Towards Givenness and Relevance-Theoretic Open World Reference Resolution

Tom Williams
MIRRORLab
Colorado School of Mines
Golden, CO, USA
twilliams@mines.edu

Evan Krause, Bradley Oosterveld, Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University
Medford, MA, USA
{firstname.lastname}@tufts.edu

Abstract—**Robots participating in natural dialogue may need to discuss, reason about, or initiate actions concerning dialogue-referenced entities. To do so, the robot must first identify or create new representations for those entities, a capability known as reference resolution. We previously presented GH-POWER: an algorithm that used a Givenness Hierarchy theoretic approach to resolving definite, indefinite, anaphoric, and deictic noun phrases in uncertain and open worlds. In this work, we introduce GROWLER: a new reference resolution algorithm which enables more robust reference resolution by extending GH-POWER with a model of *relevance*, and discuss how this extension is able to handle some cases not handled by our original algorithm.**

I. Introduction and Motivation

The ability for robots to engage in natural language dialogue with human interlocutors is crucial for many domains of interest to the field of Human-Robot Interaction [43], such as eldercare robotics, education robotics, space robotics, and urban search-and-rescue robotics. Two crucial facets of natural language capability are the twin tasks of *understanding* and *generating* referring expressions. Consider, for example, a natural-language-capable robot wheelchair operating in an eldercare facility. If the wheelchair's user says "Bring me to the recreation room", the robot must be able to understand what location is being referred to by "the recreation room". If the robot knows of multiple recreation rooms, it may need to generate referring expressions that describe those different rooms, in order to ask "Do you mean the recreation room near the garden, or the recreation room near the dining hall?"

The first problem, known as *language grounding*, can be broken into two subtasks: *reference resolution*, in which referring expressions are associated with symbolic representations, and *symbol grounding*, in which symbolic representations are associated with continuously represented percepts [55]. In our work, we are particularly interested in *domain-independent open-world reference resolution*: the association of referring expressions with symbolic representations under the relaxation of two common assumptions: (1) candidate referents (i.e., the "real world entities" referred to in natural language) are not assumed to be known at resolution time, and thus new symbols may need to be *hypothesized*; and (2) candidate referents are not assumed to be drawn from a single domain, and thus may be distributed across multiple heterogeneous knowledge bases.

The relaxation of these assumptions is particularly important for realistic human-robot interaction scenarios. Imagine the following command to an urban search-and rescue robot:

(1) The east wing needs to be evacuated. Please tell that to all personnel.

In order to handle these utterances, the robot needs a *domain-independent* reference resolution algorithm that can resolve references to both *locations* (e.g., "the east wing") and *people* (e.g., "all personnel"). Furthermore, the robot needs an *open world* reference resolution algorithm, because the robot should be able to reason about and carry out the supervisor's request even if it did not previously know that the building being discussed had an "east wing".

In previous work, we developed GH-POWER (Givenness Hierarchy-theoretic Probabilistic Open World Entity Resolution) [57, 55]: a domain-independent open-world reference resolution algorithm which made use of the *Givenness Hierarchy* (GH) [22] framework of reference. GH-POWER was able to resolve the majority of references found in a corpus of human-robot and human-human dialogues, including anaphoric and deictic expressions. However, as we will later discuss, there are some cases in which GH-POWER makes counterintuitive decisions due to how it rules out referential candidates. In this paper, we present a new reference resolution algorithm, GROWLER, which resolves these issues using a model of *relevance*.

In the next section (Section II-A), we provide an overview of previous reference resolution work in robotics that has *not* taken a GH-theoretic approach. In Section II-C, we then provide an overview of the GH. In Section II-D we then discuss previous Gh-theoretic approaches, including GH-POWER. In Section III, we present the GROWLER reference resolution algorithm. In Section IV, we present a proof-of-concept demonstration of GROWLER, and demonstrate how it avoids the occasional counterintuitive decisions made by GH-POWER. Finally, in Section VI we conclude with a discussion of possible directions for future work.

II. Related Work

In this section we will discuss previous approaches to reference resolution in robotics. In Section II-A we provide a broad overview of previous *non-GH-theoretic* approaches. In Section II-B, we discuss concerns related to the problem of reference resolution, such as ability to handle anaphoric and deictic references, domain independence, and the ability to operate in uncertain and open worlds; we then provide a deeper analysis of these previous approaches with respect to those concerns. In Section II-C we provide an overview of the *Givenness Hierarchy* (GH). In Section II-D we discuss previous approaches to reference resolution (including our own) which have taken a GH-theoretic approach.

A. Previous Approaches

While there has been significant work on open-world *directive grounding* [33, 32], in which utterances are translated directly into *action sequences* (thus bypassing the need to ground constituent noun phrases) these has been relatively little work in open-world *reference resolution*. In this section we will discuss both closed-world and open-world approaches.

Work on reference resolution in robotics can be traced back to Terry Winograd's SHRDLU system [58], in which a simulated robot used a *procedural semantics* approach to natural language understanding in order to carry out commands in a simple environment. Under this approach, words were associated with short procedures, such as searching through objects in the scene, which were executed when those words were encountered. This approach inspired several modern models of reference resolution, especially those presented by Gorniak and Roy [19, 38] and Kruijff et al. [29].

Other researchers have also taken a "knowledge-based approach" in which properties are assessed based on the information stored in a centralized knowledge base. For example, Lemaignan uses a semantic parser to translate utterances into lists of RDF triples [26]; for example, "the yellow banana" is translated into {((?obj type banana) (?obj hasColor yellow))}. These triples can then be used to query a central knowledge base populated by input from perception systems, thus producing the set of entities in that knowledge base that satisfy the conjunction of triples [30].

Similarly, Zender et al., who focus on reference resolution in the domain of large-scale topological spaces such as rooms and hallways (as opposed to the domain of objects used by the previous approaches), parse utterances into SPARQL queries [37] (a particular form of RDF query) [59]. This approach also differs from the approach used by Lamaignan through the use of a dedicated *co-reference resolution* step, which attempts to add the references found in an utterance to clusters of references found in past utterances – a step which results in resolution of some anaphoric expressions.

