ed. D Fox, LE Kavraki, H Kurniawati, pap. 5. N.p.: Robot. Sci. Syst. Found.
88. Thomason J, Zhang S, Mooney R, Stone P. 2015. Learning to interpret natural language commands through human-robot dialog. *Proceedings of the Twenty-Fourth International Conference on Artificial Intelligence*, pp. 1923–29. Palo Alto, CA: AAAI Press
89. Fasola J, Matarić MJ. 2013. Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 143–50. Piscataway, NJ: IEEE
90. Brooks DJ, Lignos C, Finucane C, Medvedev MS, Perera I, et al. 2012. *Make it so: continuous, flexible natural language interaction with an autonomous robot*. Tech. Rep. WS-12-07, Assoc. Adv. Artif. Intell., Palo Alto, CA

91. Arumugam D, Karamcheti S, Gopalan N, Wong LLS, Tellex S. 2017. Accurately and efficiently interpreting human-robot instructions of varying granularities. In *Robotics: Science and Systems XIII*, ed. N Amato, S Srinivasa, N Ayanian, S Kuindersma, pap. 56. N.p.: Robot. Sci. Syst. Found.
92. Huang AS, Tellex S, Bachrach A, Kollar T, Roy D, Roy N. 2010. Natural language command of an autonomous micro-air vehicle. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2663–69. Piscataway, NJ: IEEE
93. Deits R, Tellex S, Thaker P, Simeonov D, Kollar T, Roy N. 2013. Clarifying commands with information-theoretic human-robot dialog. *J. Hum.-Robot Interact.* 2:58–79
94. Tellex S, Knepper R, Li A, Rus D, Roy N. 2014. Asking for help using inverse semantics. In *Robotics: Science and Systems X*, ed. D Fox, LE Kavraki, H Kurniawati, pap. 24. N.p.: Robot. Sci. Syst. Found.
95. Paul R, Arkin J, Roy N, Howard TM. 2016. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *Robotics: Science and Systems XII*, ed. D Hsu, N Amato, S Berman, S Jacobs, pap. 37. N.p.: Robot. Sci. Syst. Found.
96. Roy DK, Pentland AP. 2002. Learning words from sights and sounds: a computational model. *Cogn. Sci.* 26:113–46
97. Guadarrama S, Rodner E, Saenko K, Darrell T. 2015. Understanding object descriptions in robotics by open-vocabulary object retrieval and detection. *Int. J. Robot. Res.* 35:265–80
98. Blukis V, Misra D, Knepper RA, Artzi Y. 2018. Mapping navigation instructions to continuous control actions with position-visitation prediction. In *Proceedings of the 2nd Conference on Robot Learning*, ed. A Billard, A Dragan, J Peters, J Morimoto, pp. 505–18. Proc. Mach. Learn. Res. Vol. 87. N.p.: PMLR
99. Billard A, Dautenhahn K, Hayes G. 1998. Experiments on human-robot communication with Robota, an imitative learning and communicating doll robot. In *Socially Situated Intelligence: A Workshop Held at SAB'98, August 1998, Zürich*, ed. B Edmonds, K Dautenhahn, pp. 4–16. Zürich: Univ. Zürich
100. Cangelosi A, Hourdakis E, Tikhanoff V. 2006. Language acquisition and symbol grounding transfer with neural networks and cognitive robots. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pp. 1576–82. Piscataway, NJ: IEEE
101. Ahn H, Ha T, Choi Y, Yoo H, Oh S. 2018. Text2Action: generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation*, pp. 5915–20. Piscataway, NJ: IEEE
102. Gopalan N, Arumugam D, Wong L, Tellex S. 2018. Sequence-to-sequence language grounding of non-Markovian task specifications. In *Robotics: Science and Systems XIV*, ed. H Kress-Gazit, S Srinivasa, T Howard, N Atanasov, pap. 67. N.p.: Robot. Sci. Syst. Found.
103. Andreas J, Klein D. 2015. Alignment-based compositional semantics for instruction following. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1165–74. Stroudsburg, PA: Assoc. Comput. Linguist.
104. Das A, Datta S, Gkioxari G, Lee S, Parikh D, Batra D. 2018. Embodied question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–10. Piscataway, NJ: IEEE
105. Hermann KM, Hill F, Green S, Wang F, Faulkner R, et al. 2017. Grounded language learning in a simulated 3D world. arXiv:1706.06551 [cs.CL]
106. Misra D, Langford J, Artzi Y. 2017. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1004–15. Stroudsburg, PA: Assoc. Comput. Linguist.
107. Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, ed. F Pereira, CJC Burges, L Bottou, KQ Weinberger, pp. 1097–105. Red Hook, NY: Curran
108. Johnson J, Karpathy A, Fei-Fei L. 2016. DenseCap: fully convolutional localization networks for dense captioning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4565–74. Piscataway, NJ: IEEE
109. Karpathy A, Fei-Fei L. 2015. Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–37. Piscataway, NJ: IEEE
110. Shridhar M, Hsu D. 2018. Interactive visual grounding of referring expressions for human-robot interaction. In *Robotics: Science and Systems XIV*, ed. H Kress-Gazit, S Srinivasa, T Howard, N Atanasov, pap. 28. N.p.: Robot. Sci. Syst. Found.

111. Hatori J, Kikuchi Y, Kobayashi S, Takahashi K, Tsuboi Y, et al. 2018. Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE International Conference on Robotics and Automation*, pp. 3774–81. Piscataway, NJ: IEEE
112. Steels L, Kaplan F. 2000. AIBO's first words: the social learning of language and meaning. *Evol. Commun.* 4:3–32
113. Chao C, Cakmak M, Thomaz AL. 2011. Towards grounding concepts for transfer in goal learning from demonstration. In *2011 IEEE International Conference on Development and Learning*, Vol. 2. Piscataway, NJ: IEEE. **https://doi.org/10.1109/DEVLRN.2011.6037321**
114. Krening S, Harrison B, Feigh KM, Isbell CL, Riedl M, Thomaz A. 2016. Learning from explanations using sentiment and advice in RL. *IEEE Trans. Cogn. Dev. Syst.* 9:44–55
115. Forbes M, Rao RP, Zettlemoyer L, Cakmak M. 2015. Robot programming by demonstration with situated spatial language understanding. In *2015 IEEE International Conference on Robotics and Automation*, pp. 2014–20. Piscataway, NJ: IEEE
116. Artzi Y, Forbes M, Lee K, Cakmak M. 2014. Programming by demonstration with situated semantic parsing. In *Artificial Intelligence for Human-Robot Interaction: Papers from the 2014 AAAI Fall Symposium*, pp. 33–35. Palo Alto, CA: AAAI Press
117. Peng B, Loftin R, MacGlashan J, Littman ML, Taylor ME, Roberts DL. 2015. *Language and policy learning from human-delivered feedback*. Paper presented at the Machine Learning for Social Robotics Workshop, 2015 International Conference on Robotics and Automation, Seattle, WA, May 26–30
118. Loftin R, Peng B, MacGlashan J, Littman ML, Taylor ME, et al. 2014. Learning something from nothing: leveraging implicit human feedback strategies. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 607–12. Piscataway, NJ: IEEE
119. Loftin R, Peng B, MacGlashan J, Littman ML, Taylor ME, et al. 2016. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Auton. Agents Multi-Agent Syst.* 30:30–59
120. Thomaz AL, Breazeal C. 2008. Teachable robots: understanding human teaching behavior to build more effective robot learners. *Artif. Intell.* 172:716–37
121. Cakmak M, Chao C, Thomaz AL. 2010. Designing interactions for robot active learners. *IEEE Trans. Auton. Mental Dev.* 2:108–18
122. Thomason J, Padmakumar A, Sinapov J, Hart J, Stone P, Mooney RJ. 2017. Opportunistic active learning for grounding natural language descriptions. In *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 67–76. Proc. Mach. Learn. Res. Vol. 78. N.p.: PMLR
123. Padmakumar A, Stone P, Mooney RJ. 2018. Learning a policy for opportunistic active learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1347–57. Stroudsburg, PA: Assoc. Comput. Linguist.
