Situated Open World Reference Resolution for Human-Robot Dialogue

Tom Williams
Saurav Acharya
Human-Robot Interaction Laboratory
Tufts University, Medford MA, USA
{williams,sachar01}@cs.tufts.edu

Stephanie Schreitter
Austrian Research Institute
for Artificial Intelligence
Vienna, Austria
stephanie.schreitter@ofai.at

Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University, Medford MA, USA
mscheutz@cs.tufts.edu

***Abstract*—A robot participating in natural dialogue with a human interlocutor 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 algorithms for resolving references occurring in definite noun phrases. In this paper we present GH-POWER: an algorithm for resolving references occurring in a wider array of linguistic forms, by making novel extensions to the *Givenness Hierarchy*, and evaluate GH-POWER on natural task-based human-human and human-robot dialogues.**

***Index Terms*—natural language processing, human-robot interaction**

I. Introduction

A robot participating in natural dialogue with a human interlocutor 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 algorithms for resolving references in definite noun phrases. Those algorithms were designed to handle the open worlds and uncertain contexts commonplace in natural human-robot interaction (HRI) scenarios [21]–[24]. In this paper, we present an open-world reference resolution algorithm which can handle a wider array of linguistic forms by using the *Givenness Hierarchy* (GH) [8], a linguistic framework which associates the form of a referential expression (e.g., pronominal, definite noun phrase, indefinite noun phrase) with a presumed "cognitive status" (e.g., focus of attention, short term memory, long term memory). This significantly advances the state of the art of natural-language based HRI, by (1) increasing the breadth and complexity of referential expressions understandable by robots, (2) allowing robots to understand such expressions in *open* and *uncertain* worlds, and (3) bringing robot natural language understanding closer in line with an established linguistic framework (i.e., the GH). What is more, it is significant in its extension of the GH itself, through the addition of guidelines which clarify how the GH should be computationalized.

The rest of the paper proceeds as follows. In Section II we discuss previous work: the GH, previous implementations thereof, and an explanation of how those implementations might be improved if clear guidelines for using the GH could be crafted. In Section III we suggest such guidelines for the GH and present GH-POWER: an algorithm which uses those guidelines to improve on previous approaches. In Section IV we evaluate GH-POWER using data from an empirical human-human and human-robot experiment. Finally, we discuss directions for future work in Section V and conclude in Section VI.

II. Previous Work

As this work is a GH-theoretic extension of our previous work on open world reference resolution [21]–[24], we will focus primarily in this section on the GH itself and on previous GH-based reference resolution algorithms.

However, this work builds upon a large body of work on human-robot dialogue processing [12], [14], and more specifically, situated reference resolution and language grounding [4], [13], [20]. We thus direct the reader to the works cited above, as well as to our own previous work, for a broader understanding of how our approach relates to previous work.

A. The Givenness Hierarchy

Gundel et al.'s *Givenness Hierarchy* (GH), contains six levels or tiers at which a piece of information may be cognitively accessible [6]. As seen in Fig. 1, these are nested such that if a piece of information attains some level of cognitive accessibility, it also attains all lower tiers. For example, any information that is in focus is also activated (i.e., in short term memory), familiar (e.g., previously referenced in the current dialogue, or culturally salient), can be uniquely identified, can be referred to, and can have its type identified, whereas information that is *at most* familiar is also uniquely identifiable, referential and type identifiable, but not in focus or activated.

In focus ⊂ *Activated* ⊂ *Familiar* ⊂ *Uniquely identifiable* ⊂ *Referential* ⊂ *Type identifiable*

Fig. 1. The Givenness Hierarchy

Each cognitive status in the GH is assumed to be cued by a different set of linguistic forms, as seen in Table I. For example the use of "it" to refer to an entity signifies that the speaker believes that entity to be in her interlocutor's focus of attention (as seen in Row 1), and the use of "that N" (for some noun-phrase N) signifies that the speaker believes the entity to be "familiar" to her interlocutor (i.e., that it is *at least* in her interlocutor's long-term memory).

TABLE I
COGNITIVE STATUS AND FORM IN THE GIVENNESS HIERARCHY

Level	Cognitive Status	Form
In focus	in focus of attention	*it*
Activated	in working memory	*this,that,this* N
Familiar	in LTM	*that* N
Uniquely id-able	in LTM or new	*the* N
Referential	new	indef. *this* N
Type id-able	new or hypothetical	*a* N

While the GH itself does not make claims about how a piece of information might acquire a particular cognitive status, Gundel et al. present a "coding protocol" which suggests a possible set of criteria that might be used to make such a decision [10]. For example, the protocol suggests that the syntactic topic of the immediately preceding sentence should be *in focus*, the sentence's speech act and the targets of concurrent gestures or sustained eye gaze should be at least *activated*, and any entity mentioned previously in the current dialogue should be at least *familiar*.

The GH and its coding protocol thus provide:

1) data structures necessary for reference resolution
2) guidelines for how those data structures are populated
3) guidelines for how those data structures are accessed

These concepts represent a powerful framework for reference resolution with strong experimental justification [7]. It is thus unsurprising that there have been several attempts to use the GH to inform reference resolution algorithms in the fields of Human-Robot and Human-Agent Interaction. We will now describe the two implementations which, until now, have made the most extensive use of the GH.