Meyer uses tightly coupled co-reference resolution and reference resolution algorithms to jointly resolve anaphoric and non-anaphoric references [35]. The reference resolution algorithm used by Meyer uses a Markov Logic Network whose weights are learned based on the connections between lexical items and the taxonomic classes of possible referents.

Chai et al. also use a co-reference resolution pre-processing step. After this step, Chai et al. use incoming utterances and perceived deictic gestures to build up a graph representing the relations between the entities mentioned in conversation, and perform reference resolution by finding the best partial match between this graph to a similar graph that represents the relations between entities observed in the world [5, 14, 31, 6].

A different approach is taken by Fasola and Mataric [15], through their work on semantic fields. Fasola and Mataric use a simple reference resolution procedure in which a knowledge base of labels is checked when particular nouns are used – their approach is interesting, however in how they process relations. When a noun is ambiguous, if that noun is a constituent of a prepositional phrase it is disambiguated using a semantic field: a data-driven model of the preposition that produces a probability distribution over coordinates in the environment; the referent whose location has the highest probability value according to this distribution is selected as the referent.

In our own previous (non GH-theoretic) work [53, 54], we presented a probabilistic approach to open world reference resolution where natural language is parsed into a set of logical formulae that are used to guide a best-first search through the space of possible assignments from known entities to variables occurring in those formulae. This approach was later integrated into the GH-theoretic framework we discuss below.

A probabilistic approach is also taken by a number of Bayesian modelers. Kennington and Schlangen present an incremental Bayesian model in which each word is used to modulate the probability of reference for each entity in a scene [25]. Similarly, Tellex and Kollar's Generalized Grounding Graph (G^3) approach uses utterances (after a co-reference resolution pre-processing step) to instantiate probabilistic graphical models that are used to resolve references [47, 48]. This approach has been extended by Tellex and Kollar's colleagues through the Hierarchical Distributed Correspondance Graph approach, which differs in that it uses the "type" associated with each observed noun to restrict the set of possible values associated with each noun-node in the resulting graphical model [8] (see also [49]).

Finally, similar to all three of these approaches, Matuszek et al. present an approach in which utterances are parsed into lambda expressions associated with visual classifiers used to identify objects (which return confidence values that given objects satisfy those expressions).

B. Other Concerns

Each of the approaches mentioned in the previous section addresses, at the least, the *classic reference resolution problem* (cp. the *classic REG problem*[50]): given a definite description, a set of candidate referents from a common domain, and a set of properties held by each of those referents, determine the candidate referent associated with each entity mentioned in the definite description.

But solutions to this "classic" problem framing are not sufficient for robots operating in realistic human-robot interaction scenarios. First, robots cannot assume that referring expressions will always come in the form of definite descriptions: interlocutors may use *anaphoric* expressions (e.g., "it") that reference entities previously mentioned in dialogue; or they may use *deictic expressions* (e.g., "this") that reference entities based on their joint situated perspective with the robot. Second, robots cannot assume that candidate referents will be drawn from a single domain; interlocutors may refer in a single utterance to some combination of locations, objects, people, utterances, ideas, actions, and so on. Third, robots cannot assume that candidate referents will even be known *a priori*; interlocutors may refer to entities that were previously unknown to the robot. Finally, robots cannot assume that they will have perfect knowledge regarding the properties of objects: they may only have confidence *to some extent* that a certain property or relation holds for a certain object or set of objects. In this section, we will analyze the previous approaches and assess the extent to which they address each of these four additional concerns.

1) Anaphoric and Deictic Reference: Many of the discussed approaches handle anaphoric reference to at least a limited extent. Winograd associated anaphoric expressions such as "it" with special procedures that gave preference to elements considered to be "in focus" [58] (see also [36]); a simpler procedure is used by Gorniak and Roy [19]. Kruijff et al. also select items based on focus when "this" is used, and use occurrences of "it" to constrain search to the domain of objects [29]. Lemaignan et al.[30] and Fasola and Mataric [16] both handle anaphora by replacing anaphoric references with the last entity in the dialogue history that matches the animacy and gender constraints imposed by that referent. As previously discussed, a number of previous approaches (e.g., [59, 35, 6, 48]) handle anaphora through dedicated co-reference resolution pre-processing stages. Kennington and Schlangen handle anaphora by attributing a special property to entities selected in dialogue, and then statistically associating pronouns with that special property through training [25].

Few approaches handle deictic references. Kruijff et al. use deictic references to impose *preference orderings* over candidate referents [29]. Lemaignan et al. resolve deictic references to the last entity in the dialogue history that was the focus of simultaneous eye gaze and gesture [30]. Chai et al. incorporate gestural information into their dialogue graph structures [5]. Kennington and Schalngen [25] and Matuszek et al. [34] handle deixis and gaze by combining the probability of reference given an utterance with the probability of reference given gaze and the probability of reference given gesture.

2) Domain Independence: The majority of the examined approaches are dependent on a particular domain. The majority of these approaches were designed to operate only on visible objects [58, 19, 30, 6, 15, 25, 34], while others operate in the domain of large-scale topological locations [59, 56].

Meyer appears to consider objects and units of time, with entities from both domains stored in a single, centralized knowledge base [35]. Similarly, Kruijff et al.'s approach understands references to both objects and small-scale locations (i.e., local points in space), with information from both domains stored in a single, centralized knowledge base (but informed by a set of independent sensory systems) [29].

The approaches of Tellex et al. [47] and ourselves [53] make steps forward with respect to these previous approaches. The approach presented by Tellex et al. is not hand tailored to a particular domain, but appears to handle references to entities from whatever dataset it is trained on – so long as they are physically extant and can be grounded to coordinates in Cartesian space. We expect that this assumption is also true of the work presented by Chung et al. [8]. Similarly, our own previous approach uses a domain-independent framework into which domain-dependent algorithms can be used as "consultants" [52]. Their approach does not make any assumptions about the physical existence or nonexistence of candidate entities.

