Reference in Robotics: a Givenness Hierarchy Theoretic Approach

Tom Williams and Matthias Scheutz
{williams,mscheutz}@cs.tufts.edu
Human-Robot Interaction Laboratory
Tufts University
200 Boston Ave.
Medford MA, 02145

1 Introduction

As robots become increasingly prevalent in our society, it becomes increasingly important to endow them with natural language capabilities. Natural language capabilities are especially important for robots designed to operate in domains such as *eldercare robotics*, *education robotics*, *space robotics*, and *urban search-and-rescue* robotics. In eldercare robotics and education robotics, it may simply be too *cognitively burdensome* for the target population to learn to interact with their would-be caregiving or educational assistants through some other modality. In space robotics and urban search-and-rescue robotics, it may be too *physically burdensome* for the target population to interact with their would-be assistants or rescuers, due to, e.g., lack of gravity, or trapped limbs. In urban search-and-rescue

environments, victims are also not likely to have the time or inclination to learn another control modality to interact with their would-be rescuers. It is thus important that robots operating in these and other domains be taskable through control modalities, like natural language, that the general public is already familiar and proficient with.

A crucial aspect of natural language communication is the ability to *refer*: the capability to which this volume is dedicated. Robots must thus be able to both *understand* so-called *referring expressions*, and *generate* them as well. In this chapter we will focus on the task of referring expression *understanding*. There are a number of unique challenges that present themselves to robots seeking to understand referring expressions due to robots' status as *situated agents*: agents (entities capable of autonomously acting to achieve their own goals (Jennings, 2000)) that are embedded in an environment that is perceivable and manipulable by themselves and other agents with whom they can interact (Smith & Gero, 2005).

While a software entity operating within a non-situated domain such as text mining or document summarization may need to associate entities referenced in a text with previous portions of that text, a robot must instead associate entities referenced in dialogue with its own *mental representations* resulting not only from dialogue and inference, but also from interpretation of sensory data gathered by its perceptual systems. The robot's knowledge of perceived entities will almost certainly be uncertain

(as robots do not have perfect perception of the world) and incomplete (as robots cannot presume to be familiar with every entity in the world).

In robotics, the problem of identifying what real-world entities are the referents of referring expressions goes by many names, including *language grounding* (Steels & Hild, 2012), *reference resolution* (Popescu-Belis, Robba, & Sabah, 1998), and *entity resolution* (Meyer, 2013). While these names are sometimes used to denote the same concept, they carry different connotations. Specifically, *reference resolution* connotes association of a referring expression with a discretely represented entity, while *language grounding* further connotes grounding that discrete representation to continuously represented perceptual data (Harnad, 1990). In our work, we are specifically interested in *reference resolution* that is *domain independent* (i.e., in which it is not assumed that referents will be of one particular type (c.f. *spatial reference resolution*)) and *open world* (i.e., in which it is is not assumed that all candidate referents are perceivable or otherwise known at resolution time); assumption relaxations that are particularly important for realistic human-robot interaction scenarios. Imagine a search-and-rescue scenario, in which a supervisor says to a robot:

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

This example contains three referring expressions: 'The east wing', 'that', and 'all personnel'. We propose that a *domain-independent*

algorithm is needed to resolve these references, which are of different types, and that an *open-world* algorithm is needed as the agent should be able to understand the utterance even if it did not previously know that the building being discussed had an 'east wing'. The need for open-world reference resolution algorithms underlies our decision to study only the reference resolution half of the language grounding problem. In closed-world resolution, at resolution time all entities are either perceivable or have been previously perceived, and thus continuous perceptual data is available to adjudicate the fitness of all candidate referents. Because this is not the case in an open world, it is important to develop reference resolution algorithms that do not require symbols to be ground to perceptual data, as do language grounding algorithms.

Furthermore, We argue that it is important to focus on the two halves of the language grounding problem separately, as a robot is unlikely to generate subsymbolic representations for entities learned of through *dialogue*. In order to facilitate domain-dependent open-world reference resolution, we have developed an algorithm which makes use of the *Givenness Hierarchy* (GH) (Gundel, Hedberg, & Zacharski, 1993), which provides an elegant linguistic framework for reasoning about notions of reference in human discourse.

In the next section (Section 2), we provide an overview of the GH and discuss previous GH-theoretic approaches to reference resolution. We then describe in Section 3 our own GH-theoretic approach, the GH-POWER

algorithm, and suggest future refinements of our algorithm with respect to the theoretical commitments of the GH. In Section 4, we go on to briefly survey other prominent approaches to reference resolution in robotics, and discuss how these compare to our approach. Finally, in Section 5 we conclude with a discussion of possible directions for future work.

2 The Givenness Hierarchy

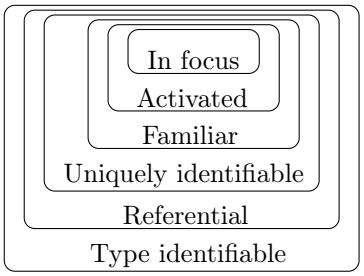

Figure 1: The Givenness Hierarchy

The GH (Gundel et al., 1993) is comprised of six hierarchically nested tiers of cognitive status, as seen in Fig. 1. If a candidate referent is marked as having one of these statuses, the hierarchical nature of this framework means that it also has all statuses lower in the hierarchy. For example, a candidate referent that is *familiar* is also *uniquely identifiable*, *referential*, and *type identifiable*. It is *possible* that the candidate referent is also *activated*, or even *in focus*, but a higher status cannot be inferred from a lower status. Each level of the GH is 'cued' by a set of linguistic forms, as seen in Table 1 for English. For example, the second row of the table shows

that when the definite 'this' is used, one can assume that *the speaker assumes* the referent of 'this' to be *at least* activated for their interlocutor.