B. GH-Based Reference Resolution Algorithms

The first implementation of the GH that we will examine is that presented by Kehler et al. [11], in which they propose the modified hierarchy seen in Fig. 2. There, Kehler et al. omit the last two levels of the GH, due to a primary interest in interfaces with which it is unlikely for one to refer to unknown or hypothetical entities. Kehler et al. used their modified hierarchy to craft four rules (presented here verbatim) they found capable of resolving all references they encountered:

1) If the object is gestured to, choose that object
2) Otherwise, if the currently selected object meets all semantic type constraints imposed by the referring expression (i.e., "the museum" requires a museum referent; bare forms such as "it" and "that" are compatible with any object), choose that object.
3) Otherwise, if there is a visible object that is semantically compatible, then choose that object (this happened three times; in each case there was only one suitable object).
4) Otherwise, a full NP (such as a proper name) was used that uniquely identified the referent.

$$\textit{In focus} \subset \textit{Activated} \subset \textit{Familiar} \subset \textit{Uniquely identifiable}$$

Fig. 2. Kehler's Modified Hierarchy

The second GH implementation we will examine, Chai et al. [1], expands on Kehler's approach in two important ways: First, Chai et al.'s implementation can identify and resolve ambiguities (Kehler's first rule is problematic if the target of a gesture is ambiguous, and Kehler's third rule is problematic if a referential expression is ambiguous). Second, Chai et al.'s implementation makes it possible to handle utterances containing multiple referential expressions or gestures. To make these advancements, Chai et al. combine a subset of the GH with Grice's theory of Conversational Implicature [5] to produce the modified hierarchy seen in Fig. 3.

$$\textit{Gesture} \subset \textit{Focus} \subset \textit{Visible} \subset \textit{Others}$$

Fig. 3. Chai's Modified Hierarchy

Chai et al.'s modified hierarchy contains four tiers: (1) "Gesture", containing entities gestured toward (because a gesture *intentionally* singles out entities), (2) "Focus", combining Gundel's *in focus* and *activated* tiers, (3) "Visible", combining Gundel's *familiar* and *uniquely identifiable* tiers, and (4) "Others", combining Gundel's *referential* and *type identifiable* tiers, although this tier does not appear to be used, perhaps due to the lack of hypothetical entities in graphical interfaces.

Chai et al. present a greedy reference resolution algorithm using their hierarchy. This algorithm first assigns a score between each referential expression X in an utterance and each entity N contained in a set of vectors (Gesture, Focus, Visible), calculated by multiplying (1) the probability of selecting N from its vector, (2) the probability of selecting that tier given the form of X, and (3) the "compatibility" between X and N. Compatibility is 1 if N has all properties mentioned in X, is of the type mentioned in X (if any), has the name mentioned in X (if any), and was gestured to when X was uttered (if any), 0 otherwise; it is thus *binary* in nature and cannot account for uncertainty. After scoring all visible entities, the algorithm greedily binds references to entities, moving downward through the hierarchy of vectors. This approach does not address all aspects of reference resolution found in typical human-robot dialogues (nor does any other current approach). There are, in particular, five aspects of human-robot dialogue not captured by this approach.

First, the algorithm assumes complete certainty as to entities' properties. In realistic HRI scenarios, an agent may only be able to say that an entity has a certain property *with some probability*. Furthermore, an agent could be aware that it simply *does not know* whether an entity has a certain property. Second, consider the following command:

"Get **my laptop** from **my office**, and if you see **a charger** bring that too."

The three bolded referential expressions present issues for Chai et al.'s approach. **My laptop** is (presumably) not currently visible, a condition common in many HRI scenarios, but one which cannot currently be handled using Chai et al.'s algorithm. **My office** is also (presumably) not currently visible. And, it is not an object, per se, and cannot be gestured towards in the same way as can be objects or icons. It is unclear whether Chai' et al.'s modified hierarchy could handle references to locations, which are common in many HRI scenarios. **A charger** is also (presumably) not currently visible. And, it is not even known to exist, as it is *hypothetical*. In order to resolve such references, one must assume an *open world* in which new entities may be added through experience or dialogue. While many HRI scenarios are open-world in nature, Chai et al.'s algorithm operates in a closed world.

Third, a robot may need to resolve references to events, speech acts, or other entities that *cannot* physically exist, as seen in Examples 1 and 2. However, Chai et al.'s algorithm cannot handle references to nonexistent entities.

(1) Can you repeat it?

(2) Can you repeat that?

Fourth, because Chai's modified hierarchy combines the first two levels of the GH, Chai et al.'s algorithm cannot distinguish between Examples 1 and 2 even if it *could* handle references to physically nonexistent entities. When Example 1 is used to respond to the utterance "I'm sorry, but I failed to complete the task", "it" unambiguously refers to "the task". However, this is not the case when Example 2 is used. The GH predicts that when a form associated with the *activated* level is used, one should prefer an activated referent (such as a speech act) to an in-focus referent (such as the focus of the previous sentence), because if the speaker had meant to refer to an in-focus entity she could have used an in-focus-cueing form (e.g., "it"). Thus, while Example 2 could refer to either the speech act or failed task, the speech act should be preferred[1].