3) Operation in Uncertain Worlds: It is important to note that the Bayesian approaches do not handle uncertainty in the way we describe: the approaches presented by Kennington and Schalngen, Tellex and Kollar, and Chung et al. represent uncertainty with respect to the relationship between words and features, but not the uncertainty in whether certain entities have certain features. And in fact, representing this uncertainty would undermine the features of some of these algorithms. Chung et al., for example, use entity type to restrict the values considered for each noun-node in the models instantiated by their approach. This approach would need modification if there was uncertainty as to an entity's type.

While the Semantic Fields approach does not appear able to handle uncertain properties, it does handle uncertain spatial relations [15]. Finally, Fang et al. describes how Chai et al.'s approach handles uncertain properties by incorporating an *extent of compatibility* measure into their graph-matching scoring functions [14] ; the approach taken by Matuszek et al. is able to represent the uncertainty in the properties of the objects it reasons about, based on classifier confidences [34]; and our own framework is specifically designed to use domain-specific "consultants" that provide probability values of just this sort [53, 54].

4) Operation in Open Worlds: Of the previous approaches, only two begin to address operation in open worlds. Recent work from Duvallet et al. in the G^3 framework allows a robot to handle references to previously unknown objects described in relation to previously known objects [11]. This approach is limited, however, to spatially situated objects: the pose of the new object is sampled with respect to the other object according to a learned distribution. Our own previously presented framework is also able to hypothesize new entities, but is domain independent in nature, and thus does not have this limitation [53, 54].

5) Discussion: We have argued that a robot operating in natural human-robot interaction scenarios must use a domain-independent reference resolution algorithm capable of handling not only definite descriptions, but also anaphoric and deictic expressions, and must do so in both uncertain and open worlds. our previously presented framework [53, 54], known as POWER (Probabilistic Open World Entity Resolution), makes progress towards this as a domain-independent framework designed to operate in uncertain and open worlds. But this framework falls short in that it is not able to handle anaphora or deictic expressions. In order to extend POWER to handle this wider variety of referring expressions, we have turned to a linguistic framework known as the *Givenness Hierarchy* (GH).

C. The Givenness Hierarchy

As shown in Table I, the GH [22] is comprised of six hierarchically nested tiers of cognitive status, where information with one cognitive status can be inferred to also have all lower statuses. For example, a piece of information that is *activated* is also *familiar*, *{*uniquely identifiable}, *referential*, and *type identifiable*. Each level of the GH is "cued" by a set of linguistic forms, as seen in Table I. For example, the second row of the table shows that the definite use of "this" can be used to infer that the speaker assumes the referent to be at least activated to their interlocutor.

TABLE I
COGNITIVE STATUS AND FORM IN THE GH

Cognitive Status	Mnemonic Status	Form
In focus	in the FOA	*it*
Activated	in STM	*this,that,this* N
Familiar	in LTM	*that* N
Uniquely identifiable	in LTM *or new*	*the* N
Referential	new or hypothetical	indefinite *this* N
Type identifiable	new or hypothetical	*a* N

The GH presents an attractive framework for computational research for several reasons. First, it provides a clear mapping between linguistic form and cognitive status. But second, unlike other frameworks (c.f., e.g., Ariel's Accessibility Theory [1, 2]), the GH focuses on *means of access* rather than *salience*, and thus each status in the hierarchy evokes distinct actions taken to search or modify a discrete set of cognitive structures. Specifically, the "In Focus", "Activated" and "Familiar / Uniquely Identifiable" statuses suggest searching specific structures: the Focus of Attention (FOA), Short-Term Memory (STM), and Long-Term Memory (LTM). On the other hand, "Referential" and "Type identifiable" evoke the concept of *hypothesization*, by which these structures are modified by the creation and insertion of a *new* representation. We call the actions evoked by each tier the *mnemonic status* of the tier, as shown in the second column of Table I.

In order to determine the cognitive status ascribed to a piece of information, one can use rules based on the Coding Protocol provided by Gundel et al. [23]. To determine the most restrictive status a referent can be assumed to have (given the pronoun used to refer to it) one can consult a table such as Table I which associates linguistic forms with cognitive statuses: when "this N" is used, the *most restrictive* status is *activated* – lower statuses can also be inferred, but are less restrictive; it is possible that the referent is *in focus*, but this cannot be inferred, and is less likely as if the referent *were* in focus, the speaker could have used a more restrictive pronoun such as "it" to refer to the referent. The GH can also be used for reference resolution and referring expression generation, but does not allow these tasks to be solved *automatically*. In the next section, we will discuss previous approaches to use the GH to facilitate reference resolution.

D. GH-Theoretic Approaches to Reference Resolution

We now turn to previous GH-theoretic approaches to reference resolution. While we are not the first to draw inspiration from the GH, there have been few others. Specifically, we are aware of only three previous GH-theoretic approaches to reference resolution.

The first two such approaches are the partial GH implementations presented by Kehler [24] and Chai [4]. Of the two approaches, Chai et al.'s is the more extensive, and we direct the author to their paper, as well as our critiques of Kehler's approach [57]. We refer to both approaches as "partial implementations" as they do not attempt to handle all tiers of the GH. Chai et al., for example, use reduced four-tier hierarchy which hierarchy combines the GH's *in focus* and *activated* tiers into a single "Focus" tier, and combines the GH's *familiar* and *uniquely identifiable* tiers into a single "Visible" tier. Chai et al. also include a new top-most tier devoted entirely to Gestured-towards entities, and include a bottom-most tier "Others", which nominally combines the GH's *referential* and *type identifiable tiers*, but does not actually appear to be used.

In our previous work [57] we show that while Chai et al.'s approach may be sufficient for the multi-modal user interfaces for which it was designed, it cannot satisfactorily address aspects of reference resolution that concern the domain of *human-robot interaction*. Like the majority of non-GH-theoretic reference resolution approaches we discussed in the previous section, Chai et al.'s approach does not handle uncertain or open worlds and is restricted to the domain of objects. Furthermore, the modified hierarchy used by Chai et al. necessarily prevents important categories of linguistic forms from being differentiated between, and does not account for the GH's preference for lower tiers over higher tiers. Finally, the algorithm presented by Chai et al. that makes use of the modified hierarchy is greedy in nature; as we have discussed in previous work, however, that this may make the algorithm prone to errors when resolving references, and show how the way the algorithm is employed negates some computational benefits of the greedy approach.

