Robots That Use Language: A Survey

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

[1] Computer Science Department, Brown University, Providence, RI, USA, 02906; email: stefie10@cs.brown.edu
[2] School of Interactive Computing, Georgia Institute of Technology, Atlanta, Georgia, USA, 30332; email: ngopalan3@gatech.edu
[3] Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY, USA, 14853; email: hadaskg@cornell.edu
[4] Computer Science and Electrical Engineering Department, University of Maryland, Baltimore County, Baltimore, MD, USA, 21250; email: cmat@umbc.edu

Annual Review of Control, Robotics, and Autonomous Systems 2020. 3:1–35

https://doi.org/10.1146/((please add article doi))

Keywords

robots, language, grounding, learning, logic, dialog

Abstract

This paper 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 into 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 dialog 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.

Contents

1. INTRODUCTION

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 non-dedicated manufacturing spaces, and in the logistics industry (169, 69), 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 or touch-screen interfaces require training, and must be complex in order to enable a person to command complex robotic behavior (201). Higher level abstractions such as automata (19), programming abstractions (15), or structured language (96) offer a great degree of power and flexibility, but also require a great deal of training to use.

Natural Language Processing (NLP): Computational techniques for transforming human languages such as English into machine usable structures.

In contrast, people use language every day to direct behavior, ask and answer questions, provide information, and ask for help. Language-based interfaces require minimal user training and allow the expression of a variety of complex tasks. This paper reviews the current state of the art in natural language (NL) communication with robots, compares different approaches, and discusses the challenges of creating robust language-based human-robot interactions. The fundamental question for language-using robots is: how can words and language structures be *grounded* in the noisy, perceptual world in which a robot operates (70)?

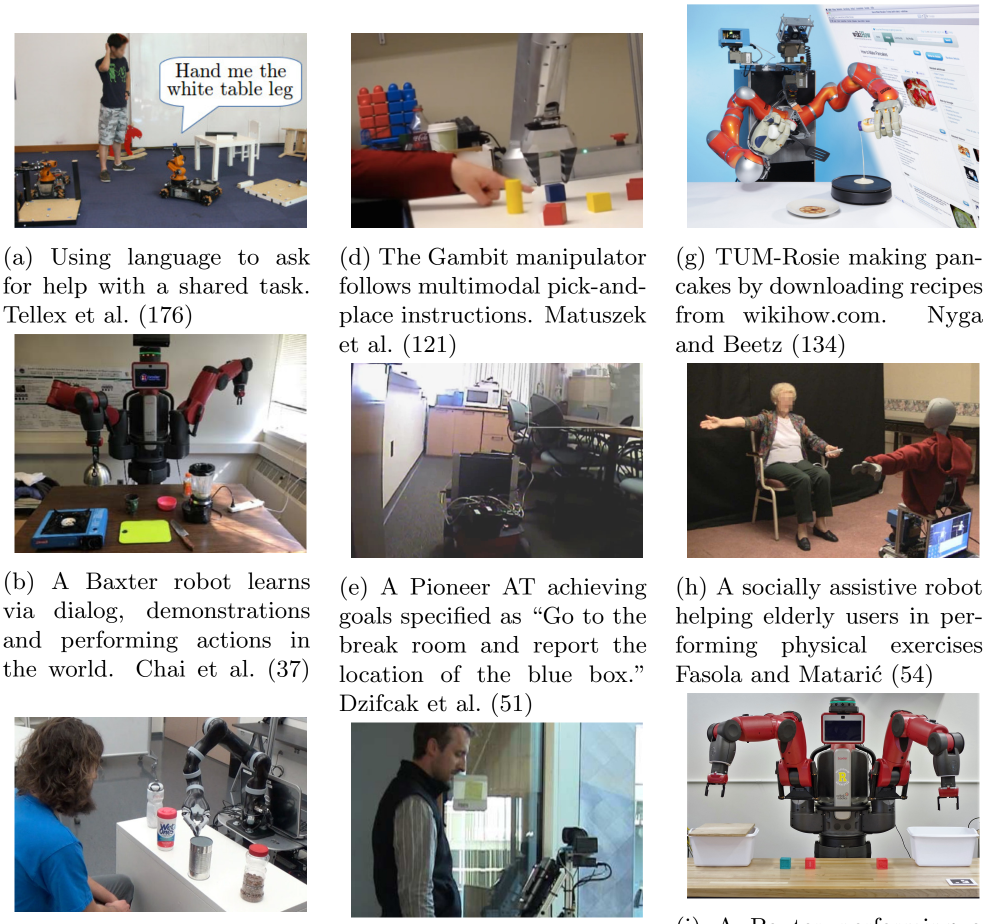
Grounded Language Understanding: Interpreting a natural language utterance in terms of the physical state of the robot and the environment.

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 gesture). Systems that address both problems enable robots to engage in *collaborative dialog*.

There is a long history of systems that try to understand natural language in physical domains, beginning with Winograd (197). Generally, language is most effective as an interface when users are untrained, are under high cognitive load, and when their hands and eyes are busy with other tasks. For example, in search-and-rescue tasks, robots might interact with human victims who are untrained and under great stress (131). The context in which language is situated can take many forms; examples include sportcasts of simulated soccer games (42), linguistic descriptions of spatial elements in video clips (174), GUI interaction (24), descriptions of objects in the world (118), spatial relationships (92), and the meaning of instructions (112). Language has also been used with a diverse group of robot platforms, ranging from manipulators to mobile robots to aerial robots. Some examples are shown in Figure 1.

Language for robotics is currently an area of significant research interest, as evidenced by many recent workshops (examples include (191, 120, 12, 6, 2)) and papers covered in this article. Other survey papers have reviewed related topics; for example, Fong et al. (59) surveys socially interactive robots, and Goodrich and Schultz (63) and Thomaz et al. (182) give broad surveys of human-robot interaction, although neither focuses on language specifically. This survey is intended for robotics researchers who wish to understand the current state of the art in natural language processing as it pertains to robotics.

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.



(a) Using language to ask for help with a shared task. Tellex et al. (176)

(b) A Baxter robot learns via dialog, demonstrations and performing actions in the world. Chai et al. (37)

(c) A Jaco arm identifying objects from attributes, here "silver, round, and empty." Thomason et al. (179)

(d) The Gambit manipulator follows multimodal pick-and-place instructions. Matuszek et al. (121)

(e) A Pioneer AT achieving goals specified as "Go to the break room and report the location of the blue box." Dzifcak et al. (51)

(f) CoBot learning to follow commands like "Take me to the meeting room." Kollar et al. (93)

(g) TUM-Rosie making pancakes by downloading recipes from wikihow.com. Nyga and Beetz (134)

(h) A socially assistive robot helping elderly users in performing physical exercises Fasola and Matarić (54)

(i) A Baxter performing a sorting task synthesized from natural language. Boteanu et al. (22)

Figure 1: Robots used for language-based interactions.

For example, the words "red block" might be converted to a formal expression such as $\lambda x : \textsc{block}(x) \wedge \textsc{red}(x)$. 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 paper 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 which organizes

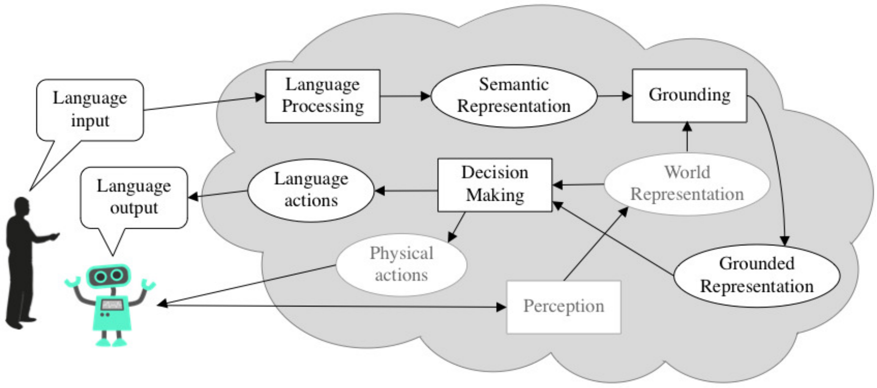


Figure 2: System diagram showing language input and output integrated into a robotic system. Many approaches include only a subset of the modules. Greyed-out modules are relevant to language interpretation but not reviewed in this paper.

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

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 later in the paper. We review the concept of grounded language, as well as the syntactic and semantic structure of language, and statistical language processing.