Fifth, natural human-robot dialogues may contain complex noun phrases such as "Do you see the red block on that blue block?" Because Chai et al. use a greedy algorithm (instead of, e.g., their previous graph matching approach), it may be unable to resolve subsequent referential expressions if the first considered referential expression is incorrectly resolved. Chai et al. argue that this approach is advantageous because it may significant prune the search space. However, their algorithm scores all entities against all referential expressions *before* its greedy approach. In a realistic HRI scenario, this may not be practical, as a robot may know of hundreds of entities. Furthermore, checking whether certain properties hold for all entities may be cost prohibitive. For example, while determining whether a given person is a man may be accomplished by a simple database look-up, determining whether two rooms are across from each other may require more expensive computation. An algorithm which performed such assessments lazily (i.e., only when needed, perhaps as the search space was pruned) could be much more efficient.

C. A Need for GH Usage Guidelines

Thus far, we have described reasons for extending Chai et al.'s modified hierarchy and algorithm. But to make the needed extensions, we must first extend the GH itself: each extension we have discussed thus far can be related to an area for which the GH lacks clear usage guidelines. No existing GH-based approach can handle uncertain information, perhaps because the GH neither specifies how uncertainty is handled nor provide guidelines for how intra-tier ambiguity is resolved.

GH-based approaches must be extended to better resolve multiple referential expressions occurring in the same utterance, in order to avoid incorrect greedy decisions. This is because the GH does not provide guidelines for how multiple related referents are simultaneously resolved.

Chai et al.'s approach cannot handle references to entities that are unknown, hypothetical, intangible or not present. This is the result of Chai et al.'s omission and combination of GH tiers, and their use of a purely top down traversal. This may have been avoided if clear guidelines had existed for traversing the tiers of the GH and for guiding intra-tier search using salience arising from linguistic, visual or gestural factors.

We thus believe that a GH-based reference resolution algorithm for human-robot dialogue requires the following.
Clear guidelines for:

1) Determining the order in which to peruse the tiers of the Givenness Hierarchy that allow gestured-towards or gazed-upon entities to take some degree of precedence.
2) Resolving complex referential expressions.
3) Choosing between candidates found within a given tier.

Assumptions of:

1) Uncertain information (i.e., the *properties* of an entity may not be certain or known)
2) An open world (i.e., the *existence* of an entity may not be certain or known)
3) Global resolution (i.e., a referential expression may refer to an entity which is not currently visible)
4) Domain independence (i.e., a referential expression may refer to *any* entity, regardless of type or tangibility).

III. Proposed Extension to GH-Based Approaches to Reference Resolution

We now present (1) our suggestions for guidelines 1-3, and (2) a reference resolution algorithm which uses those guidelines and which operates under assumptions 1-4.

A. Guidelines for the Givenness Hierarchy

In Section II-C, we noted the lack of clear guidelines for how GH tiers should be traversed, and that both strictly top-down and strictly bottom-up approaches can be problematic. We thus suggest the following set of inter-tier traversal guidelines: referential forms cuing the *in focus* tier prompt a search of *FOC*: the entities *in focus*; referential forms cuing the *activated* tier prompt a search of *ACT*: entities that are *activated* (but not *FOC*) followed by a search of *FOC*; referential forms cuing the *familiar* and *uniquely identifiable*

[1] Gundel et al. have empirically verified that these two hierarchical levels are distinguished between in a wide variety of languages beyond English, including Eegimaa, Kumyk, Ojibwe, and Tunisian Arabic (each of which is genetically and typologically unrelated to the other three.) [7]

tiers follow the same pattern as referential forms cuing the *activated* tier, and *then* search the data structure associated with the tier they cue (i.e., *FAM* or *LTM*). This is in line with [9], in which Gundel et al. argue that for definite noun phrases, referents in one's current perceptual environment are preferred to those found by searching Long Term Memory (LTM). These guidelines are summarized in Table II, which lists search plans for forms cueing the first four levels of the GH.

TABLE II
SEARCH PLANS FOR GH TIERS 1-4

Level	Search Plan
in focus	FOC
activated	ACT → FOC
familiar	ACT → FOC → FAM
uniquely id'able	ACT → FOC → LTM

Unfortunately, Table II does not differentiate between referential expressions of the form *this N* used to cue the *activated* tier (e.g., "Pick up *this spoon*") and those used to cue the *referential* tier (e.g., "This spoon I saw was amazing"). While it may be possible to use factors such as tense to tell when one is using the *referential*-cueing sense, we believe that a first step towards appropriately handling the *referential*-cueing sense would be to treat all uses of *this N* as *activated*-cueing so long as a suitable referent can be found at the *activated* or *in focus* tiers, and otherwise treating such a use as *referential* cueing. A first step might also use a single process to handle *referential* and *type-identifiable* cues, as both lead to the construction of new representations. These suggestions yield Table III. Note that in this table, it is assumed that the form "this N" is always associated with the *referential* tier, which now deals with both the activated and referential senses of this form.