In order to address these concerns, we presented the GH-POWER algorithm [57]: an extension of the previously discussed POWER algorithm which uses the GH to handle a wider class of expressions, including anaphoric and deictic expressions, as well as to increase computational efficiency. Because GROWLER (the algorithm we present in this paper) builds directly off of GH-POWER, we will now describe its design in some detail.

E. The GH-POWER Algorithm

As previously discussed, the first four tiers of the GH evoke a hierarchically nested four-tiered memory structure, comprised of the *Focus of Attention* (FOA), *Short-Term Memory* (STM), *Discourse Context* (DC) and *Long Term Memory*

(LTM). Our memory model uses just such a memory structure, wherein the first three tiers contain memory traces to representations stored in LTM, where LTM itself is a set of distributed heterogeneous knowledge bases managed through our Consultant Framework.

Because the GH does not provide guidelines for how cognitive structures are chosen for selection during reference resolution, we presented a set of “Search Plans” that specify which of these tiered memory structures to search through when different linguistic forms are used, and when to hypothesize new representations instead of searching for existing ones. For example, “activated"-cueing forms, such as “this N”, are associated with the search plan $STM \rightarrow FOA$; because STM is “preferred” to the FOA when an activated-cuing form is used, GH-POWER first searches through the STM (ignoring members that are also in the FOA) for a memory trace to an entity that matches the properties used to describe N, and if a sufficiently probable candidate cannot be found, the FOA is searched. The rationale for each of these strategies is described in our previous work [55]. Overall, this strategy serves to significantly increase the efficiency of reference resolution, as only a small set of known entities (i.e., those activated or in focus) will need to be considered in most circumstances.

In order to handle multiply-referring expressions such as “the green block that is on the blue block,” GH-POWER chooses an appropriate search-plan for each sub-expression, and uses these search plans to create a table containing each combination of search plan steps – this table is iterated through until exhausted or until the combination of search steps indicated by a particular table row (e.g., “the green block” $\rightarrow$ STM, “the blue block” $\rightarrow$ LTM) produces satisfactory referents for all subexpressions. Details of this process are described in more depth in our previous work [55].

Because the GH does not provide guidelines for how candidates are selected from within particular cognitive structures during reference resolution, we previously proposed a set of our own guidelines[57]. Specifically, we suggested that the FOA and STM be sorted according to some scoring function combining linguistic salience, visual salience, eye gaze, and gesture, and that the DC be sorted chronologically; in either case, the first sufficiently probable candidate (as assessed by the POWER framework) according to the imposed ordering would be selected as the correct referent.

We have previously demonstrated [57] that GH-POWER correctly resolved the majority of referring expressions found in a corpus of human-robot and human-human team tasks, even without considering eye gaze and gestural information, and that it was able to capture a variety of aspects of the GH that were not captured by previous GH-theoretic approaches. For example, Gundel et al. [21] present the following example:

(2) a. Alice: I failed my linguistics course.
b. Bob: Can you repeat that?

Here, “that” could either refer to the linguistics course, which should be in Alice’s FOA, or to the utterance that she just uttered, which should at least be “activated” to Alice. Gundel et al. [21] argue that it’s more likely that Bob is referring to the utterance itself, as otherwise Bob could have used “it” rather than “that” to refer to the course. GH-POWER captures this preference by checking STM before checking the FOA; since the utterance is a satisfactory match and is contained in STM, it is automatically chosen, and the contents of the FOA need not even be examined.

F. Discussion

While GH-POWER represented an advance over previous reference resolution algorithms, we have identified several reasons why further improvement is needed [55]. For example consider the following scenario, in which a human Bob instructs a robot subordinate.

(3) *Scene: A table on which sits a red box and a white box*
a. Bob: "Look at the white box"
b. Bob: "Pick that up"

Figure 1 shows the contents of the robot’s GH-theoretic data structure after hearing the first of the two commands. At this point, the white box should be in the robot’s FOA and the red box should be in its STM. When the robot hears Bob’s second command, what should it resolve “that” to? GH-POWER first considers the contents of STM, as if Bob had meant to refer to something in the FOA he could have used “it” instead of “that”. Because a suitable candidate (the red box) is in STM, it will be selected. But while “it” may indeed have been a better choice of wording to use in this scenario, choosing the red box as the referent of “that” is clearly incorrect.

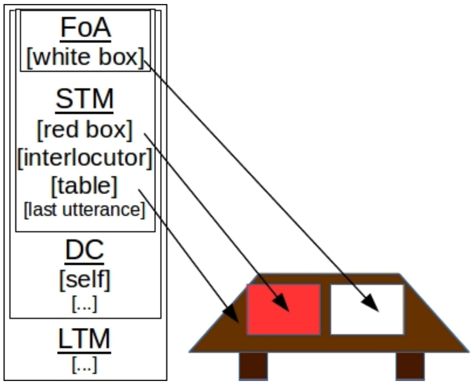

Fig. 1. Contents of GH-POWER’s hierarchical data structures during hypothetical algorithm run.

As discussed in earlier work [55], GH-POWER likely errs for two reasons: (1) it treats hierarchical preferences as *absolute*, and removes dispreferred candidates entirely from consideration; and (2) GH-POWER does not take salience or conversational relevance into account when choosing what to rule out, but instead only when choosing what to select.

GH-POWER operates by checking whether each candidate in a tier is sufficiently probable, only moving on to consider entities in another tier if no sufficiently probable candidate can be found. In this case, this behavior produces an incorrect decision: GH-POWER should accept the red box as sufficiently *probable*, but given the previous sentence, perhaps should not consider it sufficiently *relevant* to the conversation to immediately stop the reference resolution process.