124. Knox WB, Stone P, Breazeal C. 2013. Training a robot via human feedback: a case study. In *Social Robotics*, ed. G Herrmann, MJ Pearson, A Lenz, P Bremner, A Spiers, U Leonards, pp. 460–70. Cham, Switz.: Springer
125. Pillai N, Budhraja KK, Matuszek C. 2016. *Improving grounded language acquisition efficiency using interactive labeling*. Paper presented at the Workshop on Planning for Human-Robot Interaction: Shared Autonomy and Collaborative Robotics, Robotics: Science and Systems XII, Ann Arbor, MI, June 18–22
126. Kulick J, Toussaint M, Lang T, Lopes M. 2013. Active learning for teaching a robot grounded relational symbols. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 1451–57. Palo Alto, CA: AAAI Press
127. Paul R, Arkin J, Aksaray D, Roy N, Howard TM. 2018. Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms. *Int. J. Robot. Res.* 37:1269–99
128. Hayes B, Scassellati B. 2014. Discovering task constraints through observation and active learning. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4442–49. Piscataway, NJ: IEEE
129. Chao C, Thomaz A. 2016. Timed Petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration. *Int. J. Robot. Res.* 35:1330–53

130. Chao C, Lee J, Begum M, Thomaz AL. 2011. Simon plays Simon says: the timing of turn-taking in an imitation game. In *2011 IEEE International Workshop on Robot and Human Communication*, pp. 235–40. Piscataway, NJ: IEEE
131. Cakmak M, Thomaz AL. 2012. Designing robot learners that ask good questions. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 17–24. New York: ACM
132. Toh E, Poh L, Causo A, Tzuo PW, Chen I, et al. 2016. A review on the use of robots in education and young children. *J. Educ. Technol. Soc.* 19:148–63
133. Leite I, McCoy M, Lohani M, Ullman D, Salomons N, et al. 2015. Emotional storytelling in the classroom: individual versus group interaction between children and robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 75–82. New York: ACM
134. Gold K, Doniec M, Crick C, Scassellati B. 2009. Robotic vocabulary building using extension inference and implicit contrast. *Artif. Intell.* 173:145–66
135. Breazeal C, Harris PL, DeSteno D, Kory Westlund JM, Dickens L, Jeong S. 2016. Young children treat robots as informants. *Top. Cogn. Sci.* 8:481–91
136. Kory Westlund JM, Dickens L, Jeong S, Harris P, DeSteno D, Breazeal C. 2015. A comparison of children learning new words from robots, tablets, and people. In *Proceedings of New Friends 2015: The 1st International Conference on Social Robots in Therapy and Education*, ed. M Heerink, M de Jong, pp. 26–28. Almere, Neth.: Windesheim Flevoland
137. van den Berghe R, Verhagen J, Oudgenoeg-Paz O, van der Ven S, Leseman P. 2018. Social robots for language learning: a review. *Rev. Educ. Res.* 89:259–95
138. Ramachandran A, Huang CM, Gartland E, Scassellati B. 2018. Thinking aloud with a tutoring robot to enhance learning. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 59–68. New York: ACM
139. Clabaugh C, Ragusa G, Sha F, Matarić M. 2015. Designing a socially assistive robot for personalized number concepts learning in preschool children. In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, pp. 314–19. Piscataway, NJ: IEEE
140. Mutlu B, Shiwa T, Kanda T, Ishiguro H, Hagita N. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 61–68. New York: ACM
141. Mavridis N. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robot. Auton. Syst.* 63:22–35
142. Pejsa T, Andrist S, Gleicher M, Mutlu B. 2015. Gaze and attention management for embodied conversational agents. *ACM Trans. Interact. Intell. Syst.* 5:3
143. Görer B, Salah AA, Akın HL. 2017. An autonomous robotic exercise tutor for elderly people. *Auton. Robots* 41:657–78
144. Fasola J, Matarić MJ. 2015. Evaluation of a spatial language interpretation framework for natural human-robot interaction with older adults. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 301–8. Piscataway, NJ: IEEE
145. Fasola J, Matarić MJ. 2014. Interpreting instruction sequences in spatial language discourse with pragmatics towards natural human-robot interaction. In *2014 IEEE International Conference on Robotics and Automation*, pp. 2720–27. Piscataway, NJ: IEEE
146. Fasola J, Matarić MJ. 2012. Using socially assistive human–robot interaction to motivate physical exercise for older adults. *Proc. IEEE* 100:2512–26