TABLE III
SEARCH PLANS FOR COMPLETE GH

Level	Search Plan
in focus	FOC
activated	ACT → FOC
familiar	ACT → FOC → FAM
uniquely id'able	ACT → FOC → LTM
referential	ACT → FOC → HYP
type id'able	HYP

We previously noted that the GH lacks clear guidelines for choosing between candidates found in a given tier. Note that while the GH coding protocol suggests that the target of sustained eye gaze or gesture should be considered *activated*, in practice it may not be possible to disambiguate precisely which entity is being gazed or gestured towards.

We thus suggest that *all* entities in an interlocutor's field of view be considered *activated*. As *activated* is roughly equivalent to short term memory (STM), this represents the possibility of any entity in the vicinity of an interlocutor's gaze being in her STM. Suppose each entity in *FOC* and *ACT* were assigned a *salience score* calculated by function Γ, using features such as intensity, length, and recency of gaze and gesture, visual salience, and linguistic salience (c.f. [6]). Suppose further that function $\Phi(P, E)$ could determine the joint probability that properties P described entity E.

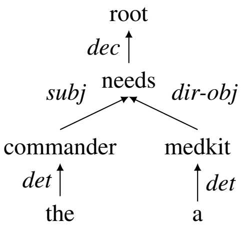

Fig. 4. Example Parser Output

Then, the best candidate c could be chosen from tier T in the following way: (1) sort T in descending order of Γ-score; (2) calculate $\Phi(S, c)$ for each candidate $c \in T$ and property set S until, e.g., a candidate with sufficient probability is found: this will be the candidate with the highest Φ score.

We now use the above guidelines in coordination with the GH to yield GH-POWER: a reference resolution algorithm which extends the state of the art of human-robot dialogue.

B. Algorithm

GH-POWER combines our proposed GH extensions with POWER, a domain-dependent open-world reference resolution algorithm [22], [23]. We will first discuss how utterances are parsed and analyzed, then describe the data structures we use. Finally, we will describe how those data structures are used to resolve references in parsed utterances. All capabilities described in these sections are performed by components of the Distributed, Integrated, Affect, Reflection and Cognition (DIARC) architecture [17], as implemented in the Agent Development Environment (ADE) [15], [16].

1) Parsing: Each utterance is first sent to the C&C parser [3], which uses the Combinatory Categorical Grammar formalism [19] to generate a dependency graph. That graph is converted into a tree such as that seen in Fig. 4, which shows the tree produced for "The commander needs a medkit".

From the structure of this tree one may extract: (1) a set of formulae representing the surface semantics of the utterance, (2) a set of "status cue" mappings for each referenced entity, and (3) the *type* of utterance which was heard. From the tree shown in Fig. 4, for example, one would extract:

1) The set of formulae $\{needs(X, Y) \wedge commander(X) \wedge medkit(Y)\}$.
2) The set of status cue mappings $\{X \rightarrow \text{uniquely id'able}, Y \rightarrow \text{type id'able}\}$.
3) The utterance type "STATEMENT" (indicated by the label "dec" on the arc pointing to the root node).

2) Data Structure Population: GH-POWER uses four data structures: *FOC*, *ACT*, *FAM*, and *LTM*, corresponding with the first four levels of the GH (levels five and six do not have associated structures, as they involve construction of new representations). Here, we describe how these data structures are populated, as summarized in Table IV. Lines marked with a star denote information which is not yet, included in each data structure, representing future work.

Before clause n of some natural language utterance is processed, the contents of *FOC* and *ACT* are reset (*FAM*

TABLE IV
CONTENTS OF RELEVANT DATA STRUCTURES

Level	Contents
FOC	Main clause subject of clause n-1 Syntactic focus of clause n-1 * Event denoted by clause n-1
ACT	* Entities visible in int.'s region of attention All other entities referenced in clause n-1 * Focus of int.'s gesture, if any * Focus of int.'s sustained eye gaze, if any * Speech act associated with clause n-1 * All propositions entailed by clause n-1
FAM	All entities referenced in clause n-1 * The robot's current location
LTM	All declarative memory

is reset after each dialogue, and *LTM* is never reset). *FOC*, *ACT* and *FAM* are then updated using the rules listed in Table IV. Linguistically, this entails placing the main clause subject, syntactic focus, and event denoted by clause n-1 into *FOC* (each of which may be extracted from the syntactic representation of clause n-1), placing the speech act and any propositions entailed by clause n-1 into *ACT*, and placing all entities referenced at all in clause n-1 into both *ACT* and *FAM*. In addition, each location visited by the robot and its interlocutor should be placed into *FAM*, and any entities within the interlocutor's region of attention should be placed into *ACT*. Each data structure is then sorted according to Γ-score. Although the ideal scoring function would account for a variety of extra-linguistic factors, we currently use the function $\Gamma(e) = \alpha_1 * m(e) + \alpha_2 * s(e) + \alpha_3 * r(e)$ where $m(e) \in [0, 1]$ represents whether e is in a main clause, $s(e)$ measures the syntactic prominence of e, $r(e)$ measures the recency of mention of e, and $\alpha_1, \alpha_2, \alpha_3$ are monotonically decreasing coefficients prioritizing the three measures.

3) Reference Resolution: To resolve the references in a given clause, that clause is first viewed as a graph whose vertices and edges are the variables and formulae used to represent the semantics of that clause[2]. This graph is then partitioned into connected components. For each partition, Alg. 1 (GH-POWER) is used to resolve all references found in that partition, producing a set of variable-entity bindings.