We thus previously suggested [55] that a successor algorithm to GH-POWER should consider *suitability* (i.e., the agent's certainty that a candidate holds all described properties, and, if the candidate is an argument to a verb, the agent's certainty that the candidate is a reasonable argument to the verb) should be used when deciding whether to *retain* a referential candidate, and should consider *relevance* (i.e., the agent's certainty that reference to a candidate would not violate, e.g., Grice's Maxim of Relevance [20]) when deciding whether or not to *extend* search to a new tier.

In addition to these concerns, we point out that there is another reason to extend the GH-POWER algorithm. As it stands, the GH-POWER uses salience to sort the entities contained in the FOA and STM. This ensures that the entity chosen as the target referent is the most salient of the sufficiently probable candidates. But this is only the case when a single candidate referent is selected: when there are multiple candidate referents, it may be more appropriate for a robot to ask for clarification rather than simply select the most salient option. If a robot considers all such sufficiently plausible referents in this way, then the effect of sorting data structures by salience is eliminated. Using salience as part of the relevance metric used to decide whether or not to extend search to additional tiers would allow salience to retain an effect even when all sufficiently probable candidates (rather than simply the single most salient) are considered in subsequent processing steps. For this reason, as well as the other concerns listed above, we have developed a new reference resolution algorithm that fulfills the design suggestions laid out in our previous work [55].

III. THE GROWLER ALGORITHM

In this section we present *Givenness- and Relevance-theoretic Open WorLd Entity Resolution*, or GROWLER: a new GH-theoretic reference resolution algorithm that seeks to address the concerns discussed in the previous section.

While GH-POWER proceeded through a list of variable-tier combinations until a sufficiently probable solution was found, GROWLER takes a different approach, as shown in Algorithm 1. GROWLER takes four arguments: (1) a set of logical formulae S encoding the surface semantics of an utterance (excepting the predicate associated with the verb), (2) a set of "status cue mappings" M associating each variable found in S to its presumed cognitive status, (3) GH: the hierarchical data structure comprised of the FOA, STM and DC, and (4) POWER, a reference resolution algorithm for querying LTM.

Given these arguments, GROWLER associates with each variable v: (1) a sequence $\Theta(v)$ of data structures to search or actions (i.e., hypothesization) to perform (Line 3), and (2) an (initially empty) list $C(v)$ of *candidate referents* (Line 4).

Algorithm 1 GROWLER (S, M, GH)

```
 1: $S$: set of formulae, $M$: set of status cue mappings, $GH$:
    FOA, STM, and FAM data structures
 2: $V = [v | v \in vars(S)]$
 3: $\Theta = create\_plan\_maps(M, GH)$
 4: $C = create\_candidate\_maps(V)$
 5: for all $v \in V$ do
 6:     while $(\nexists c \in C(v) \mid R(c) > \bar{R}) \land (\Theta(v) \neq \emptyset)$ do
 7:         $grow(S, C(v), \Theta(v))$
 8:     end while
 9: end for
10: $Q = populate\_hypothesis\_queue(C)$
11: $R = [v \in V | helpable\_variables(v, Q)]$
12: while $R \neq \emptyset$ do
13:     for all $v \in R$ do
14:         $C' = (C \setminus C(v)) \cup grow(S, C(v), \Theta(v))$
15:         $Q = Q \cup populate\_hypothesis\_queue(C')$
16:         $C(v) = C(v) \cup C'(v)$
17:     end for
18:     $R = [v \in V | helpable\_variables(v, Q)]$
19: end while
20: $Q = assess\_LTM(S, Q)$
21: $Q = assert\_LTM(S, Q)$
22: return $relevantPrefix(Q)$
```

Algorithm 2 $grow(S, C(v), \Theta(v))$

```
1: for all $e \in domain(head(\Theta(v)))$ do
2:     $P(e) = POWER.ASSESS(e, S)$
3:     if $P(e) > \bar{P}$ then
4:         $C(v) = C(v) \cup \langle e, P(e), R(e) \rangle$
5:     end if
6: end for
7: $pop(\Theta(v))$
8: return $C(v)$
```

GROWLER then finds a set of candidate referents to associate with each variable v that sufficiently satisfy the unary predicates in S that involve v (Lines 5–9). This is accomplished as follows: for each variable, GROWLER considers each mnemonic action in $\Theta(v)$ until the perusal of a structure reveals a sufficiently *relevant* candidate (i.e., for which its relevance, $R(c)$, is greater than some relevance threshold, $\bar{R}$ (Line 6)), or until it runs out of mnemonic actions to take (other than LTM queries and hypothesization, which are saved until the end of the resolution process). This process makes use of Algorithm 2, which removes from consideration all insufficiently *probable* candidates (i.e., candidates for which probability $P(e)$ is not greater than some probability threshold $\bar{P}$, as judged by a query of Long-Term Memory using POWER.ASSESS (Algorithm 2, Line 2). Note that insufficiently *relevant* candidates are not removed from consideration, but simply do not suffice to stop the search process.

For each candidate remaining at the end of this process, $C(v)$ will contain an entry $\langle ID, P, R \rangle$, where ID is a unique identifier representing a memory trace allowing access to an entity in LTM, P is the probability that entity ID satisfies unary predicates (i.e., properties) involving v, and R is the *relevance* of entity ID: a combined measure of its visual, linguistic, and conversational salience, which we currently use as an approximation of relevance. In this paper, the salience score used only minimally assesses visual salience; we leave a more robust estimate of visual salience for future work.

GROWLER must now apply the constraints imposed by higher-arity predicates (i.e., relations). Through this process, it may be determined that the existing candidate bindings for a particular variable are not compatible with the bindings to one or more other variables. If this is the case, additional mnemonic actions must be taken in order to find alternate candidate bindings for that variable. We will now describe how this additional assessment and expansion process is handled.

First a hypothesis queue Q is created, where each "hypothesis" added to Q represents a unique combination of the candidate variable bindings found in C for each variable $v \in V$ (Line 10). As part of this step, the probability of satisfaction for individual variables are multipled to calculate the joint probability of each hypothesis, and insufficiently probable hypotheses are pruned out. GROWLER then determines the set of "helpable" variables (Line 11): a variable v is "helpable" if (1) it can still be extended through the `grow` procedure, and if (2) there does not exist a hypothesis in R that contains binds v to a *sufficiently relevant entity*.