147. Brawer J, Mangin O, Roncone A, Widder S, Scassellati B. 2018. Situated human–robot collaboration: predicting intent from grounded natural language. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 827–33. Piscataway, NJ: IEEE
148. Scassellati B, Brawer J, Tsui K, Nasihati Gilani S, Malzkuhn M, et al. 2018. Teaching language to deaf infants with a robot and a virtual human. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pap. 553. New York: ACM
149. Shimizu N, Haas A. 2009. Learning to follow navigational route instructions. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, pp. 1488–93. Palo Alto, CA: AAAI Press

150. Blisard SN, Skubic M. 2005. Modeling spatial referencing language for human-robot interaction. In *IEEE International Workshop on Robot and Human Interactive Communication*, pp. 698–703. Piscataway, NJ: IEEE
151. Mooney RJ. 2019. *A review of work on natural language navigation instructions*. Invited talk presented at the Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics, Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, June 6
152. Bollini M, Tellex S, Thompson T, Roy N, Rus D. 2013. Interpreting and executing recipes with a cooking robot. In *Experimental Robotics*, ed. J Desai, G Dudek, O Khatib, V Kumar, pp. 481–95. Heidelberg, Ger.: Springer
153. Beetz M, Klank U, Kresse I, Maldonado A, Mosenlechner L, et al. 2011. Robotic roommates making pancakes. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*, pp. 529–36. Piscataway, NJ: IEEE
154. Correa A, Walter MR, Fletcher L, Glass J, Teller S, Davis R. 2010. Multimodal interaction with an autonomous forklift. In *Proceeding of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 243–50. New York: ACM
155. Tasse D, Smith NA. 2008. *Sour cream: toward semantic processing of recipes*. Tech. Rep. CMU-LTI-08-005, Carnegie Mellon Univ., Pittsburgh, PA
156. Kiddon C, Ponnuraj GT, Zettlemoyer L, Choi Y. 2015. *Mise en place*: unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 982–92. Stroudburg, PA: Assoc. Comput. Linguist.
157. Cantrell R, Talamadupula K, Schermerhorn P, Benton J, Kambhampati S, Scheutz M. 2012. Tell me when and why to do it! Run-time planner model updates via natural language instruction. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 471–78. Piscataway, NJ: IEEE
158. Walter M, Hemachandra S, Homberg B, Tellex S, Teller S. 2013. Learning semantic maps from natural language descriptions. In *Robotics: Science and Systems IX*, ed. P Newman, D Fox, D Hsu, pap. 4. N.p.: Robot. Sci. Syst. Found.
159. Pronobis A, Jensfelt P. 2012. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *2012 IEEE International Conference on Robotics and Automation*, pp. 3515–22. Piscataway, NJ: IEEE
160. Bálint-Benczédi F, Mania P, Beetz M. 2016. Scaling perception towards autonomous object manipulation—in knowledge lies the power. In *2016 IEEE International Conference on Robotics and Automation*, pp. 5774–81. Piscataway, NJ: IEEE
161. Saunders J, Lehmann H, Förster F, Nehaniv CL. 2012. Robot acquisition of lexical meaning-moving towards the two-word stage. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics*. Piscataway, NJ: IEEE. **https://doi.org/10.1109/DevLrn.2012.6400588**
162. Lai K, Bo L, Ren X, Fox D. 2012. Detection-based object labeling in 3D scenes. In *2012 IEEE International Conference on Robotics and Automation*, pp. 1330–37. Piscataway, NJ: IEEE
163. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. 2010. Sun database: large-scale scene recognition from abbey to zoo. In *2010 IEEE conference on Computer Vision and Pattern Recognition*, pp. 3485–92. Piscataway, NJ: IEEE
164. Knepper RA, Layton T, Romanishin J, Rus D. 2013. IkeaBot: an autonomous multi-robot coordinated furniture assembly system. In *2013 IEEE International Conference on Robotics and Automation*, pp. 855–62. Piscataway, NJ: IEEE
165. Chai JY, Hong P, Zhou MX. 2004. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, pp. 70–77. New York: ACM
166. Whitney D, Eldon M, Oberlin J, Tellex S. 2016. Interpreting multimodal referring expressions in real time. In *2016 IEEE International Conference on Robotics and Automation*, pp. 3331–38. Piscataway, NJ: IEEE
167. Golland D, Liang P, Klein D. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 410–19. Stroudsburg, PA: Assoc. Comput. Linguist.