GH-POWER takes four parameters: (1) S (the semantics of clause n), (2) M (the status cue mappings for clause n), (3) GH (containing FOC, ACT, and FAM), and (4) POWER (a module for Probabilistic, Open-World Entity Resolution, to interface with LTM, as described in [23]). GH-POWER first collects the variables appearing in S and sorts them with respect to the tier they are cued towards. For example, if $X \to infocus$ and $Y \to familiar$ appear in M, then X will appear before Y (Alg. 1 line 2). GH-POWER then initiates cache-table C which stores a memoized list of variable-to-entity bindings for each combination of variables in V and tiers in {*FOC*, *ACT*, *FAM*, *HYP*} (line 3).

Algorithm 1 GH-POWER($S, GH, POWER$)

```
1: $S$: set of formulae, $M$: set of status cue mappings, $GH$: FOC, ACT, and FAM data structures, $POWER$: a Probabilistic, Open-World Entity Resolver
2: $V = [v|v \in vars(S)]$ sorted by $M(v)$
3: $C = create\_cache\_table(V, \{FOC,ACT,FAM,HYP\})$
4: $\Theta = create\_plan\_table(M)$
5: $H = \emptyset$
6: for all $P \in \Theta$ do
7:     $P_d = [p|p \in P, tier(p) = LTM]$
8:     $V_p$ = new list
9:     for all $p \in (P \setminus P_d)$ do
10:        $(v, t) = (var(p), tier(p))$
11:        if $C[v, t] == \emptyset$ then
12:            if $(t == HYP)$ then
13:                $C[v, t] = \{((v \to "?") \to 1.0)\}$
14:            else
15:                $C[v, t] = ASSESS(S, v, t, POWER)$
16:            end if
17:        end if
18:        $V_p = v \cup V_p$
19:        $H = ASSESS\_ALL(S, V_p, (H \times C[v, t]), POWER)$
20:        if $H == \emptyset$ then
21:            BREAK
22:        end if
23:    end for
24:    if $P_d! = \emptyset$ then
25:        for all $h \in H$ do
26:            $h = resolve(POWER, bind(S, h), order(vars(P_d))))$
27:        end for
28:    end if
29:    $H = [h|h \in H, prob(h) >= \tau_{resolve}]$
30:    if $|H| > 0$ then
31:        BREAK
32:    end if
33: end for
34: if $|H| \neq 1$ then
35:    return $H$ // AMBIGUOUS or UNRESOLVEABLE
36: else
37:    return $assert(POWER, bind(S, H[0]))$
38: end if
```

Algorithm 2 ASSESS($S, V, T, POWER$)

```
1: $S$: set of formulae, $V$: variable of interest, $T$: tier of interest, $POWER$: a Probabilistic, Open-World Entity Resolver
2: $S_v = [s|s \in S, vars(s) = \{V\}]$
3: $H = \emptyset$
4: for all $t \in members(T)$ sorted by $\Gamma(t)$ do
5:     $h = (V \to t) \to \prod_{s \in S_v}$ assess(POWER, bind(s, (V $\to t$)))
6:     if $prob(h) >= \tau_{assess}$ then
8:         $H = H \cup h$
9:     end if
10: end for
11: return $H$
```

Algorithm 3 ASSESS_ALL($S, V, H, POWER$)

```
1: $S$: set of formulae, $V$: variables of interest, $H$: set of hypotheses, $POWER$: a Probabilistic, Open-World Entity Resolver
2: $S_v = [s|s \in S, head(V) \in vars(S), [\exists v \in tail(V)|v \in vars(s)]]$
3: $H' = \emptyset$
4: for all $h \in H$ do
5:     $prob(h) = prob(h) * \prod_{s \in S_v} assess(POWER, bind(s, h))$
6:     if $prob(h) >= \tau_{assess}$ then
7:         $H' = H' \cup h$
8:     end if
9: end for
10: return $H'$
```

Before GH-POWER begins trying different variable-entity

[2] To properly handle declarative and imperative utterances, we omit the formula associated with the main clause verb from consideration. Future work will consider the main clause verb using common-sense reasoning.

assignments, it must determine in which data structures to look for those entities, determined by the *plan* associated with each level of the hierarchy seen in Table III.

(3) The ball in this red box

To handle multi-variable expressions, GH-POWER creates a table Θ, storing all multi-variable plan combinations. For example, if the referential expression seen in Example 3 is parsed as: $\{ball(X) \wedge box(Y) \wedge red(Y) \wedge in(X,Y)\}$ with status cue mappings $\{X \rightarrow$ *{uniquely id'able*, $Y \rightarrow$ *referential*$\}$, then Table V of joint search plans will be created. After Θ is created (line 4), an empty set of candidate hypotheses H is created. GH-POWER then examines Θ one row at a time until a solution is found or the end of the table is reached. For each table entry P, GH-POWER first separates variables for which it must query LTM from all other variables (line 7). It then initializes an empty list V_p to hold variables that have been examined thus far for entry P (line 8). Next, it iterates over each (variable, tier) pair in that row, as we now describe.