As long as there exist variables that can be "helped" in this way, GROWLER tries to "help" each using the following loop: First, GROWLER uses the `grow` algorithm to find additional bindings to the variable v in need of help, and creates $C\prime$: a copy of C in which the set of candidate bindings for v are replaced by this new set of bindings (Line 14). Next, GROWLER updates the hypothesis queue Q with all sufficiently probable combinations of variable bindings that can be created using $C\prime$ in the same way that Q was initialized using C (Line 15). Finally, the new bindings for v found in the first step are added to the full list of sufficiently probable bindings for v stored in $C(v)$ (Line 16).

Finally, previously set-aside mnemonic actions (i.e., LTM queries and/or hypothesization) are executed (Lines 20-21); all remaining combinations of candidate bindings in Q deemed sufficiently relevant *with respect to the most relevant remaining combination of bindings* are returned (Line 22). Currently, we return all bindings that have a relevance score at least half as large as the most relevant candidate; a deeper investigation of possible metrics will be a topic for future work.

IV. Demonstration

In this section, we present a proof-of-concept demonstration of our proposed algorithm, implemented as a component of the ADE [41] implementation of the DIARC architecture [46, 44, 40]. The Distributed Integrated Affect Reflection Cognition (DIARC) Architecture is a component-based architecture that has been under development for over 15 years, and which focuses on enabling robust open-world spoken language understanding. The ADE (Agent Development Environment) middleware in which DIARC is implemented provides a well-validated infrastructure for enabling agent architectures through parallel distributed processing.

For our demonstration scenario, the following architectural components were used: Speech Recognition (using the Sphinx4 Speech Recognizer [51]), Parsing (which uses the most recent iteration [45] of the TLDL Parser [12]), the Dialogue and Pragmatics Components [3, 17], the Goal Manager [39], the Belief Component (which provides a Prolog Knowledge Base [9]), the Resolver Component [53, 54], the GROWLER HyperResolver Component, the Vision Component [27, 28] (which serves as a Consultant), and the PR2 Component, which controls a Willow Garage PR2 [10]. In front of the PR2 is placed a table, on top of which are placed two objects, a mug and a knife. The Vision Component identifies these two objects in the robot's field of view, and records them as `object_0` and `object_1`, respectively.

A human teammate (hereafter "Jim") approaches the robot, and states "Find the knife". This is recognized by the Speech Recognizer, and parsed by the Parser into an utterance of type `INSTRUCT`, with semantics $findObject(self, X)$ and supplemental semantics $knife(X)$. This is then translated by Pragmatics into the intention $want(jim, did(findObject(self, X)))$, i.e., that Jim wants the robot to have achieved the goal of having found X (which can be identified by its supplemental semantics).

Because "the" was used, X is denoted as being assumed by the speaker to be at least *Uniquely Identifiable*. Accordingly, a plan to search through memory for the target object is made by GROWLER (Alg. 1, Line 3): ACT, FOC, FAM, LTM, POSIT. GROWLER next identifies an initial set of candidates that satisfy the unary properties found in the supplemental semantics (i.e., $\{knife(X)\}$). This process begins by GROWLER searching through the set of activated entities, where it finds both objects, which the Vision Component has claimed should be activated because they are the only objects detected in the visual scene. It uses the POWER algorithm [54] to determine that property $knife(X)$ holds for `object_1` (with probability 1.0), but not for `object_0`, accordingly, the hypothesis $X \rightarrow object\_1$ is maintained (Alg. 2, Line 4). However, because this object was not previously referenced in conversation and is not regarded by the Vision Component as being terribly salient, the relevance score for `object_1` is very low (0.06), and accordingly, GROWLER continues its initial expansion, considering FOC and FAM buffers, which, being initially empty, yield no additional candidates. Because there are no higher-arity predicates to consider, GROWLER's job is essentially done, and is able to return the hypothesis $X \rightarrow object\_1$. This is used to create the bound semantics $want(jim, did(findObject(self, object\_1)))$. An action to achieve the indicated goal is submitted to the Goal Manager, Vision reports it is able to find the requested object, and the robot responds "okay."

Next, Jim says "Grab that." This is recognized by the Speech Recognizer, and parsed by the Parser into an utterance of type `INSTRUCT`, with semantics $graspObject(self, X)$ and supplemental semantics $that(X)$.

This is then translated by Pragmatics into the intention $want(jim, did(graspObject(self, X)))$, i.e., that Jim wants the robot to have achieved the goal of having grasped X (which can be identified by its supplemental semantics).

Because “that” was used, X is denoted as being assumed by the speaker to be at least *Familiar*. Accordingly, a plan to search through memory for the target object is made by GROWLER (Alg. 1, Line 3): ACT, FOC, FAM. GROWLER next identifies an initial set of candidates that satisfy the unary properties found in the supplemental semantics(i.e., $\{that(X)\}$). This process begins by GROWLER searching through the set of activated entities, where it finds `object_0` (because `object_1` was prominently mentioned in the previous utterance, it has been promoted to being held in focus). GROWLER then uses the POWER algorithm [54] to (trivially) determine that property $that(X)$ holds for `object_0` (with probability 1.0). Accordingly, the hypothesis $X \rightarrow object\_0$ is maintained (Alg. 2, Line 4). However, because this object was not previously referenced in conversation and is not regarded by the Vision Component as being terribly salient, the relevance score for `object_0` is very low (0.06), and accordingly, GROWLER continues its initial expansion, considering the FOC buffer, which yields the candidate `object_1`. This candidate is also (trivially) found by POWER to hold property $that(X)$. But because `object_1` was mentioned in the main clause of the previous sentence, this factor of linguistic salience is used to yield a higher relevance score (0.31), which GROWLER deems sufficiently high to cease its' search. Because a relevant candidate binding ($X \rightarrow object_1$) was found, it is returned, with the irrelevant hypothesis ($X \rightarrow object_0$) discarded. This is used to create the bound semantics $want(jim, did(graspObject(self, object\_1)))$. An action to achieve the indicated goal is submitted to the Goal Manager, the PR2 Component reports it is able to plan a trajectory to grasp the requested object, the robot responds “okay,” and grasps the object. Note that this utterance would have been incorrectly handled by GH-POWER, as evidenced by the example described in Sec. II-F. A video of this demonstration can be found at https://www.youtube.com/watch?v=E5y7hNwzo3o.