2.1. Grounded Language

Grounded (or situated, 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 (129, 70). Possible groundings range from low-level motor commands to perceptual inputs to complex sequences of actions. Grounded language *acquisition* is the process of learning these connections between percepts and actions. For example, if a person instructs a robot to "Pick up the cup," the robot must map the words "cup" to a particular set of percepts in its high dimensional sensor space. It must recognize that a particular pattern in its camera sensor, for example, corresponds to the word "cup." Then, to follow the command, it must produce a plan or a policy 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 causes the water to not 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 primarily used 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

Natural Language	Possible Sensor/Actuator	Category	Grounding/ Interpretation
“Turn left”	Wheels, Legs	Command understanding	Contra-rotate 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 robot
“Hand me the orange mug on the left”	Manipulator + all sensors above	Combined	All of the above

Table 1: Examples of natural language and possible groundings. *Column 1:* Natural language that might occur when instructing or informing a robot. *Column 2:* Possible sensors or actuators providing the physical context. *Column 3:* The underlying task or reasoning problem implicitly encoded in the language. *Column 4:* The physically situated, or *grounded*, meaning of the language.

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; a key research question is what the grounding process should look like, and how this mapping should be carried out.

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

2.2. Syntactic Representations and Analysis

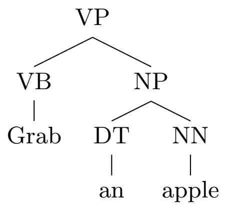
Natural language has a hierarchical, compositional syntax (73) which is studied in linguistics and cognitive semantics. This structure enables people to understand novel sentences by combining individual words in new ways (146). This *syntactic* structure can be used to help extract *semantic* representations 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 (77). CFGs and their many variants are used to describe the syntactic structure of natural language. A formal definition of CFGs can be found in Sipser (163), while Figure 3 gives an example CFG for a small subset of English with an associated parse tree. Many variants of CFGs exist. Pre-trained parsers are a common tool, many of them (41, 89) trained using the Penn Treebank (115), a corpus of text manually annotated with parse trees. Other parsers are trained on corpora of text, such as newspaper articles. This data is 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 structures that represents the syntactic decomposition of an utterance in the form of a rooted tree.

Many robotics applications use Combinatory Categorial Grammars (CCGs) (166). 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

$$
\begin{aligned}
VP &\rightarrow VB \quad NP\\
NP &\rightarrow DT \quad NN\\
DT &\rightarrow \text{the}|\text{a}|\text{an}\\
NN &\rightarrow \text{block}|\text{ball}|\text{apple}\\
VB &\rightarrow \text{grab}|\text{move}
\end{aligned}
$$

(a) CFG for a small subset of English.



(b) Parse tree and sentence produced from the grammar. The structure defines compositional relations among word meanings.

Figure 3: Grammar and parse tree for the English sentence "Grab an apple".

approach which 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 (200, 7, 76), typically learning from pairs of NL sentences and sentence meanings. CCGs have been applied to robotic language understanding in many contexts (51, 121, 7, 98), reviewed in all of 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 $\exists x(\textsc{apple}(x) \wedge \textsc{grab}(x))$ (that states "there exists an x that is an apple and that is being grabed"). Given a consistent formal meaning for, e.g., $\textsc{grab}()$, this expression can be interpreted and used for understanding actions in the world. There is extensive work on symbolic representations of semantics, for example (82, 194, 170, 73). 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 semantics 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 $\textsc{grab}()$ could be grounded to a manipulation action.[1]

Temporal logics are modal logics that contain temporal operators (52), 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) vs. branching (formulas

[1] Additional reading regarding the formal syntax and semantics of first-order predicate logic can be found in logic texts, for example Huth and Ryan (81).

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.

are defined over trees of executions), and whether they are deterministic, or include probabilistic operators and reasoning. The survey (97) describes the use of several temporal logics in the context of robot control.

2.4. Statistical NLP and Deep Learning

Substantial progress in NLP has been made by eschewing the explicit modeling of linguistics structures. For example, n-gram models which focus on counting words (113) robustly capture aspects of language use without requiring a full understanding of syntax or meaning, by leveraging the statistics of word co-occurrance. Shallow parsing or chunking has been shown to be useful to capture aspects of syntax and semantics without performing a complete analysis of the sentence (85). 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 (LSTMs) (75) combined with Word2Vec (124) or Paragragh Vector (104). 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 robotic 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 survey paper 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 logic, 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-centric 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, but rather to provide a possible framework for considering the overall research space.

3.1. Lexically-Grounded Methods

In this section, we describe 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; frequently temporal logics are used, as there are algorithms to transform the resulting formulas into behavior that provides guarantees on performance and correctness (97). 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 either be learned or manually prescribed, but as opposed 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, it is common to use existing knowledge bases and cognitive architectures to automatically enrich the lexicon using a base set of manual groundings.

Knowledge Bases and Ontologies. Many existing knowledge bases provide real-world "common sense" knowledge that can be used to create language-using robots. WordNet (58) 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 (159) 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 (107). Similarly, FrameNet (10) created a dataset of verb meanings with parameterized actions. ImageNet (50) is an image database organized along nouns in the WordNet hierarchy. This dataset has been used extensively in computer vision and provides information that could enable a robot to detect objects and ground noun phrases. Datasets that are specific to a particular type of grounding task also exist, such as RefCOCO (114) for referring expressions.

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 there has been work done with Soar (102, 165), DIARC (157, 158), and ACT-R (185), among others.

Soar (102, 165) 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 (154), 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 dialog with a person. Building on this work, Instructo-Soar was introduced by Huffman and Laird (80), enabling grounding new instructions to procedures in Soar. Instructo-Soar assumes simple imperative sentences which are straightforward to parse and instantiate as a new operator template. Language groundings can also be learned from mixed-initiative human-robot interactions that include language, gestures and perceptual information (128). 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 (Adaptive Character of Thought-Rational) and ACT-R/E (Adaptive Character of Thought-Rational/Embodied), introduced by Trafton et al. (185), 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 Distributed Integrated Cognition Affect and Reflection (DIARC) Architecture (157, 158), 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. In the DIARC Architecture (158, 32, 33, 94), researchers created a system that incrementally processes natural language utterances, creates goals for a planner, and executes the instructions, shown in Figure 1e (51). In that work, the lexicon is labeled with both syntactic annotations from a combinatory categorial grammar (CCG (166, 7)) and semantic annotations in the form of λ-expressions related to the temporal logic CTL* (52), and first-order dynamic logic (FDL). 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 employs pragmatic inference to enable more complex language interaction where the meaning of the utterances may be implicit and where context and semantics are combined (195, 196).

Probabilistic Action Cores (PRAC) (134), while not a cognitive architecture *per se*, generalize the notion of a knowledge base by creating a system that enables inferring over, disambiguating, and completing vague or under specified natural language instructions by using information from existing lexical databases and by drawing on background knowledge from WordNet and framenet, among others. From this information the robot can infer a motor action that causes a source object to end up in a goal location. An image from this work is shown in Figure 1g.

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 datasets 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 semantics 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 dialog manager or a planner that executes situated robot actions. Broadly speaking, the following approaches to language interactions follow a similar pipeline: NL utterances in the form of text are syntactically parsed, 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 use formal representations for natural language interactions were Grace and George, robots that competed in the AAAI robot challenges. At AAAI 2004, Grace acted as an information kiosk providing information about the conference and giving directions, while George physically escorted people to their destination (162). Both robots utilized the Nautilus parser (142), which uses a context-free grammar 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 (111) was created to interpret route instructions given in NL, combining syntactic and semantic structures with information from the perception system regarding the environment.

Grounding and executing NL instructions from websites such as wikiHow.com was explored by Tenorth et al. (177). The system uses the Stanford parser (48) in which a probabilistic context-free grammar is used to syntactically parse instructions. These instructions are grounded using WordNet (58) and Cyc (106) and are captured as a set of instructions in a knowledge base. Later work (88) discussed controlled natural language as a way to repair missing information through explicit clarification. Nyga et al. (135) used a similar probabilistic model for using relational knowledge to fill in gaps for aspects of the language missing from the workspace.

In Raman et al. (149), Lignos et al. (107), high-level natural language tasks are grounded to Linear Temporal Logic (LTL) (52) formulas by using part-of-speech tagging and parsing to create the syntactic structure. VerbNet (159) 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 (22, 23), semantic mappings are learned using the distributed correspondence graph (78) framework; an image from this work is shown in Figure 1i.