TABLE V
Sample Joint Search Plan Table

Y	X
ACT	ACT
ACT	FOC
ACT	LTM
FOC	ACT
FOC	FOC
FOC	LTM
HYP	ACT
HYP	FOC
HYP	LTM

Consider row one of Table V. GH-POWER would first examine the first entry in this row, which says to look for Y's referent in ACT. If C does not already contain hypotheses for $var(p)$ and $tier(p)$ (i.e., Y and ACT), a new one is created: if $tier(p) = HYP$, this hypothesis binds $var(p)$ to "?". Otherwise, GH-POWER uses *ASSESS* to search $tier(p)$ for the most likely entity to assign to $var(p)$ (line 15).

ASSESS takes four parameters: (1) S (the set of formulae), (2) V (the variable of interest), (3) T (the tier in which to look for possible referents for V), and (4) POWER. ASSESS creates, for each entity $t \in T$, a new hypothesis which maps V to t, with probability equal to the product of probabilities of each formula $s \in S$ which only refers to V (Alg. 2 lines 2-6). For example, if Example 3 is heard and there is one entity in ACT (e.g., obj_13), *ASSESS* would consult POWER to see to what degree obj_13 could be considered to be a box, and to what degree it could be considered to be red, and then create a hypothesis mapping Y to obj_13 with probability equal to the product of the two probabilities returned by POWER.

Once all formulae containing only $var(p)$ are examined, all those containing both $var(p)$ *and* any other previously examined variables are examined (line 19) using Alg. 3 (*ASSESS-ALL*). For Example 3, this would involve inquiring to what degree the candidate entities for X could be considered to be "in" each candidate entity for Y. After each variable is considered, all candidate bindings whose likelihoods fall below a certain threshold are removed. If this leaves no hypotheses with probability above τ_assess, GH-POWER breaks out of its loop and considers the next row of the table.

For example, if resolving Y produces hypothesis list

$$\{((Y \rightarrow obj\_13) \rightarrow 0.8), ((Y \rightarrow obj\_12) \rightarrow 0.75)\},$$

and resolving X produces the hypothesis list

$$\{((X \rightarrow obj\_5) \rightarrow 0.9)\},$$

these are combined into:

$$\{((Y \rightarrow obj\_13, X \rightarrow obj\_5) \rightarrow 0.72), \\ ((Y \rightarrow obj\_12, X \rightarrow obj\_5) \rightarrow 0.675)\}.$$

If ASSESS determines that $in(X, Y)$ has probability 0.2 for the first of these hypotheses and 0.9 for the second, the two hypotheses are updated to

$$\{((Y \rightarrow obj\_13, X \rightarrow obj\_5) \rightarrow 0.144), \\ ((Y \rightarrow obj\_12, X \rightarrow obj\_5) \rightarrow 0.6075)\}.$$

If τ_assess is set to 0.6, for example, then the first of these hypotheses would be removed.

GH-POWER now considers all variables set aside to be searched for in LTM. If any such variables exist, GH-POWER considers each candidate binding in H (line 26). For each, S is bound using h's variable bindings, and an ordering of the variables V_h to be queried in LTM is created based on the prepositional attachment observed in S. The bound semantics and variable ordering are then used by the POWER algorithm [23] to determine (1) whether any of the variables in V_h refer to unknown entities, and (2) which entities in LTM are the most probable referents for each other variable in V_h. The set of hypotheses H is then updated using these results.

Finally, once a solution is found or all table rows are exhausted, the *number* of remaining hypotheses is examined. If more or less than one hypothesis was found, GH-POWER returns the set of solutions. This signifies that the referential expression was either ambiguous or unresolvable. If only one hypothesis remains, GH-POWER uses that hypothesis' variable bindings to update the set of semantics S, and then uses POWER to assert a new representation for each variable bound to "?" (line 38). For example, if resolving Example 3 produces a single hypothesis with probability 0.7 in which X is bound to obj_4 and Y is bound to "?", POWER will create a new object (perhaps with identifier 5) with properties $\{box(obj\_5), red(obj\_5), in(obj\_4, obj\_5)\}$ and return $\{((Y \rightarrow obj_5, X \rightarrow obj_4) \rightarrow 0.7)\}$. Once all partitions have been processed in this way, the results are combined into a comprehensive set of candidate binding hypotheses.

IV. Validation and Evaluation

In this section, we verify that the proposed algorithm and GH extensions do indeed improve on previous approaches, and then perform an experimental evaluation on real-world human-human and human-robot dialogues collected by Schreitter et al. [18]. In those dialogues, human *instructors* demonstrated to human or robot *listeners* how to connect two sections of tubing and then affix the tubing to a box.

A. Validation

We first evaluated several test cases within the previously described experimental context, to demonstrate the success of GH-POWER in addressing our concerns with previous GH-based approaches to reference resolution:. In each case, the algorithm was provided with a knowledge base containing information about the robot's environmental and task context (possibly modified according to that case), and was incrementally fed the relevant utterances for that case.