V. Discussion

The performance of GROWLER depends in large part upon a number of factors, each of which require improvement upon in future work. First, GROWLER depends on the accurate calculation of relevance scores. Currently, we use a rudimentary measure of relevance comprised of linearly combined weighted measures of linguistic salience (considering whether an entity has been prominently and recently mentioned) and visual salience (considering whether an entity is present within the current visual context). However, what would be more helpful would be to have a robust measurement of conversational relevance [18] beyond these local factors. Moreover, the weighting factors currently used to calculate relevance, as well as the thresholds for “sufficient relevance” and “sufficient probability” are arbitrarily chosen: it would be more valuable to have weightings and thresholds learned from data.

GROWLER also presents the opportunity to investigate a number of interesting new research questions. For example, it presents the opportunity to investigate the Givenness Hierarchy's account of references accompanying deictic gestures. Consider, for example, the utterance “Pick that up”, accompanied by a gesture towards a nearby object. While these were not addressed algorithmically by GH-POWER, we previously posited that these forms could be resolved using the GH coding protocol [23], which suggests that entities that have just been gestured or gazed at have *activated* status. Extending on this account from the perspective of GROWLER, we might expect that gesture or gaze might increase the relevance of a given object, allowing it to be selected from among other activated entities. Evaluating this account, however, will require the development of models of the impact of gesture and gaze upon relevance, and careful study to tease out the gaze-and-gesture related factors that increase relevance versus those that activate an entity in the first place. We believe that studying such topics will be promising directions for future work.

VI. Conclusion

In this chapter, we began by outlining the *language grounding* problem, and its constituent parts: *reference resolution* and *symbol grounding*. We then described GH-POWER, in which the task of *symbol grounding* is considered the purview of the *distributed heterogeneous knowledge bases* that comprise long term memory, and in which the task of *reference resolution* is performed by a GH-theoretic algorithm that makes use of *consultants* which provide access to those knowledge bases. Next, we discussed some theoretical concerns which provide motivation for future work, and discussed GH-POWER in relation to other approaches to Reference Resolution within robotics.

In addition to the modifications proposed in previous sections, there are a number of directions for future work within our framework. Our algorithm should be extended to handle references to *sets*, and references to *non-discrete* entities (e.g., vague regions of space). We should integrate common-sense affordance-based reasoning capabilities [7] and *incrementalize* and *parallelize* our algorithm, to come in line with psycholinguistic literature [13], similar to previous work from our lab [42] and others [25]. We are also interested in using this approach to *generate* referring expressions in a GH-theoretic manner. And we are interested in more deeply integrating GH-POWER with other components within our architecture (e.g., Vision Processing) so that our within-structure processes can better account for eye gaze and gesture. Finally, and more generally, it is our hope that the framework discussed in this paper will serve as a jumping-off point for much further study of the interaction of language, memory, and attention, not only for algorithmic purposes in the development of integrated systems, but for cognitive modeling purposes as well.

VII. Acknowledgments

This work was in part funded by grant N00014-14-1-0149
from the US Office of Naval Research.

References

[1] Mira Ariel. Referring and accessibility. *Journal of Linguistics*, 24(01):65–87, 1988.

[2] Mira Ariel. Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8:29–87, 2001.

[3] Gordon Briggs and Matthias Scheutz. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of the twenty-seventh AAAI Conference on Artificial Intelligence*, 2013.

[4] Joyce Chai, Zahar Prasov, and Shaolin Qu. Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research*, 27:55–83, 2006.

[5] Joyce Y Chai, Pengyu Hong, and Michelle X Zhou. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of IUI*, 2004.

[6] Joyce Y. Chai, Lanbo She, Rui Fang, Spencer Ottarson, et al. Collaborative effort towards common ground in situated human-robot dialogue. In *HRI*, 2014.

[7] Craig Chambers, Michael Tanenhaus, and James Magnuson. Actions and affordances in syntactic ambiguity resolution. *J. Exp. Psych: Lear., mem., and cog*, 2004.

[8] Istvan Chung, Oron Propp, Matthew R Walter, and Thomas M Howard. On the performance of hierarchical distributed correspondence graphs for efficient symbol grounding of robot instructions. In *IROS*, 2015.

[9] William Clocksin and Christopher S Mellish. *Programming in PROLOG*. Springer Science & Business Media, 2003.

[10] Steve Cousins. Ros on the pr2 [ros topics]. *IEEE Robotics & Automation Magazine*, 17(3):23–25, 2010.

[11] Felix Duvallet, Matthew R Walter, Thomas Howard, Sachithra Hemachandra, et al. Inferring maps and behaviors from natural language instructions. In *ISER*, 2014.

[12] Juraj Dzifcak, Matthias Scheutz, Chitta Baral, and Paul Schermerhorn. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, May 2009.

[13] Kathleen M Eberhard, Michael J Spivey-Knowlton, et al. Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of psycholinguistic research*, 24(6), 1995.

[14] Rui Fang, Changsong Liu, and Joyce Chai. Integrating word acquisition and referential grounding towards physical world interaction. In *Proceedings of ICMI*, 2012.

[15] Juan Fasola and Maja J Matarić. Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots. In *Proceedings of IROS*, 2013.

[16] Juan Fasola and Maja J Matarić. Interpreting instruction sequences in spatial language discourse with pragmatics towards natural human-robot interaction. In *ICRA*, 2014.

[17] Felix Gervits, Gordon Briggs, and Matthias Scheutz. The pragmatic parliament: A framework for socially-appropriate utterance selection in artificial agents. In *39th Annual Meeting of the Cognitive Science Society, London, UK*, 2017.

[18] Rachel Giora. Discourse coherence and theory of relevance: Stumbling blocks in search of a unified theory. *Journal of pragmatics*, 1997.