Table 1: Cognitive Status and Form in the GH

Cognitive Status	Linguistic Form
In focus	*it*
Activated	*this,that,this* N
Familiar	*that* N
Uniquely identifiable	*the* N
Referential	indefinite *this* N
Type identifiable	*a* N

The GH is attractive to computational researchers not only because it suggests a clear mapping between linguistic form and cognitive status, but because, due to its focus on *means of access* rather than *salience*, each status evokes a particular *mnemonic actions* (i.e., actions involving selecting or creating mental representations) upon an agent's cognitive structures.

When the linguistic form of an expression explicitly signals that its referent is type identifiable or referential (but not necessarily uniquely identifiable), this suggests the action of *hypothesization*: creating a *new* mental representation, and then selecting that representation as the target referent.

When the linguistic form of an expression signals that its referent can also be uniquely identified (but is not necessarily familiar), this suggests *either* the action of hypothesizing a referent *or* selecting an existing referent from memory. When the linguistic form of an expression signals

that its referent is also familiar, this suggests that the referent should be able to be found by searching through memory and selecting an existing representation.

When the linguistic form of an expression signals that its referent is also activated or in focus, this suggests that the referent should be able to be found by searching through a subset of memory (the subset of activated entities and the subset of activated entities that are in focus, respectively) and selecting a referent from that subset.

The GH can directly solve certain computational problems: To determine the cognitive status ascribed to a candidate referent, one need only check which forms explicitly encode which statuses on the GH in a given language (see also the Coding Protocol provided by Gundel et al. (2006)). And, when Speaker S uses linguistic form F to refer to entity E when speaking to hearer H, it is easy to determine the *most restrictive* status that H can rationally assume S to ascribe to E. For example, when S uses 'it', we can assume that S believes E to be in the subset of H's memory that is *in focus*: any information that could *not* plausibly be *in focus* can be ruled out, as such an interpretation *would not be possible* given the cognitive status conventionally signaled by 'it'; when S uses 'this', we can assume that S believes E to be at least in the subset of H's memory that is currently *activated*. E may also be in the subset of those entities that are *in focus*, but we can not assume this; and in fact, it is unlikely that S believes E to be in that subset, as otherwise S could have used the more

informative 'it'. Furthermore, information that could not plausibly be in the *activated* subset of H's memory can be ruled out, as such an interpretation *would not be possible* given the cognitive status conventionally signaled by 'this'.

However, within the GH framework, choices among referents that meet cognitive status restrictions are made through *interaction* of the GH with general pragmatic principles operative in language interpretation, such as Grice's maxims or Relevance theory. As a result, the GH can only *facilitate*, but not *directly produce* solutions for the aforementioned computational problems of reference resolution (i.e., determining, when S uses linguistic form F when speaking to H, what entity E is most likely being referenced) and referring expression generation (i.e., determining, when S wishes to refer to E when speaking to H, what linguistic form F should be used (c.f.(Krahmer & Van Deemter, 2012; Van Deemter, 2016)). As previously discussed, this chapter will focus on a GH-theoretic approach to the reference resolution problem.

There have been several previous partial implementations of the GH (e.g., (Kehler, 2000; Chai, Prasov, & Qu, 2006)) for use in reference resolution algorithms, the most extensive of which is that presented by Chai et al. (Chai et al., 2006). Chai et al. were interested in handling multi-modal referring expressions within the context of multi-modal user interfaces, and combined the GH with ideas from Grice's theory of Conversational Implicature (Grice, 1970) to produce the reduced four-tier

hierarchy seen in Fig. 2. In this hierarchy, 'Focus' subsumes the GH's *in focus* and *activated* tiers; 'Visible' subsumes the *familiar* and *uniquely identifiable* tiers; and 'Others' subsumes the *referential* and *type identifiable tiers*, but crucially, does not appear to be used. These three tiers are topped by a new 'Gesture' tier which specifically handles gestured-towards entities.

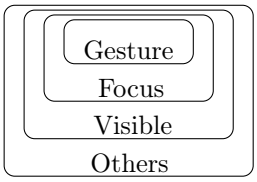

Figure 2: Chai's Modified Hierarchy

In previous work (Williams, Acharya, Schreitter, & Scheutz, 2016), we discuss how this reduced hierarchy, and its accompanying algorithm presented by Chai et al., do not address all aspects of reference resolution found in typical human-robot dialogues. To summarize our concerns, the approach taken by Chai et al. (1) assumes complete certainty in the properties of entities, (2) appears to only handle references to objects, not locations, people, or less concrete entities (e.g., utterances) (3) operates under the closed-world assumption, (4) does not account for the GH's preference for lower tiers over higher tiers (e.g., in 'Could you repeat *that*', 'that' is more likely to refer to an *activated* entity (e.g., the previous utterance) than an *in-focus* entity (e.g., the topic of the previous utterance), (5) cannot differentiate between the tiers that they join together, and (6) may be prone to errors and inefficiencies when resolving complex noun

phrases. To address our concerns, we presented the GH-POWER algorithm (Williams et al., 2016), which we outline in the next section.

3 The GH-POWER Algorithm

The GH-POWER reference resolution algorithm dictates how the referent of a referring expression should be searched for, given a memory model organized in a specific, hierarchical way that parallels the organization of the GH. In this section, we will first discuss the memory structure used in our approach. Next, we will discuss the *between-structure processes* by which GH-POWER algorithm chooses which structures to search. We will then discuss the *within-structure processes* by which GH-POWER selects suitable referents from a given structure. Finally, we will discuss the shortcomings of this approach which we are seeking to address in our current work. In all cases we choose to describe our approach at a high level, and do not provide pseudocode or optimization details. The interested reader should consult our previous work for such details, as cited throughout the section.

3.1 The GH-POWER Memory Model

The memory model used by GH-POWER aligns well with Nelson Cowan's conceptualization of working memory (Cowan, 1998). According to Cowan, working memory and long-term memory are not disjoint structures.