(1) Previous approaches could not handle *uncertainty*. We confirmed that when the robot believed there was 70% probability that one tube could be referred to as flexible, and 40% probability that the other tube could be referred to as flexible, GH-POWER resolved "The flexible tube" to the first tube. (2) Previous approaches could not handle *open worlds*. We confirmed that when the robot only knew of red and yellow markers, GH-POWER posited a new entity when resolving "Find the blue marker." (3) Previous approaches could not handle references to hypothetical entities. We confirmed that when the robot knew of a box on a table in front of it and was asked to resolve "Imagine a box." and "Describe the box", "the box" was resolved to the imaginary box. (4) Previous approaches could not resolve references to unobservable entities. We confirmed that when the robot believed it was learning a task, GH-POWER correctly resolved "the task" in "describe the task". (5) Previous approaches have had trouble resolving *complex noun phrases*. We confirmed that when a tube on a triangular table was in "familiar" and a tube on a round table was in "activated", GH-POWER correctly resolved "the tube" in "Pick up the tube that is on the triangular table".

B. Evaluation

In addition to validating that GH-POWER significantly extended the set of cases handled compared to previous algorithms, we evaluated it on the corpus of human-human and human-robot dialogues collected by Schreitter et al. As participants' utterances in that experiment were originally in German, these were first translated to English. As we are not currently attempting to handle disfluencies, these utterances were then "cleaned up", removing disfluencies and parenthetical statements. For example, an utterance with word-for-word translation "So then put you the grasp you here at the marker at the red and yellow one" was "cleaned up" to "So then you grasp here at the red and yellow marker."

A knowledge base containing the relevant properties of the 16 objects and agents involved in the task was constructed and provided to GH-POWER. Then, each task-relevant utterance (excepting, e.g., "Hello.") was provided to GH-POWER in sequence, and the results of resolution were compared against "gold standard" resolution results provided by human annotators. The human-robot corpus contained 32 task-relevant utterances, the human-human corpus contained 110.

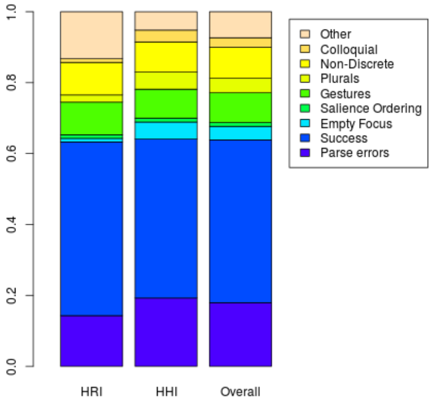

Fig. 5. Reference Resolution Results

Overall, GH-POWER correctly resolved 48 of the 98 (48.98%) references found by the C&C parser in the human-robot dialogues (HRDs), and 121 of the 270 (44.81%) in the human-human dialogues (HHDs), for a net 45.92% accuracy. However, 17.93% of references found by C&C (14.29% in HRDs, 19.26% in HHDs) were not references at all, but artifacts or parse errors. For example, the parser frequently decided that utterances like "Right, so" referred to entities on the right. Discarding these parse errors, GH-POWER correctly resolved 55.96% of references (57.14% in HRDs, 55.50% in HHDs). The remaining 44.04% of references could not be resolved due to several reasons, shown in Fig. 5:

4.97% of references (2.38% in HRDs, 5.96% in HHDs) were *plurals* (e.g. "the tubes"). GH-POWER was unable to resolve these as it is designed to handle *singular* references. Future work will be needed to generate likely groupings of entities to which plurals might be resolved.

10.60% of references (10.71% in HRDs, 10.55% in HHDs) referred to non-discrete entities, e.g., regions or sections of tube. Future work will be needed to generate likely *regions* or *portions* of entities to which such references might be resolved.

10.26% of references (10.71% in HRDs, 10.09% in HHDs) needed gestural information to be disambiguated; while it is an explicit design aim for GH-POWER to handle this facet of multi-modal interaction, we do not yet make use of such information. Future work will be needed to use gesture and eye gaze to correctly bias entities' salience scores.

4.64% of references (1.19% in HRDs, 5.96% in HHDs) were incorrectly resolved due to inconsistencies regarding the "beginning" of the task. For example, participants sometimes started interactions with utterances similar to "I will now describe *it* to you". Because speaker and listener shared a joint context at the start of the task, *the task* may have been in the listener's focus of attention. However, in the evaluation, the system never "heard" the experimenter giving instructions, and thus "the task" was considered at most activated.

3.31% of references (1.19% in HRDs, 4.13% in HHDs) were idiomatic or colloquial. For example, "that was it" was used to indicate task completion. This suggests that GH-POWER may need tighter integration with pragmatic inference.

1.32% of references (1.19% in HRDs, 1.32% in HHDs) were incorrectly resolved because the linguistic salience score we used did not sufficiently boost the target. Future work will be needed to investigate other salience scoring functions.

The remaining 8.94% of references (15.48% in HRDs, 6.42% in HHDs) were incorrectly resolved for various other reasons. For example, some participants referred to some concepts we were unprepared to handle (e.g., "The problem here…"), and some participants used indefinite noun phrases in ways we did not anticipate (e.g., "There is a pipe there").

V. Discussion

These results suggest several promising directions for future work towards the goal of truly natural HRI. First, while GH-POWER handled a surprising portion of naturally occurring references despite only using linguistic salience, this portion could be made larger by accounting for gesture and eye gaze. An obvious first step would be to utilize the annotated gestural information contained in our evaluation corpora. These results also suggest that plural and non-discrete references may be relatively common in task-based dialogues; handling each category is an interesting research question in its own right. Future work could also investigate how common-sense affordance-based reasoning capabilities might operate within the GH-POWER framework, as psycholinguistic work [2] suggests that affordance reasoning allows humans to eliminate many unlikely resolution candidates.