168. Landau B, Jackendoff R. 1993. "What" and "where" in spatial language and spatial cognition. *Behav. Brain Sci.* 16:217–65
169. Alomari M, Duckworth P, Hogg DC, Cohn AG. 2017. Natural language acquisition and grounding for embodied robotic systems. In *Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4349–56. Palo Alto, CA: AAAI Press
170. Tellex S, Roy D. 2009. Grounding spatial prepositions for video search. In *Proceedings of the International Conference on Multimodal Interfaces*, pp. 253–60. New York: ACM
171. Veloso MM, Biswas J, Coltin B, Rosenthal S. 2015. CoBots: robust symbiotic autonomous mobile service robots. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 4423–29. Palo Alto, CA: AAAI Press
172. Rosenthal S, Veloso M. 2011. Modeling humans as observation providers using POMDPs. In *IEEE International Workshop on Robot and Human Communication*, pp. 53–58. Piscataway, NJ: IEEE
173. Thomason J, Sinapov J, Mooney RJ, Stone P. 2018. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *Proceedings of the 32nd National Conference on Artificial Intelligence*, pp. 5520–27. Palo Alto, CA: AAAI Press
174. Thomason J, Sinapov J, Svetlik M, Stone P, Mooney R. 2016. Learning multi-modal grounded linguistic semantics by playing "I spy." In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 3477–83. Palo Alto, CA: AAAI Press
175. Mutlu B, Forlizzi J, Hodgins J. 2006. A storytelling robot: modeling and evaluation of human-like gaze behavior. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*, pp. 518–23. Piscataway, NJ: IEEE
176. Cascianelli S, Costante G, Ciarfuglia TA, Valigi P, Fravolini ML. 2018. Full-GRU natural language video description for service robotics applications. *IEEE Robot. Autom. Lett.* 3:841–48
177. Dale R, Reiter E. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19:233–63
178. Grice H. 1975. Logic and conversation. In *Syntax and Semantics*, Vol. 3: *Speech Acts*, ed. P Cole, JL Morgan, pp. 41–58. New York: Academic
179. Mitchell M, Van Deemter K, 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 Linguis Human Language Technologies*, pp. 1174–84. Stroudsburg, PA: Assoc. Comput. Linguist.
180. Fang R, Doering M, Chai JY. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 271–78. New York: ACM
181. Zender H, Kruijff GJM, Kruijff-Korbayová I. 2009. Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pp. 1604–9. Palo Alto, CA: AAAI Press
182. Bohus D, Horvitz E. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pap 5. New York: ACM
183. Okuno Y, Kanda T, Imai M, Ishiguro H, Hagita N. 2009. Providing route directions: design of robot's utterance, gesture, and timing. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 53–60. Piscataway, NJ: IEEE
184. Veloso M, Biswas J, Coltin B, Rosenthal S, Kollar T, et al. 2012. CoBots: collaborative robots servicing multi-floor buildings. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5446–47. Piscataway, NJ: IEEE
185. Marge M, Powers A, Brookshire J, Jay T, Jenkins OC, Geyer C. 2011. Comparing heads-up, hands-free operation of ground robots to teleoperation. In *Robotics: Science and Systems VII*, ed. H Durrant-Whyte, N Roy, P Abbeel, pp. 193–200. Cambridge, MA: MIT Press
186. Tse R, Campbell ME. 2015. Human-robot information sharing with structured language generation from probabilistic beliefs. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1242–48. Piscataway, NJ: IEEE

187. Chai JY, Gao Q, She L, Yang S, Saba-Sadiya S, Xu G. 2018. Language to action: towards interactive task learning with physical agents. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 2–9. Palo Alto, CA: AAAI Press
188. Kollar T, Perera V, Nardi D, Veloso M. 2013. Learning environmental knowledge from task-based human-robot dialog. In *2013 IEEE International Conference on Robotics and Automation*, pp. 4304–9. Piscataway, NJ: IEEE
189. Suhr A, Lewis M, Yeh J, Artzi Y. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, pp. 217–23. Stroudsburg, PA: Assoc. Comput. Linguist.
190. Johnson J, Hariharan B, van der Maaten L, Fei-Fei L, Zitnick CL, Girshick R. 2017. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1988–97. Piscataway, NJ: IEEE
191. Gordon D, Kembhavi A, Rastegari M, Redmon J, Fox D, Farhadi A. 2018. IQA: visual question answering in interactive environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4089–98. Piscataway, NJ: IEEE
192. Anderson P, Wu Q, Teney D, Bruce J, Johnson M, et al. 2018. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–83. Piscataway, NJ: IEEE
193. Blukis V, Brukhim N, Bennett A, Knepper RA, Artzi Y. 2018. Following high-level navigation instructions on a simulated quadcopter with imitation learning. In *Robotics: Science and Systems XIV*, ed. H Kress-Gazit, S Srinivasa, T Howard, N Atanasov, pap. 66. N.p.: Robot. Sci. Syst. Found.