Rather, working memory can be regarded as the subset of entities in long term memory that are currently activated. Cowan further posits an additional substructure, the focus of attention, which is a subset of those activated entities that is limited in size to at most four elements, comprised of those items of which an agent is consciously aware. There is clearly a strong parallel between Cowan's *Focus of Attention* $\subset$ *Set of Activated Entities* $\subset$ *Long Term Memory* structures and Gundel's *In Focus* $\subset$ *Activated* $\subset$ *Familiar* statuses, and observing this connection will facilitate understanding the connection between our own memory structure and the statuses of the GH.

Our approach consists of *four* hierarchical data structures: the *Focus of Attention* (FOA), *Set of Activated Entities* (ACT), *Set of Familiar Entities* (FAM) and *Long Term Memory* (LTM). These four data structures are hierarchically organized such that FOA $\subset$ ACT $\subset$ FAM $\subset$ LTM. At the *computational level* of analysis (Marr, 1982), the FOA, ACT, and LTM data structures are identical to Cowan's three memory structures. But in a robot architecture, all of a robot's knowledge is not typically located in a single, monolith knowledge base. Instead, it may be distributed across a *set* of knowledge bases that may be located on different machines, may use different knowledge representation schemes, and may have different ways of accessing and modifying the knowledge contained within them. Thus, at the *algorithmic level*, our LTM data structure is really a set of domain-specific *distributed, heterogeneous knowledge bases*. Because LTM is

not a single coherent knowledge base, the FOA and ACT also must differ at the algorithmic level; instead of being literal subsets of the mental representations distributed across LTM, the FOA and ACT instead contain *memory traces* that allow access to certain of those mental representations. Note that these three structures are not intended to serve as the agent's *actual* cognitive structures; instead, they serve to model what an interlocutor might *believe* to be in those structures, and thus as a model of *common ground*.

Finally, for the sake of convenience and efficiency, we introduce the FAM structure, a minor point of departure from both the GH and Cowan's model of Working Memory, which we make for practical rather than theoretical reasons. FAM contains memory traces for entities in LTM that are likely to be referenced, such as entities mentioned at some point in the robot's current dialogue, recently visited locations, and recently visited objects, including all entities in ACT (and by extension, in the FOA). Because searching all of LTM is potentially expensive, when LTM needs to be searched for an entity that matches some criteria, that search is preempted by a search of FAM: if a match can be found there, LTM need not be searched.

To summarize, our model consists of four hierarchically nested data structures: a distributed LTM data structure containing mental representations of known entities, and three smaller data structures that contain memory traces allowing fast access to entities in LTM (i.e., FOA,

ACT, and FAM). These three data structures are populated periodically (e.g., after an utterance is processed) according to rules inspired by the GH Coding Protocol. In the next two sections, we will describe how these structures are used during reference resolution: in Section 3.2 we discuss how the linguistic form of a referring expression is used by the GH-POWER algorithm to determine *which* of these structures to examine; in Section 3.3, we discuss how GH-POWER chooses whether a particular candidate referent within one of those structures is the target referent.

3.2 Between-Structure Processes

The GH alone does not specify how cognitive structures are selected for perusal during reference resolution. For example, suppose Speaker S uses the pronoun'that' to refer to entity E when speaking with Hearer H. The GH suggests that H can assume that S *assumes* that E is at least in H's ACT, and thus *may or may not* also be in H's FOA.

Several strategies could be used to search ACT and the FOA. The agent could consider entities in the FOA, then out-of-focus entities in short term memory (a top down approach), or she could consider out-of-focus entities in ACT, then in-focus entities (a bottom up approach).

While some previous approaches (e.g., (Chai et al., 2006)) have used a global top-down approach, this may violate certain predictions of the GH. For example, the Givenness Hierarchy framework (i.e., the GH when working in conjunction with general cognitive principles such as Grice's

maxim as Quantity) suggests that in the example above, while the referent of 'that' *could be* assumed to be in *H*'s FOA, it is more likely to be assumed to be in *H*'s ACT *but not in H's FOA*, as otherwise *S* could have used 'it' to refer to the referent. If a purely top-down approach is used, this effect may not be captured. On the other hand, consider the utterance 'Pick up the box'. The bottom-up approach would inappropriately prioritize inactive boxes from LTM over an activated box in front of the listener. Since neither a purely top-down or purely bottom-up approach seems adequate, we developed a hybrid approach, in which a unique search strategy is used for each GH tier. These strategies, refinements of those presented in (Williams et al., 2016) are seen in Table 2. In that table, FOA denotes a search through memory traces found in the FOA; ACT denotes a search through memory traces found in ACT *but not in* the FOA; FAM denotes a search through memory traces found in FAM *but not in ACT* LTM denotes a search through all of LTM; HYP denotes hypothesization. We will now explain the rationale for each strategy.

Table 2: Search Plans for Complete GH

Level	Search Plan
in focus	FOA
activated	ACT → FOA
familiar	ACT → FOA → FAM → LTM
uniquely identifiable	ACT → FOA → FAM → LTM → HYP
referential	ACT → FOA → HYP
type identifiable	HYP

3.2.1 In Focus

In the case of an 'in focus' cuing form (e.g., 'it'), we only search the FOA, as it would be otherwise inappropriate to use such a form.

3.2.2 Activated Entities

In the case of an 'activated' cuing form (e.g., 'this'), search is expanded to include out-of-focus entities in ACT. For the reasons discussed above, we proceed bottom-up, first searching the out-of-focus entities in ACT, then searching the FOA. However, this process is modified in the case of 'This N', as we discuss below.

3.2.3 Familiar Entities

In the case of a 'familiar' cuing form (e.g., 'that N'), search is expanded to include all entities in memory. As it is inappropriate to *prioritize* entities in LTM over those in ACT, we still perform our search through ACT and the FOA first, and then move on to search through LTM. As previously discussed, we first search through FAM, the subset of most probable referents in LTM (not including those referents also found in ACT), and only search *all* of LTM if this search fails, using the DIST-POWER algorithm described in (Williams & Scheutz, 2016). DIST-POWER is a distributed extension of the cognitive model proposed in (Williams & Scheutz, 2015a) and computationalized in (Williams & Scheutz, 2015b). This algorithm has two main features relevant to GH-POWER. First, it is able to simultaneously

resolve all parts of a complex definite description. The second feature will be discussed in the following subsection.