We are also interested in the interaction between GH-POWER and natural language *generation*, both with respect to the generation of referring expressions themselves, as well as the generation of clarification requests, either to indicate that the robot does not know of any suitable referent, or that it needs to know which of several referents is correct.

Most importantly, however, this work will allow us to study the interaction between natural language processing and cognitive processes such as memory and attention, all of which are integrated within the GH framework. We hope that GH-POWER will serve as a valuable starting point for studying the interaction of these processes, both for integrated system algorithm development and cognitive modeling research.

VI. Conclusion

We have presented GH-POWER: an open-world reference resolution algorithm for human-robot dialogue based on the *Givenness Hierarchy*, and shown its ability to handle the majority of references naturally occurring in task-based interactions. GH-POWER improves upon the GH by formalizing inter-tier traversal, salience-based intra-tier candidate selection, and multiple resolution. This allowed us to make significant theoretical extensions to the GH and to extend the state of the art in human-robot dialogue in several important ways. First, GH-POWER uses the *complete* GH to handle linguistic forms and resolution phenomena which could not be captured by previous approaches. Second, GH-POWER is able to handle the *uncertain* and *open worlds* commonplace to HRI scenarios. Finally, GH-POWER provides a starting point for studying the integration of cognitive processes in both robots and humans.

VII. Acknowledgments

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

References

[1] J. Chai, Z. Prasov, and S. Qu. Cognitive principles in Robust multimodal interpretation. *Journal of A.I. Research*, 27, 2006.

[2] C. Chambers, M. Tanenhaus, and J. Magnuson. Actions and affordances in syntactic ambiguity resolution. *Journal of experimental psychology: Learning, memory, and cognition*, 30(3), 2004.

[3] S. Clark and J. Curran. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Comp. Linguistics*, 33(4), 2007.

[4] S. Coradeschi, A. Loutfi, and B. Wrede. A short review of symbol grounding in robotic and intelligent systems. *Künstliche Intelligenz*, 27(2):129–136, 2013.

[5] H. P. Grice. Logic and conversation. In *Syntax and Semantics 3*. 1970.

[6] J. Gundel. Reference and Accessibility from a Givenness Hierarchy Perspective. *International Review of Pragmatics*, 2(2), 2010.

[7] J. Gundel, M. Bassene, B. Gordon, L. Humnick, and A. Khalfaoui. Testing predictions of the Givenness Hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, 42(7), 2010.

[8] J. Gundel, N. Hedberg, and R. Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 1993.

[9] J. Gundel, N. Hedberg, and R. Zacharski. Underspecification of cognitive status in reference production: Some empirical predictions. *Topics in cognitive science*, 4(2), 2012.

[10] J. Gundel, N. Hedberg, R. Zacharski, A. Mulkern, T. Custis, B. Swierzbin, A. Khalfoui, L. Humnick, B. Gordon, M. Bassene, and S. Watters. Coding protocol for statuses on the givennes hierarchy. unpublished manuscript, May 2006.

[11] A. Kehler. Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction. In *Proc. of 14th AAAI Conf. on AI*, 2000.

[12] G.-J. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, I. Kruijff-Korbayová, and N. Hawes. Situated dialogue processing for human-robot interaction. In *Cognitive Systems*, pages 311–364. 2010.

[13] S. Lemaignan, R. Ros, E. A. Sisbot, R. Alami, and M. Beetz. Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics*, 4(2), 2012.

[14] D. Roy and E. Reiter. Connecting language to the world. *Artificial Intelligence*, 167(1):1–12, 2005.

[15] M. Scheutz. ADE - steps towards a distributed development and runtime environment for complex robotic agent architectures. *Applied A.I.*, 2006.

[16] M. Scheutz, G. Briggs, R. Cantrell, E. Krause, T. Williams, and R. Veale. Novel mechanisms for natural human-robot interactions in the diarc architecture. In *Proc. of AAAI W.S. on Intelligent Robotic Systems*, 2013.

[17] M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson. First steps toward natural human-like HRI. *Autonomous Robots*, 22(4), May 2007.

[18] S. Schreitter and B. Krenn. Exploring inter-and intra-speaker variability in multi-modal task descriptions. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014.

[19] M. Steedman. *The syntactic process*, volume 24. MIT Press, 2000.

[20] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, 32(4):64–76, 2011.

[21] T. Williams, R. Cantrell, G. Briggs, P. Schermerhorn, and M. Scheutz. Grounding natural language references to unvisited and hypothetical locations. In *Proc. of 27th AAAI Conf. on Artificial Intelligence*, 2013.

[22] T. Williams and M. Scheutz. A domain-independent model of open-world reference resolution. In *Proceedings of the 37th annual meeting of the Cognitive Science Society*, 2015.

[23] T. Williams and M. Scheutz. POWER: A domain-independent algorithm for probabilistic, open-world entity resolution. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.

[24] T. Williams and M. Scheutz. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *Proc. of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.