[19] Peter Gorniak and Deb Roy. Grounded semantic composition for visual scenes. *Journal of AI Research*, 2004.

[20] Herbert P Grice. Logic and conversation. *Syntax and semantics*, 3:41–58, 1970.

[21] Jeanette K Gundel. Reference and accessibility from a givenness hierarchy perspective. *I.R. Pragmatics*, 2010.

[22] Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307, 1993.

[23] Jeanette K Gundel, Nancy Hedberg, Ron Zacharski, Ann Mulkern, et al. Coding protocol for statuses on the givenness hierarchy. unpublished manuscript, May 2006.

[24] Andrew Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *AAAI*, 2000.

[25] Casey Kennington and David Schlangen. A simple generative model of incremental reference resolution for situated dialogue. *Comp. Speech & Language*, 2017.

[26] Graham Klyne and Jeremy Carroll. Resource description framework (RDF): Concepts and abstract syntax, 2006.

[27] Evan Krause, Rehj Cantrell, Ekaterina Potapova, Michael Zillich, and Matthias Scheutz. Incrementally biasing visual search using natural language input. In *Proceedings of the 2013 international conference on autonomous agents and multi-agent systems*, pages 31–38. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

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

[29] Geert-Jan M Kruijff, Pierre Lison, Trevor Benjamin, Henrik Jacobsson, and Nick Hawes. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Symp. Lang.& Robots*, 2007.

[30] Séverin Lemaignan, Raquel Ros, Rachid othersAlami, and Michael Beetz. What are you talking about? grounding dialogue in a perspective-aware robotic architecture. In *Proceedings of RO-MAN*, 2011.

[31] Changsong Liu, Rui Fang, Lanbo She, and Joyce Chai. Modeling collaborative referring for situated referential grounding. In *Proceedings of SIGDIAL*, 2013.

[32] Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. Walk the talk: Connecting language, knowledge, and action in route instructions. In *AAAI*, 2006.

[33] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox. Learning to parse natural language commands to a robot control system. In *Proceedings of ISER*, 2012.

[34] Cynthia Matuszek, Nicholas Fitzgerald, et al. A joint model of language and perception for grounded attribute learning. In *Proceedings of ICML*, 2012.

[35] Florian Meyer. *Grounding Words to Objects: A Joint Model for Co-reference and Entity Resolution Using Markov Logic for Robot Instruction Processing*. PhD thesis, Hamburg University of Technolog, 2013.

[36] Ruslan Mitkov. *Anaphora Resolution: the State of the Art*. University of Wolverhampton, 1999.

[37] Eric Prud'Hommeaux and Andy Seaborne. SPARQL query language for RDF. *W3C recommendation*, 2008.

[38] Deb Roy, Kai-Yuh Hsiao, Nikolaos Mavridis, and Peter Gorniak. Ripley, hand me the cup! (sensorimotor representations for grounding word meaning). In *ASRU*, 2003.

[39] Paul W Schermerhorn and Matthias Scheutz. The utility of affect in the selection of actions and goals under real-world constraints. In *IC-AI*, pages 948–853, 2009.

[40] Paul W Schermerhorn, James F Kramer, Christopher Middendorff, and Matthias Scheutz. Diarc: A testbed for natural human-robot interaction. In *Proceedings of the twentieth AAAI conference on artificial intelligence*, pages 1972–1973, 2006.

[41] Matthias Scheutz. Ade: Steps toward a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence*, 20(2-4):275–304, 2006.

[42] Matthias Scheutz, Kathleen Eberhard, and Virgil Andronache. A real-time robotic model of human reference resolution using visual constraints. *Conn. Sci.*, 2004.

[43] Matthias Scheutz, Paul Schermerhorn, James Kramer, and David Anderson. First steps toward natural human-like HRI. *Autonomous Robots*, 22(4):411–423, May 2007.

[44] Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Tom Williams, and Richard Veale. Novel mechanisms for natural human-robot interactions in the diarc architecture. In *Proceedings of AAAI Workshop on Intelligent Robotic Systems*, 2013.

[45] Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1378–1386. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[46] 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, João S.Sequeira, and Rodrigo Ventura, editors, *Cognitive Architectures*. 2018 (in press).

[47] Stefanie Tellex, Thomas Kollar, et al. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, 32(4):64–76, 2011.

[48] Stefanie Tellex, Pratiksha Thaker, et al. Toward a probabilistic approach to acquiring information from human partners using language. Technical report, MIT, 2012.

[49] Mycal Tucker, Derya Aksaray, Rohan Paul, Gregory J Stein, and Nicholas Roy. Learning unknown groundings for natural language interaction with mobile robots. In *International Symposium on Robotics Research (ISRR)*, 2017.

[50] Kees Van Deemter. *Computational Models of Referring: A Study in Cognitive Science*. MIT Press, 2016.

[51] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. Sphinx-4: A flexible open source framework for speech recognition. Technical report, Sun Microsystems, Inc., Santa Clara, CA, 2004.

[52] Tom Williams. A consultant framework for natural language processing in integrated robot architectures. *IEEE Intelligent Informatics Bulletin*, 2017.

[53] Tom Williams and Matthias Scheutz. POWER: A domain-independent algorithm for probabilistic, open-world entity resolution. In *Proceedings of (IROS)*, 2015.

[54] Tom Williams and Matthias Scheutz. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Proc. AAAI*, 2016.

[55] Tom Williams and Matthias Scheutz. Reference resolution in robotics: A givenness hierarchy theoretic approach. In Jeanette Gundel and Barbara Abbott, editors, *The Oxford Handbook of Reference*. 2018 (Forthcoming).

[56] Tom Williams, Rehj Cantrell, Gordon Briggs, Paul Schermerhorn, and Matthias Scheutz. Grounding natural language references to unvisited and hypothetical locations. In *Proceedings of AAAI*, 2013.

[57] Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. Situated open world reference resolution for human-robot dialogue. In *HRI*, 2016.

[58] Terry Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical report, MIT, 1971.

[59] Hendrik Zender, Geert-Jan M. Kruijff, and Ivana Kruijff-Korbayová. Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of IJCAI*, 2009.