3.2.4 Uniquely Identifiable

In the case of a 'uniquely identifiable' cuing form (e.g., 'the N'), search is extended to allow for the possibility that the speaker is referencing a previously unknown entity. This begins by searching through the four tiers of the GH-POWER memory model, as performed with familiar entities. However, when searching through LTM, we take advantage of DIST-POWER's second important feature DIST-POWER's 'hypothesization mode'. When run in this mode, if DIST-POWER is unable to find a mental representation that satisfies all semantic criteria of a definite description, it attempts to find a subset of that description that it *can* successfully resolve, and automatically hypothesizes representations for remaining entities.

3.2.5 Referential

Gundel et al. suggest that the *indefinite* form of 'this N' (as in 'This dog I saw was enormous!') cues the 'referential' tier[1], resulting in the hypothesization of a representation. As a simplification (i.e., so that we do not need to decide whether each use of 'This N' is definite or indefinite), GH-POWER deals with both forms at the referential tier. To do so, we begin with the standard 'activated' search strategy (i.e., a bottom-up search starting from ACT), and hypothesize a representation only if this search

fails. We acknowledge that there may be cases in which this does not produce the correct behavior. For example, if one says 'This dog I saw was enormous!' while standing in front of a dog, 'This dog' may be incorrectly resolved to the co-present canine.

3.2.6 Type Identifiable

In the case of a linguistic form that *only* cues the Type Identifiable tier (e.g., 'a N'), we immediately hypothesize a representation in the same way as is performed in the previous subsection. Note that this does not imply that the robot does not yet have a representation for the intended referent. For example, suppose the robot is looking at a box, and its interlocutor says to it remotely, 'You should see a box: Bring it to me.' In this case, the robot's interlocutor actually intends to refer to a particular box, and the robot in fact already knows of this box. Even in such a case, we still create a new mental representation for a new box. It will be up to subsequent processing stages to recognize the meaning of the sentence, find the two representations, verify that they match, consolidate them into a single representation, and of course, bring the box to the interlocutor.

3.2.7 Complex Referring Expressions

The GH framework also does not specify how to resolve syntactically complex referring expressions, i.e., referring expressions containing multiple referents described in relation to each other, such as those in Example 2:

(2) *Scene: A table upon which sits a large green block and a large blue block (towards the front of the table), and a greenish-yellow block on a bluish-purple block (in a far corner of the table).*

a. Pick up the green block that is on the blue block.

b. Pick up the one on the blue block.

Chai et al. resolve references of this sort using a a *greedy algorithm* in which locally optimal choices are sequentially made for each sub-expression. However, in cases like that seen in Example 2a, this is likely to incorrectly resolve whichever referential expression is considered first, due to the decreased salience, prototypicality, and proximity of the targets. Greedily resolving Example 2b will likely be even less successful due to the underspecification of 'the one'.

We would thus argue that syntactically complex referring expressions should not be considered greedily in a GH-theoretic reference resolution algorithm. How, then, should search plans (i.e., from Table 2) for an expression's constituent parts be jointly examined? We decided to handle this problem by 'crossing' the search plans for the constituent parts, that is, considering all possible combinations of search plans sorted in search plan order. For example, crossing $ACT \rightarrow FOA \rightarrow HYP$ with $ACT \rightarrow FOA$ yields Table 3.

The rows of this table are successively examined until a sufficiently probable solution is found or the table is exhausted. Here, for example,

Table 3: Sample Joint Search Plan Table

Y	X
ACT	ACT
ACT	FOA
FOA	ACT
FOA	FOA
HYP	ACT
HYP	FOA

GH-POWER would begin by searching for a pair of memory traces contained in ACT which fit the given description. If no such pair can be found, GH-POWER will proceed to the next line of the table, and search for a pair of memory traces, one from ACT and one from the FOA, which fit the given description, and so forth. Two decisions were made in designing this subroutine.

First, rows are considered in left-to-right order. For example, when searching for a pair of referents on the second line, GH-POWER would first try to find candidate referents from ACT to associate with Y, and then try to find candidate referents from FOA to associate with X *given the set of restricted candidates for* Y.

Second, the action of hypothesization (denoted HYP) is postponed until the search process is successfully terminated; a new representation should only be generated if sufficiently probable referents are found for all other entries in a row, halting the search process. For example, when line five is considered, GH-POWER will begin by associating Y with a dummy referent '?'. A new representation will be created for this referent *if and*

only if a sufficiently probable referent can be found in ACT to associate with X.

3.3 Within-Structure Processes

The GH does not specify how candidates are selected from *within* cognitive structures during reference resolution. Despite what is often assumed (c.f. (Brown-Schmidt, Byron, & Tanenhaus, 2005)), Gundel et al. state that the GH is *not* a hierarchy of salience or accessibility, and that it is necessary to model salience *independently* of tier of cognitive status (Gundel, 2010). We will now describe how the proposed model handles degree of salience and uncertainty, and how these measures are used to select candidates.

3.3.1 Focus of Attention and Activated Entities

In order to account for salience without relying on, e.g., a dedicated gestural tier (c.f. (Chai et al., 2006)), GH-POWER uses a multi-modal salience score to assign a 'degree of activation' to entities contained in the FOA and ACT. The entities returned by the *assess* methods associated with the FOA and ACT structures are then the set of all *sufficiently probable* entities within those tiers, ordered by activation such that the most salient candidate will be chosen if multiple are available.

3.3.2 Familiar Entities and Long Term Memory

In the proposed model, the Set of Familiar Entities is equivalent to a 'highly salient' LTM cache; we would argue that the 'Familiar' and 'Uniquely Identifiable' tiers can be viewed as different means of accessing the same structures, with different worst-case conditions. This is consistent with Gundel's claim (Gundel, 2010) that:

> "[F]orms that encode cognitive status on the GH are not markers of *degree* of accessibility. Rather, they provide procedural information about *manner* of accessibility, how and where to mentally access an appropriate representation."