Siskind (164) presents 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.

Data and Domains for Learning Methods. Learning-based approaches use a wide variety of datasets, 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. Often, an annotated symbolic representation is 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. Some of the common datasets being used currently in language grounding and robotics are listed in Table 2 along with the type of sensor, language and annotation data.

We accompany Table 2 with a brief example of applying a dataset for a robotic task. The MARCO dataset (111) of navigation instructions is the most widely used of the existing datasets (111, 117, 92, 175, 7). Beyond being one of the earliest available datasets in this space, its wide uptake is partly because it contains not only route directions, but a complete simulation environment in which to navigate. Thus a potential user of the dataset does 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 on the same problem by using the NL instructions in MARCO, then testing in the same simulated environment.

For example, ten years after the original work used a hand-crafted grammar to explicitly model language (111), Mei et al. (123) use a long short-term memory recurrent neural net (LSTM-RNN) to learn to follow directions. This work estimates action sequences from NL 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 datasets is the mismatch between the data provided and the 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 dataset uses butterflies as a landmark object; most real environments do not have these butterflies, but have other landmarks that may not appear in MARCO. Learning more general concepts such as 'landmarks' is a key 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 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 (73). However, in more machine learning oriented work, there is a trend towards systems that learn the representation space itself from data, leading towards systems that do not need a designer to pre-specify a fully populated set of symbols and allowing robots to adapt to unexpected input. For example, Matuszek et al. (118) and Pillai and Matuszek (143) 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 and Matuszek (150) extend that work to creating symbols that are not category-limited.

We divide the following approaches into those that primarily use pre-defined languages (section 3.2.1), those that are more concerned with discovering the domain (section 3.2.2), and recent work on using deep neural networks for language understanding (section 3.2.3). In practice, work in this area falls along a spectrum, ranging from formal methods approaches which use completely manually defined word meanings (111), to learning mappings between words and a prespecified formal language (42, 119, 22), to learning new symbols from data while specifying perceptually motivated features (172), to learning new features from data as well as a mapping between word meanings and those features (118).

3.2.1. 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 (111, 117, 42). The same approach can be used for the inverse problem of *generating* natural language descriptions of formally represented events, such

Dataset	Type of Data	Link to dataset
MARCO dataset (111)	Navigation instructions given to a robot to navigate a map, and the route followed.	`www.cs.utexas.edu/users/ml/clamp/navigation/`
Scene dataset(98)	Images and descriptions of objects in the image.	`rtw.ml.cmu.edu/tacl2013_lsp/`
Cornell NLVR dataset (168)	Pairs of images and logical statements about them which are true or false.	`lic.nlp.cornell.edu/nlvr/`
CLEVR dataset (84)	Images and question-answer pairs.	`cs.stanford.edu/people/jcjohns/clevr/`
Embodied Question Answering (47)	Pairs of questions and answers in simulated 3D environments. The agent needs to search the environment to find the answer.	`embodiedqa.org`
Visual Question Answering in Interactive Environments (65)	Pairs of questions and answers in different simulated 3D environments.	`github.com/danielgordon10/thor-iqa-cvpr-2018`
Room-to-Room (R2R) Navigation (4)	Panoramic views in real buildings, paired with instructions to be followed.	`bringmeaspoon.org/`
H2R lab language grounding datasets (9, 64)	Predicate based sub-goal conditions paired with natural language instructions.	`github.com/h2r/language_datasets`
Cornell Instruction Following Framework (17, 125)	Data for three separate navigation domains in 3D environments, containing instructions paired with trajectories.	`github.com/clic-lab/ciff`
MIT Spatial Language Understanding dataset (92, 172)	Pairs of language command and trajectories for navigation and mobile manipulation.	`people.csail.mit.edu/stefie10/slu/`

Table 2: Datasets used in Language Grounding and Robotics

as RoboCup soccer games (43). MacGlashan et al. (110) 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 (28); 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. (126) learns 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. (119) created a system that learns to parse NL directions into RCL, a robot 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 and Zettlemoyer (7) created a system for learning semantic parses for mapping instructions to actions in order to follow natural language route instructions, while Thomason et al. (178) learn semantic parse information and grounded word meanings from dialog interactions with users. Fasola and Mataric (55) used a probabilistic approach to learn mappings between commands and a space of actions of service robots, including models for spatial prepositions. Brooks et al. (27), Boteanu et al. (22, 23) and Arumugam et al. (9) ground language to objects and specifications expressed in Linear Temporal Logic. 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.

3.2.2. Learning to Map to Undefined Spaces.

We draw a distinction between learning to map between predefined symbol spaces and approaches which 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 framework (G^3) (172) was introduced to interpret natural language commands given to a robotic forklift, as well as to interpret route instructions for a wheelchair (92) and a micro-air vehicle (79). 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 NL questions that clarify ambiguous commands (49, 175), and then to enable robots to ask for help (176). It has been extended by Howard et al. (78) to create an efficient interface for interpreting grounded language by mapping to planning formalisms; this approach dramatically increases the speed that words can be interpreted by the robot. Building on this framework, Paul et al. (138) 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 and Pentland (152) 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. (118) created a system for learning word meanings for words by automatically creating new features for visual object attributes, while Pillai and Matuszek (143) learned to select negative examples for grounded language learning. Guadarrama et al. (68) 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 dataset. Blukis et al. (18) learn 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.

3.2.3. Grounding Language using Deep Learning. Modern deep learning based approaches of Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Deep-Q Networks (DQN) 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 like Billard et al. (14), Cangelosi et al. (31) learned robot behavior from demonstrations neural networks and mapped language to these behaviors. Roy and Pentland (153) 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. (123, 1)). In this paper we do not survey these methods in great detail, but 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.

Instruction following with Seq2Seq. Some of the earliest progress was made in the area of instruction following (123, 9). This is a supervised problem where given a natural language command, a sequence of actions is performed by the agent. 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 RNN based Sequence to Sequence (seq2seq) approaches (123). Other have abstracted the problem to learn the grounding from natural langauge to sub-goals or goals (64, 5). These methods have been implemented on robots only when the abstract fixed grounding symbols have been provided (9).

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. (24). More recently it has been applied to richer environments with visual inputs (47, 74, 125). A common strategy is to model the agent and its environment as a Markov Decision Process (MDP), and encode the instruction given to the agent as the state of the environment. Such agents have been able to answer question about the properties of objects, or navigate to objects in simulation. This approach is hard to implement on a physical robot given the number of episodes required to learn behaviors.

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 started as classifiers to recognize an object class within an image (99). This work then progressed to captioning images densely, that is, recognizing all objects within an image (83, 86). A general approach, first described in (86) 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 (83). These methods have been implemented in physical robots in an object retrieval setting by training the robot on simulated images (161, 71).

Grounding control from robot perception. Blukis et al. (18) learned to map between navigation instructions and low level control actions, mediated by the robot's sensor input and control actions. This work aims to perform end-to-end learning from language to control actions and has since been demonstrated on physical robots.

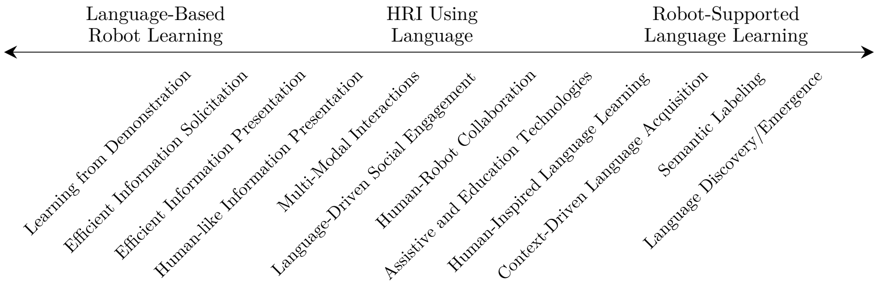

Language-Based Robot Learning
HRI Using Language
Robot-Supported Language Learning
Learning from Demonstration
Efficient Information Solicitation
Efficient Information Presentation
Human-like Information Presentation
Multi-Modal Interactions
Language-Driven Social Engagement
Human-Robot Collaboration
Assistive and Education Technologies
Human-Inspired Language Learning
Context-Driven Language Acquisition
Semantic Labeling
Language Discovery/Emergence

Figure 4: A categorization of work using language for human-robot interaction. This visualization spans efforts using language to support efficient robot learning, efforts to use language in order to maximize the effectiveness of human-robot interactions, and using robots as physically situated agents to support language learning.