The entities returned by the *assess* method associated with the FAM are its *sufficiently probable* entities, ordered in *reverse chronological order*; the entities returned by the *assess* method associated with LTM are its *sufficiently probable* entities, ordered in decreasing order of *likelihood*.

3.4 Discussion

In previous work, we demonstrated how GH-POWER was able to resolve the majority of references occurring in a corpus of human-human and human-robot team tasks (Williams et al., 2016). While, there were a number of cases that GH-POWER was unable to handle, it was able to capture several aspects of the GH missing from previous GH-theoretic

approaches. Consider, for example, the following example presented by Gundel et al. (Gundel, 2010):

(3) a. Alice: I failed my linguistics course.

b. Bob: Can you repeat that?

Before resolving 'that', the referent of 'my linguistics course' should be in the agent's FOA, while the utterance itself should be in the agent's ACT, but not in the agent's FOA since, as Gundel et al. (1993) note, speech acts are activated, but not brought into focus just by being uttered. Gundel et al. suggest that if Bob had meant to refer to the course, he would have used *it* instead of *that*, because 'it' explicitly picks out an in focus referent, whereas 'that' only signals that the referent is activated and therefore could be in focus, and thus the course should be dispreferred to the sentence itself. This effect is captured through GH-POWER's between-structure processes: When 'that' is used, ACT is first checked; and because the utterance is in ACT, it is chosen. FOA is not even examined, because any options residing therein should be dispreferred. However, consider Example 4:

(4) *Scene: A table on which sits a black box and a white box*

a. Bob: Look at the white box

b. Bob: Pick that up

Before resolving 'that', the white box should be in the agent's FOA, and the black box is likely to only be in the agent's ACT, as depicted in Fig. 3. Following the logic of Example 3, if Bob had meant to refer to the

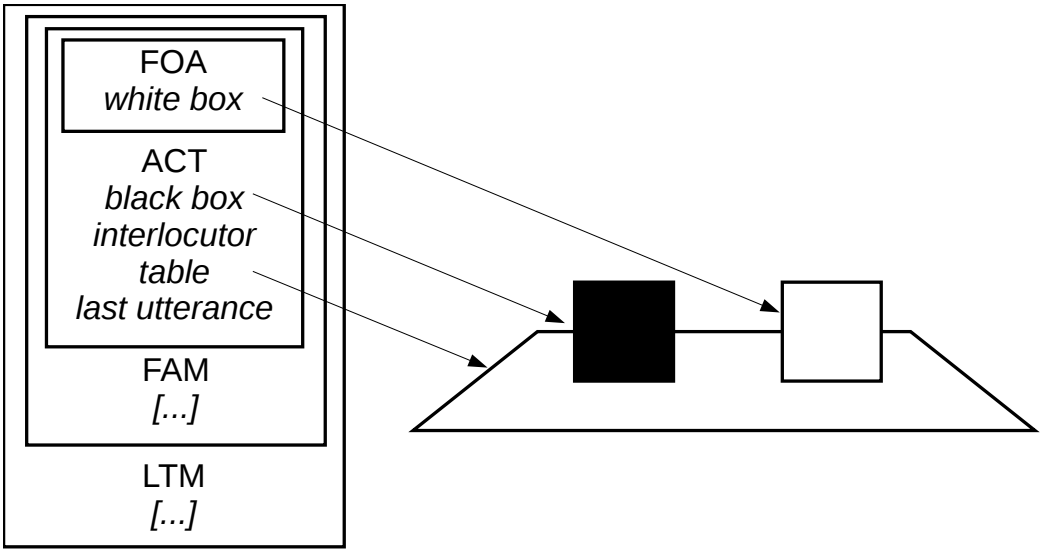

Figure 3: Contents of GH-POWER's cognitive structures during hypothetical algorithm run. Structures are arranged to depict their hierarchical nature (i.e., an entity in one structure is also in all lower structures). [...] indicates the wide variety of entities contained in the set of familiar entities and in long term memory which are not immediately relevant to this example.

white box, he could have used *'it'* instead of *'that'*, and thus the white box should be *dispreferred*. Yet while 'it' would have been more natural in referring to the white box, choosing the white box as the referent of 'that' is clearly not wrong and probably preferred in this situation in the absence of any gesture indicating a shift in attention.

In this scenario, GH-POWER errs for two reasons. First, it treats hierarchical preference as *absolute*, whereas dispreferred entities should be just that: dispreferred, not removed from consideration. Second, GH-POWER does not take *conversational relevance* into account. These factors were initially overlooked because the GH does not specify how

relevance factors influence search. GH-POWER checks whether resolution candidates are suitable, i.e., whether they satisfy all described properties, and only moves on to consider entities in another cognitive structure if no resolution candidates in the current structure are deemed *suitable*. In this case, however, this is insufficient. In order for GH-POWER to perform correctly in this scenario, it should recognize not only that both boxes are *suitable*, but that the white box is conversationally more relevant than the black box; there is no clear reason why the agent would be asked to look at the white box and then pick up the black box.

In order to address this issue, GH-POWER should operate in the following way: When ACT is examined, the low conversational relevance of the black box should be noted. This should result in search extending to the FOA while retaining the black box as a resolution candidate. The white box should then be selected using an equation that takes relevance, suitability, and other factors into account.

To be precise, at least three factors must be considered in the within-structure processes of future versions of GH-POWER: (1) suitability (i.e., the agent's certainty that a candidate holds all described properties), (2) relevance, (i.e., the agent's certainty that reference to a candidate would not violate, e.g., Grice's Maxim of Relevance (Grice, 1970)) and (3) common-sense judgments (here, e.g., the agent's certainty that a candidate *can be picked up*). Note that each of these factors may be used differently: while a candidate must score highly on all three factors for the search to

cease, only low suitability will likely result in a candidate's complete removal from consideration. Furthermore, to respect the Theory-of-Mind considerations of the GH, this process must consider the extent to which the speaker would have been cognizant of each of these factors. We are currently in the process of developing a new algorithm that takes these factors into account.