3.3. HRI-Centered Approaches

The final broad category of work we consider is that which lies primarily in the area of human-robot interaction. While work in the previous sections are 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 as 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 support robots learning other tasks. On the other, 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 human-robot interaction (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 learning, especially from human teachers. Natural language provides an accessible, rich 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. In this section, we describe current work on developing robot systems that learn about the world from natural language inputs. This includes efforts on: 1) Learning from demonstration; 2) Learning reward functions from language; 3) Active learning; and 4) 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 *learning from demonstration,* or 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 and Kaplan (167) used language and camera percepts to learn novel instance based objects and their association with words. Billard et al. (14) used LfD to ground language with a constrained vocabulary to a sequences of actions demonstrated by the teacher. Chao et al. (39) 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. (95) 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 Forbes et al. (60). Programming by demonstration can also rely on more complex semantic parsing, as in Artzi et al. (8), 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 less situation-specific learning (110). This approach can allow a non-specialist to give an agent explicit reward signals (141), or can model implicit feedback strategies inherent in human teaching (108, 109).

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 (183) is balanced with selecting data that optimizes learning, as queries to a user are a sharply limited resource (30). In general, incorporating active learning makes learning more efficient and makes it possible to learn from fewer data points (145, 180, 137). This form of learning can be implemented in a domain-independent way, as in Knox et al. (91), and can improve efficiency on learning tasks, including both explicit language grounding (144) and more general robotics problems, such as learning conceptual symbols (100)), spatial concepts (139), or task constraints (72).

Another topic in learning from language provided by nonspecialists is how to correctly elicit information and demonstrations from people. Chao and Thomaz (38), explore conducting dialog correctly, with appropriate multi-modal timing, turn-taking, and responsiveness behavior (40). It also means figuring out what questions to ask; Cakmak and Thomaz (29) studied how humans asked questions and designed an approach to asking appropriately targeted questions for learning from demonstration, while Pillai and Matuszek (143) demonstrates 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. Human-robot interaction, or 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 human-robot interaction is a broad, active field of study. In this section, we attempt to provide an overview of some of the categories of current HRI/language research.

Childhood education is a significant area of research for human-robot interaction studies (184), both because there is a chronic shortage of personnel in education and childcare, and because increasing the role of technology in childhood education is a critical part of attracting a larger and more diverse population into STEM fields. Research in this area largely focuses on the role of interactive play in child development. This can take the form of acting out stories between children and robots (105), assisting with language development (61, 26, 192, 187), or serving as intelligent tutoring systems (148, 44).

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 (133). There is a growing body of work in which language is incorporated into multimodal human-robot interactions (122). Matuszek et al. (121) 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. (79) use modeling to evaluate robots' use of gesture. In the inverse direction, Pejsa et al. (140) use people's speech, gaze, and gestures to learn a multimodal interaction model, which is 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 (66, 57), incorporating language pragmatics and anaphor resolution (56, 54) as well as verbal feedback. Verbal robots have also been explored in the context of autism support (25) and tutoring for Deaf infants (156).

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. Specific subproblems include command understanding, where a person gives a command, as well as a person providing information to the robot.

4.1.1. Giving Robots Instructions. *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 (111) created a large dataset of route directions in simulation, which has been used in a number of papers (42, 7). Kollar et al. (92)

used a statistical approach to interpret instructions for a robotic wheelchair. Shimizu and Haas (160) used a CRF approach to learn word meanings, and Matuszek et al. (117) 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 (92, 111), robotic UAVs (79), and mobile robots (16). Understanding navigational commands remains a significant and ongoing area of research (130).

A second class of problems is interpreting natural language commands for manipulator robots. This problem has been studied in the subdomain of interpreting textual recipes (21, 13), following instructions for a robotic forklift (172), interpreting instructions to a tabletop arm (121, 177), and a Baxter robot (22, 23). Such language may refer only to the robot's motion; for example, Correa et al. (45) 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 semi-constrained language of recipes. Beetz et al. (13) used a reasoning system to interpret recipes and cook pancakes. Tasse and Smith (171) created a dataset of recipes mapped to a formal symbolic representation, while Kiddon et al. (87) created an unsupervised hard EM approach to automatically mapping recipes to sequenced action graphs; neither system used robots. Bollini et al. (21) created a system for interpreting recipes, but did not ground ingredients into perception. Although the language of recipes is constrained, it remains a challenging problem to understand, in part because ingredients combine into new things that do not exist at the time of original interpretation. Flour, eggs, water and sugar are transformed to a batter, which is then transformed to 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." This perceptual 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. Telling Robots About the World. A second element of language interpretation is enabling robots to use language to improve their knowledge of the world. Compared to instruction-following, this topic is a less studied area, but there is nonetheless a rich array of approaches. Cantrell et al. (34) created a system that updates its planning model based on human instructions, while the system of Walter et al. (190) incorporates information from language into a semantic map of the environment. Pronobis and Jensfelt (147) describe a multi-modal probabilistic framework incorporating 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").

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 follows:

- General language about object *characteristics*, such as color, shape, or size. (121, 11, 155)
- Descriptions of objects at the *type*, or category membership, level. This encompasses approaches that tie language into object recognition. (101, 198, 143, 153)
- Language about particular *instances* of objects, such as "my mug." (92, 172, 196, 167)

These categories often overlap. As examples, 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 live in" (121), 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, then turn the cap" to refer to them separately. For assembly tasks, a part such as a screw and a table leg may combine to form a completed assembly, the table (176, 90). Grounding these sorts of expressions is an open problem.

Referring Expressions: Natural language expressions that uniquely denote objects, areas, or people to which the speaker is referring. (85)

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." (172, 196)) Chai et al. (36) created a system that interprets multimodal referring expressions using a graph-based approach. Whitney et al. (193) and Matuszek et al. (121) merge 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. (121) is shown in Figure 1d. Golland et al. (62) generate spatial descriptions using game theory to generate human-interpretable referring expressions in a virtual environment.

Spatial Relationships. Interpreting spatial relationships is a well-known, complex problem in NLP. For route instructions, this may take the form of "the door near the elevators," or "past the kitchen." In object descriptions, it may be "at the top left corner." Understanding these 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 only has a few and new ones are very 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 (170, 103). Some work has focused specifically on spatial prepositions (3, 174, 173, 139). 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, to convey information, and to provide context. People have studied the problem of enabling a robot to produce natural language, either answering questions, asking for help, or providing instructions. This problem is the inverse problem from language understanding: the robot desires to communicate something to the person and must find words to speak that convey its ideas. Subproblems include robots instructing people, robots asking questions, and robots informing people.

4.2.1. Robots Instructing and Querying People. Often a robot might use language to try to get a person to do something, typically by asking for help or asking them 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 and Matarić (54) used templates to generate language to motivate physical exercise for older adults, shown in Figure 1h. 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. (90) generate 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 from whom to ask for assistance (189).

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. (49) presented an algorithm to generate targeted questions based on information theory to reduce confusion. Rosenthal and Veloso (151) modeled humans as information providers, using a POMDP to ask questions when the robot encountered problems. Thomason et al. (181) 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 (179), shown in Figure 1c. Cakmak and Thomaz (29), Pillai et al. (144), and others use of 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. (43) 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. (132) created a storytelling robot that uses language as well as gaze to engage a human listener. Cascianelli et al. (35) created a system for enabling a robot to learn to describe events in a video stream and released a dataset for service robotic 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 as a robot may need to understand references to things in its environment (see 4.1.2), a robot may need to generate referring expressions about objects, landmarks, or people. Dale and Reiter (46) 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 (67) and focusing on computational cost. This approach assumes a symbolic representation of context, rather than grounding to perception. Golland et al. (62) generated spatial descriptions using game theory to produce referring expressions in a virtual environment that are interpretable by their human partner. Mitchell et al. (127) generated expressions that refer to visible objects that a robot might observe with its camera. Tellex et al. (176) provided an inverse-semantics algorithm for generating requests for help, including expressions such as “the black leg on the white table” (shown in Figure 1a). Golland et al. (62) generated spatial descriptions to objects in a virtual environment using a game-theoretic approach to find the best language to pinpoint the object. Fang et al. (53) 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. (199) created a system for enabling a mobile robot to generate natural language referring expressions to objects in the environment, as well as 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 production in order 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 dialog. Such a robot must interpret a person’s communicative acts as well as generating communicative actions of its own. Two-way communication requires more than simply combining language understanding and generation. A robot must reason about uncertainty in its own percepts, retain conversational state, react quickly to a person’s input, and work towards 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 dialog 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 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 comprises one communication mode in the overall interaction.