4 Related Work

In the previous sections, we presented our own approach to reference resolution, and described how it compared to other GH-theoretic approaches (which had been developed within the context of multi-modal user interfaces). In this section, we will discuss how our approach relates to other approaches to reference resolution within robotics. While our approach is *open-world* in nature, there has been very little work in that area[2], and thus we will be primarily discussing *closed-world* approaches to reference resolution. We will also not discuss approaches that assume that a *unique name* (i.e., a *rigid designator*) is always provided (e.g., (Khayrallah, Trott, & Feldman, 2015)).

Reference resolution has been a topic of interest in robotics for nearly a half century, beginning with Winograd's SHRDLU system (Winograd, 1971). In the SHRDLU system, a simulated robot could be issued an array of natural language commands in order to move blocks around in its

simulated environment. Winograd introduced a *procedural semantics* view of natural language understanding in which each lexical item was associated with a short procedure to be executed: adjectives and nouns, for example, were associated with procedures which considered each entity in the scene and decided whether or not that entity fit the property denoted by the lexical item. SHRDLU was also able to perform anaphora resolution: anaphoric lexical items were associated with procedures that considered the plausibility of anaphoric reference to each entity in SHRDLU's world, giving preference to elements considered to be 'in focus' (see also (Mitkov, 1999)).

Decades later, SHRDLU continues to inspire researchers. Gorniak and Roy, for example, employ a similar approach (Gorniak & Roy, 2004; Roy, Hsiao, Mavridis, & Gorniak, 2003). In their work, a simpler approach to anaphora resolution is used, but they make important steps forward in other respects: whereas SHRDLU's procedural attachments consulted a knowledge base populated with hand-assigned symbolic properties, Gorniak and Roy's procedural attachments effect the incremental and greedy application of composable *visual models*. That is, properties of objects are assessed by consulting the continuous perceptual features of those objects, thus *grounding* internal symbols to the physical world. This thus represents a solution to the full *language grounding* problem. In our work, we only address the *reference resolution* half of the language problem, leaving the *symbol grounding* half to POWER's distributed knowledge bases. But unlike Gorniak and Roy's approach, GH-POWER is domain-independent,

and operates in uncertain and open worlds.

Just as Gorniak and Roy incrementally execute *procedures* associated with particular lexical items, Kruijff et al. (Kruijff, Lison, Benjamin, Jacobsson, & Hawes, 2007) incrementally employ a set of *comparators* that can make true/false judgments as to whether certain entities satisfy certain properties. But while Gorniak and Roy focus on grounding, do not address deixis, and only narrowly address anaphora, Kruijff et al. treat grounding as a separate process, and address many aspects of deixis and anaphora. Furthermore, Kruijff et al. use a central *symbolic* knowledge base that is *informed* by subsymbolic perceptual systems, which is not dissimilar from our own decision to use a set of distributed knowledge bases. Several other researchers have used knowledge-based approaches similar to Kruijff et al., differing in the way in which their knowledge bases are queried.

Lemaignan et al. use semantic parsing techniques to translate utterances into knowledge base queries.

Lemaignan et al. handle anaphora and deixis to a limited extent: anaphoric expressions are replaced by the last entity in the dialogue history that match animacy and gender constraints; deictic expressions are resolved to the most recent focus of simultaneous eye gaze and gesture. Like that of Kruijff, this approach does not handle uncertain or open worlds.

Zender et al. take a similar approach, but apply their approach to the *spatial domain* of large-scale topological spaces (e.g., rooms and hallways) rather than the *object* domain used by all previous approaches (Zender,

Kruijff, & Kruijff-Korbayová, 2009). Zender et al. also use semantic parsing techniques to translate utterances into knowledge base queries. As Zender et al. are specifically targeting references to large-scale locations, they did not attempt to handle deixis and eye gaze. They do, however, handle anaphora through a dedicated *co-reference resolution*[3] pre-processing step, similar to other approaches we will examine.

Meyer, for example, performs co-reference resolution to resolve anaphora; but this step is tightly coupled with his Markov Logic theoretic reference resolution system such that anaphoric and other referential expressions can be resolved as part of a joint model (Meyer, 2013). Deictic expressions are not handled by this approach, and referents are restricted to objects.

A slightly different approach is taken in recent work by Chai et al. (Fang, Liu, & Chai, 2012; Liu, Fang, She, & Chai, 2013; Chai et al., 2014) (in work distinct from their GH-theoretic approach discussed above). As dialogue unfolds, Chai et al. build up a *graph* representing the relations between discussed entities. When each utterance is heard, Chai et al. use a graph matching algorithm to find the best partial overlapping region between the dialogue graph and a separate *vision graph* representing the relations between currently *observed* entities. Anaphora is handled in this approach by a co-reference resolution pre-processing step; and while deictic expressions are not discussed in this work, Chai et al. show in earlier work how gestural information can be integrated into their dialogue graph

structure (Chai, Hong, & Zhou, 2004). Furthermore, while more recent publications do not discuss it, older work suggests that they are able to handle *uncertain properties* (Fang et al., 2012). The previous approaches we have discussed assume that sensors provide straightforward true-or-false judgments on whether entities in the world have certain properties; but in realistic situated interactions, an agent may not always be certain whether certain entities in the world hold certain properties. Chai et al.'s approach begins to address this, by incorporating *extent of compatibility* into their graph-matching scoring functions. While Chai et al.'s approach is, like GH-POWER, able to handle uncertain properties, it is unable to handle open worlds.