For example, Bohus and Horvitz (20) 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. (121) created a system for learning from unscripted deictic gesture combined with language in order to perform manipulations. Okuno et al. (136) created a robot for giving route directions by integrating language utterances, gestures, and timing. Fasola and Matarić (54) created a socially assistive robot system designed to engage elderly users in physical exercise. Veloso et al. (188) created the CoBots, mobile robots that engage in tasks in an office environment such as fetching objects. Marge et al. (116) created a heads-up hands-free approach for controlling a pack-bot as it moved on the ground. Tse and 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 (37), shown in Figure 1b. The CoBot systems learned to follow commands like “Take me to the meeting room,” engaging in dialog with humans in its environment to improve its ability (shown in Figure 1f) (93).

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 inter-agent 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

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 paper 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 present 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.

5.1. Open Questions

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 learning approaches to 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 terror of language is its ability to construct arbitrarily fine-grained and specific sentences applying to all 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 dialog by the robot. Similarly, the robot must combine verbal and non-verbal 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 the potential for building flexible, interactive, and robust systems by bringing the fields together.

LITERATURE CITED

1. Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5. IEEE, 2018.
2. James F. Allen, Jacob Andreas, Jason Baldridge, Mohit Bansal, Archna Bhatia, Yonatan Bisk, Asli Celikyilmaz, Bonnie J. Dorr, Parisa Kordjamshidi, Matthew Marge, and Jesse Thomason. NAACL Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP), 2019. URL `https://splu-robonlp.github.io/`.
3. Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. Natural language acquisition and grounding for embodied robotic systems. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
4. Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
5. Jacob Andreas and Dan Klein. Alignment-based compositional semantics for instruction following. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1165–1174, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
6. Jacob Arkin, Andrea F. Daniele, Nakul Gopalan, Thomas M. Howard, Jesse Thomason, Matthew R. Walter, and Lawson L.S. Wong. Robotics: Science and Systems Workshop on Models and Representations for Natural Human-Robot Communication, 2018. URL `http://www2.ece.rochester.edu/projects/rail/mrhrc2018/`.
7. Yoav Artzi and Luke Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics (TACL)*, 1:49–62, 2013.
8. Yoav Artzi, Maxwell Forbes, Kenton Lee, and Maya Cakmak. Programming by demonstration with situated semantic parsing. In *2014 AAAI Fall Symposium Series*, 2014.
9. Dilip Arumugam, Siddharth Karamcheti, Nakul Gopalan, Lawson L. S. Wong, and Stefanie Tellex. Accurately and efficiently interpreting human-robot instructions of varying granularities. In *Robotics: Science and Systems XIII, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, July 12-16, 2017*, 2017. . URL `http://www.roboticsproceedings.org/rss13/p56.html`.
10. Collin F Baker, Charles J Fillmore, and John B Lowe. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.
11. Ferenc Bálint-Benczédi, Patrick Mania, and Michael Beetz. Scaling perception towards autonomous object manipulation—in knowledge lies the power. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5774–5781. IEEE, 2016.

12. Mohit Bansal, Cynthia Matuszek, Jacob Andreas, Yoav Artzi, and Yonatan Bisk. ACL Language Grounding for Robotics, 2017. URL `http://projects.csail.mit.edu/spatial/workshop`.

13. Michael Beetz, Ulrich Klank, Ingo Kresse, Andres Maldonado, Lorenz Mosenlechner, Dejan Pangercic, Thomas Ruhr, and Moritz Tenorth. Robotic roommates making pancakes. In *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, pages 529–536. IEEE, 2011.

14. Aude Billard, Kerstin Dautenhahn, and Gillian Hayes. Aibos first words: The social learning of language and meaning. In *Proceedings of Socially Situated Intelligence Workshop held within the Fifth Conference on Simulation of Adaptive Behavior (SAB98). Centre for Policy Modelling technical report series: No. CPM9838*, 1998.

15. Douglas Blank, Deepak Kumar, Lisa Meeden, and Holly Yanco. The Pyro toolkit for AI and robotics. *AI magazine*, 27(1):39, 2006.

16. Samuel N. Blisard and Marjorie Skubic. Modeling spatial referencing language for human-robot interaction. In *IEEE International Workshop on Robot and Human Interactive Communication*, pages 698–703, 2005.

17. Valts Blukis, Nataly Brukhim, Andrew Bennett, Ross A. Knepper, and Yoav Artzi. Following high-level navigation instructions on a simulated quadcopter with imitation learning. In *Proceedings of the Robotics: Science and Systems Conference*, 2018.

18. Valts Blukis, Dipendra Misra, Ross A Knepper, and Yoav Artzi. Mapping navigation instructions to continuous control actions with position-visitation prediction. In *Proceedings of the 2^{nd} Conference on Robot Learning*, 2018.

19. Jonathan Bohren, Radu Bogdan Rusu, E Gil Jones, Eitan Marder-Eppstein, Caroline Pantofaru, Melonee Wise, Lorenz Mösenlechner, Wim Meeussen, and Stefan Holzer. Towards autonomous robotic butlers: Lessons learned with the PR2. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 5568–5575. IEEE, 2011.

20. Dan Bohus and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, page 5. ACM, 2010.

21. Mario Bollini, Stefanie Tellex, Tyler Thompson, Nicholas Roy, and Daniela Rus. Interpreting and executing recipes with a cooking robot. In *Experimental Robotics*, pages 481–495. Springer, 2013.

22. Adrian Boteanu, Thomas Howard, Jacob Arkin, and Hadas Kress-Gazit. A model for verifiable grounding and execution of complex natural language instructions. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016)*, 2016.

23. Adrian Boteanu, Jacob Arkin, Siddharth Patki, Thomas Howard, and Hadas Kress-Gazit. Robot-initiated specification repair through grounded language interaction. In *AAAI Fall Symposium 2017, Natural Communication for Human-Robot Collaboration*, 2017.

24. S.R.K. Branavan, Harr Chen, Luke S Zettlemoyer, and Regina Barzilay. 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: Volume 1-Volume 1*, pages 82–90. Association for Computational Linguistics, 2009.

25. Jake Brawer, Olivier Mangin, Alessandro Roncone, Sarah Widder, and Brian Scassellati. Situated human–robot collaboration: Predicting intent from grounded natural language. In *Intelligent Robots and Systems (IROS)*, 2018.

26. Cynthia Breazeal, Paul L Harris, David DeSteno, Jacqueline M Kory Westlund, Leah Dickens, and Sooyeon Jeong. Young children treat robots as informants. *Topics in cognitive science*, 8 (2):481–491, 2016.

27. Daniel J Brooks, Constantine Lignos, Cameron Finucane, Mikhail S Medvedev, Ian Perera, Vasumathi Raman, Hadas Kress-Gazit, Mitch Marcus, and Holly A Yanco. Make it so: Con-

tinuous, flexible natural language interaction with an autonomous robot. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

28. Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.
29. Maya Cakmak and Andrea L Thomaz. Designing robot learners that ask good questions. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 17–24. ACM, 2012.
30. Maya Cakmak, Crystal Chao, and Andrea L Thomaz. Designing interactions for robot active learners. *Autonomous Mental Development, IEEE Transactions on*, 2(2):108–118, 2010.
31. Angelo Cangelosi, Emmanouil Hourdakis, and Vadim Tikhanoff. Language acquisition and symbol grounding transfer with neural networks and cognitive robots. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1576–1582. IEEE, 2006.
32. Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and Xuan Wu. Robust spoken instruction understanding for HRI. In *Proceedings of the 2010 Human-Robot Interaction Conference*, pages 275–282, March 2010.
33. Rehj Cantrell, Paul Schermerhorn, and Matthias Scheutz. Learning actions from human-robot dialogues. In *Proceedings of the 2011 IEEE Symposium on Robot and Human Interactive Communication*, July 2011.
34. Rehj Cantrell, Kartik Talamadupula, Paul Schermerhorn, J. Benton, Subbarao Kambhampati, and Matthias Scheutz. Tell me when and why to do it!: Run-time planner model updates via natural language instruction. In *Proceedings of the 2012 Human-Robot Interaction Conference*, Boston, MA, March 2012.
35. Silvia Cascianelli, Gabriele Costante, Thomas A Ciarfuglia, Paolo Valigi, and Mario L Fravolini. Full-gru natural language video description for service robotics applications. *IEEE Robotics and Automation Letters*, 3(2):841–848, 2018.
36. Joyce Y Chai, Pengyu Hong, and Michelle X Zhou. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, pages 70–77. ACM, 2004.
37. Joyce Y. Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. Language to action: Towards interactive task learning with physical agents. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 2–9, 2018. .
38. Crystal Chao and Andrea Thomaz. Timed Petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration. *The International Journal of Robotics Research*, 35(11):1330–1353, 2016.
39. Crystal Chao, Maya Cakmak, and Andrea L Thomaz. Towards grounding concepts for transfer in goal learning from demonstration. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1–6. IEEE, 2011.
40. Crystal Chao, Jinhan Lee, Momotaz Begum, and Andrea L Thomaz. Simon plays Simon says: The timing of turn-taking in an imitation game. In *RO-MAN, 2011 IEEE*, pages 235–240. IEEE, 2011.
41. Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139. Association for Computational Linguistics, 2000.
42. David L Chen and Raymond J Mooney. Learning to interpret natural language navigation instructions from observations. In *AAAI*, volume 2, pages 1–2, 2011.
43. David L Chen, Joohyun Kim, and Raymond J Mooney. Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research*, pages 397–435, 2010.
44. Caitlyn Clabaugh, Gisele Ragusa, Fei Sha, and Maja Matarić. Designing a socially assistive

robot for personalized number concepts learning in preschool children. In *Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2015 Joint IEEE International Conference on*, pages 314–319. IEEE, 2015.

45. Andrew Correa, Matthew R. Walter, Luke Fletcher, Jim Glass, Seth Teller, and Randall Davis. Multimodal interaction with an autonomous forklift. In *Proceeding of the 5th ACM/IEEE International Conference on Human-robot Interaction*, pages 243–250, Osaka, Japan, 2010. ACM.
46. Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263, 1995.
47. Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
48. Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.
49. Robin Deits, Stefanie Tellex, Pratiksha Thaker, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, 2013.
50. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
51. J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *IEEE International Conference on Robotics and Automation*, pages 4163–4168, 2009.
52. E. Allen Emerson. Temporal and modal logic. In Jan van Leeuwen, editor, *Handbook of Theoretical Computer Science (Vol. B)*, pages 995–1072. MIT Press, Cambridge, MA, USA, 1990. ISBN 0-444-88074-7. URL `http://dl.acm.org/citation.cfm?id=114891.114907`.
53. Rui Fang, Malcolm Doering, and Joyce Y Chai. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 271–278. ACM, 2015.
54. Juan Fasola and Maja J Matarić. Using socially assistive human–robot interaction to motivate physical exercise for older adults. *Proceedings of the IEEE*, 100(8):2512–2526, 2012.
55. Juan Fasola and Maja J Mataric. Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 143–150. IEEE, 2013.
56. Juan Fasola and Maja J Mataric. Interpreting instruction sequences in spatial language discourse with pragmatics towards natural human-robot interaction. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 2720–2727. IEEE, 2014.
57. Juan Fasola and Maja J Matarić. Evaluation of a spatial language interpretation framework for natural human-robot interaction with older adults. In *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*, pages 301–308. IEEE, 2015.
58. Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.
59. Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.
60. Maxwell Forbes, Rajesh PN Rao, Luke Zettlemoyer, and Maya Cakmak. Robot programming by demonstration with situated spatial language understanding. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 2014–2020. IEEE, 2015.

61. Kevin Gold, Marek Doniec, Christopher Crick, and Brian Scassellati. Robotic vocabulary building using extension inference and implicit contrast. *Artificial Intelligence (AIJ)*, 173(1): 145–166, 2009.

62. Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419. Association for Computational Linguistics, 2010.

63. Michael A Goodrich and Alan C Schultz. Human-robot interaction: A survey. *Foundations and trends in human-computer interaction*, 1(3):203–275, 2007.

64. Nakul Gopalan, Dilip Arumugam, Lawson Wong, and Stefanie Tellex. Sequence-to-Sequence Language Grounding of Non-Markovian Task Specifications. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. .

65. Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: Visual question answering in interactive environments. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

66. Binnur Görer, Albert Ali Salah, and H Levent Akın. An autonomous robotic exercise tutor for elderly people. *Autonomous Robots*, 41(3):657–678, 2017.

67. HP Grice. Logic and conversation. In *Syntax and Semantics Volume 3: Speech Acts*. Academic Press, New York, 1975.

68. Sergio Guadarrama, Erik Rodner, Kate Saenko, and Trevor Darrell. Understanding object descriptions in robotics by open-vocabulary object retrieval and detection. *The International Journal of Robotics Research*, page 0278364915602059, 2015.

69. Erico Guizzo and Evan Ackerman. The rise of the robot worker. *IEEE Spectrum*, 49(10): 34–41, 2012.

70. Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1): 335–346, 1990.

71. Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781. IEEE, 2018.

72. Bradley Hayes and Brian Scassellati. Discovering task constraints through observation and active learning. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 4442–4449. IEEE, 2014.

73. Irene Heim and Angelika Kratzer. *Semantics in generative grammar*, volume 13. Blackwell Oxford, 1998.

74. Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. Grounded language learning in a simulated 3d world. *CoRR*, abs/1706.06551, 2017.

75. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

76. Julia Hockenmaier and Mark Steedman. Generative models for statistical parsing with combinatory categorial grammar. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 335–342. Association for Computational Linguistics, 2002.

77. John E Hopcroft, Rajeev Motwani, and Jeffrey D Ullman. Introduction to automata theory, languages, and computation. *ACM SIGACT News*, 32(1):60–65, 2001.

78. Thomas M Howard, Stefanie Tellex, and Nicholas Roy. A natural language planner interface for mobile manipulators. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 6652–6659. IEEE, 2014.

79. A. S Huang, S. Tellex, A. Bachrach, T. Kollar, D. Roy, and N. Roy. Natural language command of an autonomous micro-air vehicle. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2663–2669, October 2010.

80. Scott B. Huffman and John E. Laird. Learning procedures from interactive natural language instructions. In *Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993*, pages 143–150. Morgan Kaufmann, 1993. ISBN 1-55860-307-7.

81. Michael Huth and Mark Ryan. *Logic in Computer Science: Modelling and Reasoning About Systems*. Cambridge University Press, New York, NY, USA, 2004. ISBN 052154310X.

82. Ray S. Jackendoff. *Semantics and Cognition*, pages 161–187. MIT Press, 1983.

83. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

84. Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

85. D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, Pearson Education International, 2^{nd} edition edition, 2008.

86. Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

87. Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. Mise en place: Unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 982–992, 2015.

88. Nicholas Hubert Kirk, Daniel Nyga, and Michael Beetz. Controlled Natural Languages for Language Generation in Artificial Cognition. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, Hong Kong, China, May 31-June 7 2014.

89. Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

90. Ross A Knepper, Todd Layton, John Romanishin, and Daniela Rus. Ikeabot: An autonomous multi-robot coordinated furniture assembly system. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 855–862. IEEE, 2013.

91. W Bradley Knox, Peter Stone, and Cynthia Breazeal. Training a robot via human feedback: A case study. In *Social Robotics*, pages 460–470. Springer, 2013.

92. Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, pages 259–266, 2010.

93. Thomas Kollar, Viraga Perera, Damiano Nardi, and Marco Veloso. Learning environmental knowledge from task-based human-robot dialog. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 4304–4309. IEEE, 2013.

94. Evan Krause, Michael Zillich, Thomas Williams, and Matthias Scheutz. Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In *Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

95. Samantha Krening, Brent Harrison, Karen M Feigh, Charles Lee Isbell, Mark Riedl, and Andrea Thomaz. Learning from explanations using sentiment and advice in RL. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55, 2016.

96. Hadas Kress-Gazit and Georgios E Fainekos. Translating structured English to robot controllers. *Advanced Robotics*, 22:1343–1359, 2008.

97. Hadas Kress-Gazit, Morteza Lahijanian, and Vasumathi Raman. Synthesis for robots: Guarantees and feedback for robot behavior. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:211–236, 2018.

98. Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting

natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013.

99. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
100. Johannes Kulick, Marc Toussaint, Tobias Lang, and Manuel Lopes. Active learning for teaching a robot grounded relational symbols. In *IJCAI*, 2013.
101. Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Detection-based object labeling in 3D scenes. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1330–1337. IEEE, 2012.
102. John E. Laird. *The Soar Cognitive Architecture.* The MIT Press, 2012. ISBN 0262122960, 9780262122962.
103. Barbara Landau and Ray Jackendoff. "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–265, 1993.
104. Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.
105. Iolanda Leite, Marissa McCoy, Monika Lohani, Daniel Ullman, Nicole Salomons, Charlene K Stokes, Susan Rivers, and Brian Scassellati. Emotional storytelling in the classroom: Individual versus group interaction between children and robots. In *HRI*, pages 75–82, 2015.
106. Douglas B Lenat. CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.
107. Constantine Lignos, Vasumathi Raman, Cameron Finucane, Mitchell Marcus, and Hadas Kress-Gazit. Provably correct reactive control from natural language. *Auton. Robots*, 38: 89–105, 2015. ISSN 0929-5593. . URL `http://dx.doi.org/10.1007/s10514-014-9418-8`.
108. Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jie Huang, and David L Roberts. Learning something from nothing: Leveraging implicit human feedback strategies. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 607–612. IEEE, 2014.
109. Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous Agents and Multi-Agent Systems*, 30(1):30–59, 2016.
110. J MacGlashan, M Babes-Vroman, M desJardins, M Littman, S Muresan, S Squire, S Tellex, D Arumugam, and L Yang. Grounding English commands to reward functions. In *Robotics: Science and Systems*, 2015.
111. Matthew Tierney MacMahon. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1475–1482, 2006.
112. Matthew Tierney MacMahon. *Following Natural Language Route Instructions.* PhD thesis, University of Texas at Austin, Department of Electrical & Computer Engineering, August 2007. Data available at http://robotics.csres.utexas.edu/ adastra/papers/MacMahon-Route-Instruction-Corpus.tar.gz.
113. Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, 1999.
114. Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.
115. Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn treebank. *Computational linguistics*, 19(2):313–330, 1993.
116. Matthew Marge, Aaron Powers, Jonathan Brookshire, Trevor Jay, Odest C Jenkins, and

Christopher Geyer. Comparing heads-up, hands-free operation of ground robots to teleoperation. In *Robotics: Science and Systems*, 2011.

117. Cynthia Matuszek, Dieter Fox, and Karl Koscher. Following directions using statistical machine translation. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, pages 251–258, 2010.
118. Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proceedings of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June 2012.
119. Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *Proceedings of the 13th Int'l Symposium on Experimental Robotics (ISER)*, June 2012.
120. Cynthia Matuszek, Stefanie Tellex, Dieter Fox, and Luke Zettlemoyer. Grounding language for physical systems, 2012. URL `http://www.aaai.org/Library/Workshops/ws12-07.php`.
121. Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the 28th National Conference on Artificial Intelligence (AAAI)*, Québec City, Quebec, Canada, March 2014.
122. Nikolaos Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35, 2015.
123. Hongyuan Mei, Mohit Bansal, and Matthew R Walter. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
124. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
125. Dipendra Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1004–1015, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
126. Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. In *International Journal of Robotics Research (IJRR)*, 2016.
127. Margaret Mitchell, Kees Van Deemter, and Ehud Reiter. Generating expressions that refer to visible objects. In *HLT-NAACL*, pages 1174–1184, 2013.
128. Shiwali Mohan, Aaron Mininger, James Kirk, and John E. Laird. Learning grounded language through situated interactive instruction. In *Robots Learning Interactively from Human Teachers, Papers from the 2012 AAAI Fall Symposium, Arlington, Virginia, USA, November 2-4, 2012*, volume FS-12-07 of *AAAI Technical Report*. AAAI, 2012. URL `http://www.aaai.org/ocs/index.php/FSS/FSS12/paper/view/5662`.
129. Raymond J. Mooney. Learning to connect language and perception. In Dieter Fox and Carla P. Gomes, editors, *Proc. of the Twenty-Third AAAI Conf. on Artificial Intelligence, AAAI 2008*, pages 1598–1601, Chicago, Illinois, July 2008. AAAI Press.
130. Raymond J. Mooney. Invited talk: A review of work on natural language navigation instructions. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP), 2019. URL `http://tiny.cc/MooneyNAACL-RoboNLP2019`.
131. Robin R Murphy, Satoshi Tadokoro, Daniele Nardi, Adam Jacoff, Paolo Fiorini, Howie Choset, and Aydan M Erkmen. Search and rescue robotics. In *Springer Handbook of Robotics*, pages 1151–1173. Springer, 2008.
132. Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, pages 518–523. IEEE, 2006.

133. Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 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*, pages 61–68. ACM, 2009.
134. Daniel Nyga and Michael Beetz. Everything robots always wanted to know about housework (but were afraid to ask). In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura, Portugal, October, 7–12 2012.
135. Daniel Nyga, Subhro Roy, Rohan Paul, Daehyung Park, Mihai Pomarlan, Michael Beetz, and Nicholas Roy. Grounding robot plans from natural language instructions with incomplete world knowledge. In *Conference on Robot Learning*, pages 714–723, 2018.
136. Yusuke Okuno, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. Providing route directions: design of robot's utterance, gesture, and timing. In *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*, pages 53–60. IEEE, 2009.
137. Aishwarya Padmakumar, Peter Stone, and Raymond J. Mooney. Learning a policy for opportunistic active learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP-18)*, Brussels, Belgium, 2018. URL `http://www.cs.utexas.edu/users/ai-lab/pub-view.php?PubID=127713`.
138. Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M. Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *Proceedings of Robotics: Science and Systems*, AnnArbor, Michigan, June 2016. .
139. Rohan Paul, Jacob Arkin, Derya Aksaray, Nicholas Roy, and Thomas M Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms. *The International Journal of Robotics Research*, page 0278364918777627, 2018.
140. Tomislav Pejsa, Sean Andrist, Michael Gleicher, and Bilge Mutlu. Gaze and attention management for embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(1):3, 2015.
141. Bei Peng, Robert Loftin, James MacGlashan, Michael L. Littman, Matthew E. Taylor, and David L. Roberts. Language and Policy Learning from Human-delivered Feedback. In *Proceedings of the Machine Learning for Social Robotics workshop (at ICRA)*, May 2015.
142. D. Perzanowski, A. C. Schultz, and W. Adams. 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) Intell*, pages 247–252, Sept 1998. .
143. Nisha Pillai and Cynthia Matuszek. Unsupervised selection of negative examples for grounded language learning. In *AAAI*, 2018.
144. Nisha Pillai, Karan K Budhraja, and Cynthia Matuszek. Improving grounded language acquisition efficiency using interactive labeling. *Proceedings of the R:SS 2016 workshop on Planning for Human-Robot Interaction: Shared Autonomy and Collaborative Robotics*, 2016.
145. Nisha Pillai, Francis Ferraro, and Cynthia Matuszek. Active learning methods for efficient grounded language acquisition. Under Review - EMNLP 2019, 2019.
146. Steven Pinker. *The language instinct: How the mind creates language*. Penguin UK, 2003.
147. Andrzej Pronobis and Patric Jensfelt. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3515–3522, 2012.
148. Aditi Ramachandran, Chien-Ming Huang, Edward Gartland, and Brian Scassellati. Thinking aloud with a tutoring robot to enhance learning. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 59–68. ACM, 2018.
149. Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton Lee, Mitch Marcus, and Hadas Kress-Gazit. Sorry dave, i'm afraid i can't do that: Explaining unachievable robot tasks using natural language. In *Robotics: Science and Systems*, 2013.