Another approach that begins to address the uncertain nature of reality is the Semantic Fields work presented by Fasola and Mataric (Fasola & Matarić, 2013). In that work, reference resolution at the lowest level is quite simple: when nouns are used, a knowledge base is examined to determine whether it contains a unique entity with that label. If so, that entity is chosen as the target referent. If not, their approach attempts to disambiguate using *semantic field* models of spatial prepositions. This approach does not handle deixis, and seems to be restricted to operation in environments in which knowledge of objects is provided *a priori*. It does handle some anaphoric expressions however, by choosing as the target referent the most recent entity that matches gender and animacy constraints (Fasola & Matarić, 2014). While the reference resolution

portion of this work may not have all the capabilities of other approaches, and is limited to closed worlds, we believe it would be interesting to integrate the preposition models used by this approach into the GH-POWER framework. It is important to recognize that GH-POWER is not necessarily incompatible with all of the examined approaches. Many of these approaches present innovative ways for grounding or evaluating certain properties, and as long as these methods can be adapted to produce probability values, they can be integrated into the GH-POWER framework.

In contrast to the approaches examined thus far stand a number of recent *Bayesian* approaches, which seek to more formally handle uncertainty. Kennington and Schlangen present an incremental Bayesian model: as each word in a sentence is heard, the probability of each entity in a scene being the target referent is modulated based on learned probabilistic models that associate lexical units with observable properties (Kennington & Schlangen, 2017). This approach handles deixis and gaze by linearly combining the probability of reference given the utterance, the probability of reference given the speaker's gaze, and the probability of reference given the speaker's gestures. Anaphora is handled by attributing a 'selected' property to entities which become selected through dialogue: pronouns are statistically associated with this property through the same learning process used for other lexical units. While this approach does handle uncertainty with respect to the relationship between words and features, it does not handle uncertainty with respect to the

relationships between features and objects: each object in the scene has a set of properties which are known a priori to be true of that object. This approach is much more cognitively plausible than the other approaches examined, including our own; and with this cognitive plausibility come a number of computational advantages, the foremost being the increased speed of processing inherent to the incremental approach. But this approach is unable to handle uncertain and open worlds – a limitation that is not entirely shared by the other Bayesian approaches we will examine.

In the Bayesian *Generalized Grounding Graph* approach taken by Tellex and Kollar, utterance structures are used to instantiate probabilistic graphical models where certain nodes are associated with certain words in the utterance (Tellex et al., 2011). This approach operates in a manner similar to our own (c.f. Williams and Scheutz (2015b)). Deixis and gaze do not appear to be handled in this approach, and like the previous approach, the uncertainty of objects' *properties* is not represented. Anaphora is handled through a co-reference resolution pre-processing step (Tellex et al., 2012), and it appears to be usable in an incremental fashion (Manek & Tellex, 2016). This framework does improve upon previous approaches, however, in an important way: it is not limited to handling just objects, or just locations, but handles both. It appears to handle references to any physically extant entities located at particular points in space. Furthermore, recent work building on this framework has begun to address *open world* resolution. The work by Duvallet et al. allows a robot to *hypothesize* a new

object described with respect to another object(Duvallet et al., 2014). This hypothesization is, however, limited to spatially situated objects. Through GH-POWER, we are interested in the hypothesization of not only spatially situated objects, but other entities such as agents and locations.

Similarly, Chung, Propp, Walter, and Howard (2015) extend Generalized Grounding Graphs to produce the *Hierarchical Distributed Correspondance Graph* approach, which uses utterance structures to instantiate probabilistic graphical models of a similar form. While this approach is more nascent, and thus has not been incrementalized and cannot yet handle deixis, gaze, or anaphora, it improves on previous approaches in that it begins to pay attention to *what* entities are considered. Considering all entities in the world when performing reference resolution may be feasible when you need only consider a small number of entities in a visual scene; but it will likely be computationally intractable in larger, more realistic settings. In Chung's approach, only the set of entities matching the correct *type* indicated by each noun phrase are considered as possible referents for that noun phrase. We believe that the GH provides a powerful alternative – in the majority of cases, GH-POWER need only consider the limited subset of entities in ACT and the FOA. Like Chung et al., we are interested in restricting the search space considered during reference resolution – but through GH-POWER, we are able to do so under uncertainty, and in a context-sensitive manner.

Finally, Matuszek, Fitzgerald, Zettlemoyer, Bo, and Fox (2012) present

an approach similar to both those of Tellex and Kollar, Gorniak and Roy, and Kennington and Schlangen. In this approach, a semantic parser is connected to a set of visual classifiers used to identify objects. As with Kennington and Schlangen, deixis and gaze are handled through a linear combination of probabilities. Unlike the majority of previous approaches, however, Matuszek et al.'s approach is able to represent the robot's *uncertainty* in the properties of the objects it detects, based on classifier confidences. This is an important step towards enabling operation in natural, realistic settings. Like the majority of previous approaches, however, Matuszek et al.'s approach is limited to handling references to objects, and is restricted to operation in a closed world.

5 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 the information distributed across 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 (Chambers, Tanenhaus, & Magnuson, 2004) and *incrementalize* and *parallelize* our algorithm, to come in line with psycholinguistic literature (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995), similar to previous work from our lab (Scheutz, Eberhard, & Andronache, 2004) and others (Kennington & Schlangen, 2017). 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.

6 Acknowledgments

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

References

Brown-Schmidt, S., Byron, D. K., & Tanenhaus, M. K. (2005). Beyond Salience: Interpretation of Personal and Demonstrative Pronouns. *Journal of Memory and Language*, *53*(2), 292–313.

Chai, J., Hong, P., & Zhou, M. X. (2004). A Probabilistic Approach to Reference Resolution in Multimodal User Interfaces. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (pp. 70–77).

Chai, J., Prasov, Z., & Qu, S. (2006). Cognitive Principles in Robust Multimodal Interpretation. *Journal of Artificial Intelligence Research*, *27*, 55–83.

Chai, J., She, L., Fang, R., Ottarson, S., Littley, C., Liu, C., et al. (2014). Collaborative Effort Towards Common Ground in Situated Human-robot Dialogue. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 33–40).