150. Luke E. Richards and Cynthia Matuszek. Learning to understand non-categorical physical language for human-robot interactions. *Proceedings of the R:SS 2019 workshop on AI and Its Alternatives in Assistive and Collaborative Robotics (RSS: AI+ACR)*, 2019.
151. Stephanie Rosenthal and Manuela Veloso. Modeling humans as observation providers using pomdps. In *RO-MAN, 2011 IEEE*, pages 53–58. IEEE, 2011.
152. Deb K Roy and Alex P Pentland. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146, 2002.
153. Deb K Roy and Alex P Pentland. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146, 2002.
154. Robert Rubinoff and Jill Fain Lehman. Real-time natural language generation in nl-soar. In *Proceedings of the Seventh International Workshop on Natural Language Generation*, INLG '94, pages 199–206, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1641417.1641440`.
155. Joe Saunders, Hagen Lehmann, Frank Förster, and Chrystopher L Nehaniv. Robot acquisition of lexical meaning-moving towards the two-word stage. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–7. IEEE, 2012.
156. Brian Scassellati, Jake Brawer, Katherine Tsui, Setareh Nasihati Gilani, Melissa Malzkuhn, Barbara Manini, Adam Stone, Geo Kartheiser, Arcangelo Merla, Ari Shapiro, et al. 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*, page 553. ACM, 2018.
157. Paul W Schermerhorn, James F Kramer, Christopher Middendorff, and Matthias Scheutz. DIARC: A testbed for natural human-robot interaction. In *AAAI*, pages 1972–1973, 2006.
158. Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. An overview of the distributed integrated cognition affect and reflection diarc architecture. In Maria Isabel Aldinhas Ferreira, Joo S.Sequeira, and Rodrigo Ventura, editors, *Cognitive Architectures.* Springer, 2018.
159. Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon.* PhD thesis, University of Pennsylvania, 2005.
160. Nobuyuki Shimizu and Andrew Haas. Learning to Follow Navigational Route Instructions. In *Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, 2009.
161. Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.
162. Reid G. Simmons, Allison Bruce, Dani Goldberg, Adam Goode, Michael Montemerlo, Nicholas Roy, Brennan Sellner, Chris Urmson, Alan C. Schultz, William Adams, Magdalena D. Bugajska, Matt MacMahon, Jessica Mink, Dennis Perzanowski, Stephanie Rosenthal, Scott Thomas, Ian Horswill, Robert Zubek, David Kortenkamp, Bryn Wolfe, Tod Milam, and Bruce A. Maxwell. GRACE and GEORGE: Autonomous robots for the AAAI robot challenge. In *AAAI Mobile Robot Competition*, volume WS-03-01 of *AAAI Technical Report*, page 52. AAAI Press, 2003.
163. Michael Sipser. *Introduction to the Theory of Computation*, volume 2. Thomson Course Technology Boston, 2006.
164. Jeffrey Mark Siskind. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Int. Res.*, 15(1):31–90, 2001.
165. Soar. Soar website. https://soar.eecs.umich.edu, 2019.
166. Mark Steedman. *The syntactic process*, volume 24. MIT Press, 2000.
167. Luc Steels and Frederic Kaplan. Aibos first words: The social learning of language and meaning. *Evolution of communication*, 4(1):3–32, 2000.
168. Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada, July 2017. Asso-

ciation for Computational Linguistics.

169. Leila Takayama, Wendy Ju, and Clifford Nass. 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*, pages 25–32. ACM, 2008.
170. Leonard Talmy. The fundamental system of spatial schemas in language. In Beate Hamp, editor, *From Perception to Meaning: Schemas in Cognitive Linguistics*, pages 199–232. Mouton de Gruyter, 2005.
171. Dan Tasse and Noah A Smith. Sour cream: Toward semantic processing of recipes. Technical report, CMU-LTI-08-005, Carnegie Mellon University, Pittsburgh, PA, 2008.
172. S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A. Banerjee, S. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence*, 2011.
173. Stefanie Tellex and Deb Roy. Grounding spatial prepositions for video search. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI-2009)*, 2009. In press.
174. Stefanie Tellex, Thomas Kollar, George Shaw, Nicholas Roy, and Deb Roy. Grounding spatial language for video search. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, page 31:1–31:8, 2010.
175. Stefanie Tellex, Pratiksha Thaker, Robin Deits, Thomas Kollar, and Nicholas Roy. Toward information theoretic human-robot dialog. In *Robotics: Science and Systems*, 2012.
176. Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. In *Robotics: Science and systems*, volume 2, page 3, 2014.
177. M. Tenorth, D. Nyga, and M. Beetz. Understanding and executing instructions for everyday manipulation tasks from the world wide web. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 1486–1491, May 2010. .
178. Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. Learning to interpret natural language commands through human-robot dialog. *Proceedings of the Twenty-Fourth international joint conference on Artificial Intelligence (IJCAI)*, 2015.
179. Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond Mooney. Learning multi-modal grounded linguistic semantics by playing "I spy". In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.
180. Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Justin Hart, Peter Stone, and Raymond J Mooney. Opportunistic active learning for grounding natural language descriptions. In *Conference on Robot Learning*, pages 67–76, 2017.
181. Jesse Thomason, Jivko Sinapov, Raymond J Mooney, and Peter Stone. Guiding exploratory behaviors for multi-modal grounding of linguistic descriptions. In *Proceedings of the 32nd National Conference on Artificial Intelligence (AAAI)*, 2018.
182. Andrea Thomaz, Guy Hoffman, and Maya Cakmak. Computational human-robot interaction. *Foundations and Trends in Robotics*, 4(2-3):105–223, 2016. ISSN 1935-8253. . URL `http://dx.doi.org/10.1561/2300000049`.
183. Andrea L Thomaz and Cynthia Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6):716–737, 2008.
184. Emily Toh, Lai Poh, Albert Causo, Pei-Wen Tzuo, I Chen, Song Huat Yeo, et al. A review on the use of robots in education and young children. *Journal of Educational Technology & Society*, 19(2), 2016.
185. J Gregory Trafton, Laura M Hiatt, Anthony M Harrison, Franklin P Tamborello II, Sangeet S Khemlani, and Alan C Schultz. Act-r/e: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, 2(1):30–55, 2013.
186. Rina Tse and Mark E. Campbell. Human-robot information sharing with structured language generation from probabilistic beliefs. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015*, pages 1242–1248. IEEE, 2015. .

187. Rianne van den Berghe, Josje Verhagen, Ora Oudgenoeg-Paz, Sanne van der Ven, and Paul Leseman. Social robots for language learning: A review. *Review of Educational Research*, page 0034654318821286, 2018.

188. Manuela Veloso, Joydeep Biswas, Brian Coltin, Stephanie Rosenthal, Tom Kollar, Cetin Mericli, Mehdi Samadi, Susana Brandao, and Rodrigo Ventura. Cobots: Collaborative robots servicing multi-floor buildings. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 5446–5447. IEEE, 2012.

189. Manuela M Veloso, Joydeep Biswas, Brian Coltin, and Stephanie Rosenthal. Cobots: Robust symbiotic autonomous mobile service robots. In *IJCAI*, page 4423. Citeseer, 2015.

190. Matt Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. Learning semantic maps from natural language descriptions. In *Proceedings of Robotics: Science and Systems*, Berlin, Germany, June 2013.

191. Matthew R. Walter and Thomas M. Howard. Robotics: Science and systems 2016 workshop on model learning for human-robot communication, 2016. URL `http://www.ece.rochester.edu/projects/rail/mlhrc2016/`.

192. Jacqueline Kory Westlund, Leah Dickens, Sooyeon Jeong, Paul Harris, David DeSteno, and Cynthia Breazeal. A comparison of children learning new words from robots, tablets, & people. In *Proceedings of the 1st international conference on social robots in therapy and education*, 2015.

193. David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. Interpreting multimodal referring expressions in real time. In *International Conference on Robotics and Automation*, 2016.

194. Anna Wierzbicka. *Semantics: Primes and universals: Primes and universals*. Oxford University Press, UK, 1996.

195. Tom Williams, Gordon Briggs, Brad Oosterveld, and Matthias Scheutz. Going beyond command- based instructions: Extending robotic natural language interaction capabilities. In *Proceedings of AAAI*. AAAI, 2015.

196. Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. Situated open world reference resolution for human-robot dialogue. In *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction*, 2016.

197. Terry Winograd. *Procedures as a representation for data in a computer program for understanding natural language*. PhD thesis, Massachusetts Institute of Technology, 1970.

198. Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

199. Hendrik Zender, Geert-Jan M Kruijff, and Ivana Kruijff-Korbayová. Situated resolution and generation of spatial referring expressions for robotic assistants. In *IJCAI*, pages 1604–1609, 2009.

200. L.S. Zettlemoyer and M. Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *Proceedings of 21th Conf. on Uncertainty in Artificial Intelligence (UAI-2005)*, 2005.

201. Matt Zucker, Sungmoon Joo, Michael X Grey, Christopher Rasmussen, Eric Huang, Michael Stilman, and Aaron Bobick. A general-purpose system for teleoperation of the DRC-HUBO humanoid robot. *Journal of Field Robotics*, 32(3):336–351, 2015.