Chambers, C., Tanenhaus, M., & Magnuson, J. (2004). Actions and Affordances in Syntactic Ambiguity Resolution. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *30*(3), 687.

Chung, I., Propp, O., Walter, M. R., & Howard, T. M. (2015). On the Performance of Hierarchical Distributed Correspondence Graphs for Efficient Symbol Grounding of Robot Instructions. In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 5247–5252).

Cowan, N. (1998). *Attention and Memory: An Integrated Framework*. Oxford University Press.

Duvallet, F., Walter, M. R., Howard, T., Hemachandra, S., Oh, J., Teller, S., et al. (2014). Inferring Maps and Behaviors from Natural Language Instructions. In *Proceedings of the 2014 International Symposium on Experimental Robotics* (pp. 373–388).

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye Movements as a Window into Real-Time Spoken Language Comprehension in Natural Contexts. *Journal of psycholinguistic research*, *24*(6), 409–436.

Fang, R., Liu, C., & Chai, J. (2012). Integrating Word Acquisition and Referential Grounding Towards Physical World Interaction. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (pp. 109–116).

Fasola, J., & Matarić, M. J. (2013). Using Semantic Fields to Model Dynamic Spatial Relations in a Robot Architecture for Natural

Language Instruction of Service Robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 143–150).

Fasola, J., & Matarić, M. J. (2014). Interpreting Instruction Sequences in Spatial Language Discourse with Pragmatics Towards Natural Human-Robot Interaction. In *IEEE International Conference on Robotics and Automation* (pp. 2720–2727).

Gorniak, P., & Roy, D. (2004). Grounded Semantic Composition for Visual Scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470.

Grice, H. P. (1970). Logic and conversation. *Syntax and semantics*, *3*, 41–58.

Gundel, J. (2010). Reference and Accessibility from a Givenness Hierarchy Perspective. *International Review of Pragmatics*, *2*(2), 148–168.

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive Status and The Form of Referring Expressions in Discourse. *Language*, 274–307.

Gundel, J., Hedberg, N., Zacharski, R., Mulkern, A., Custis, T., Swierzbin, B., et al. (2006). *Coding Protocol for Statuses on the Givenness Hierarchy.*

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.

Jennings, N. R. (2000). On Agent-Based Software Engineering. *Artificial intelligence*, *117*(2), 277–296.

Kehler, A. (2000). Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction. In *Proceedings of the 14th AAAI*

Conference on Artificial Intelligence (pp. 685–690).

Kennington, C., & Schlangen, D. (2017). A Simple Generative Model of Incremental Reference Resolution for Situated Dialogue. *Computer Speech & Language*, *41*, 43–67.

Khayrallah, H., Trott, S., & Feldman, J. (2015). Natural Language For Human Robot Interaction. In *Proceedings of the Workshop on Human-Robot Teaming at the 10th ACM/IEEE International Conference on Human-Robot Interaction.*

Krahmer, E., & Van Deemter, K. (2012). Computational Generation of Referring Expressions: A Survey. *Computational Linguistics*, *38*(1), 173–218.

Kruijff, G.-J. M., Lison, P., Benjamin, T., Jacobsson, H., & Hawes, N. (2007). Incremental, Multi-Level Processing for Comprehending Situated Dialogue in Human-Robot Interaction. In *Symposium on Language and Robots.*

Liu, C., Fang, R., She, L., & Chai, J. (2013). Modeling collaborative referring for situated referential grounding. In *Proceedings of the 2013 SIGDIAL Conference* (pp. 78–86).

MacMahon, M., Stankiewicz, B., & Kuipers, B. (2006). Walk the Talk: COnnecting Language, Knowledge, and Action in Route Instructions. In *Proceedings of the 21st National Conference on Artificial intelligence* (pp. 1475–1482).

Manek, G., & Tellex, S. (2016). Incrementally Identifying Objects from

Referring Expressions using Spatial Object Models. In *Proceedings of the 2016 RSS Workshop on Model Learning for Human-Robot Communication.*

Marr, D. (1982). *Vision: a Computational Investigation.* San Francisco: W.H.Freedman and Company.

Matuszek, C., Fitzgerald, N., Zettlemoyer, L., Bo, L., & Fox, D. (2012). A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 1671–1678).

Matuszek, C., Herbst, E., Zettlemoyer, L., & Fox, D. (2012). Learning to Parse Natural Language Commands to a Robot Control System. In *Proceedings of the 13th International Symposium on Experimental Robotics* (pp. 403–415).

Meyer, F. (2013). *Grounding Words to Objects: A Joint Model for Co-reference and Entity Resolution Using Markov Logic for Robot Instruction Processing.* Unpublished doctoral dissertation, TUHH.

Mitkov, R. (1999). *Anaphora Resolution: the State of the Art.* School of Languages and European Studies, University of Wolverhampton.

Popescu-Belis, A., Robba, I., & Sabah, G. (1998). Reference Resolution Beyond Coreference: a Conceptual Frame and its Application. In *Proceedings of the 17th International Conference on Computational Linguistics* (pp. 1046–1052).

Roy, D., Hsiao, K.-Y., Mavridis, N., & Gorniak, P. (2003). Ripley, Hand

Me The Cup! (Sensorimotor representations for grounding word meaning). In *Proceedings of the international confwewnxw of automatic speech recognition and understanding.*

Scheutz, M., Eberhard, K., & Andronache, V. (2004). A Real-time Robotic Model of Human Reference Resolution using Visual Constraints. *Connection Science Journal*, *0091*(March), 145–167.

Smith, G. J., & Gero, J. S. (2005). What Does an Artificial Design Agent Mean by being 'Situated'? *Design studies*, *26*(5), 535–561.

Steels, L., & Hild, M. (2012). *Language Grounding in Robots*. Springer Science & Business Media.

Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S., et al. (2011). Approaching the Symbol Grounding Problem with Probabilistic Graphical Models. *AI magazine*, *32*(4), 64–76.

Tellex, S., Thaker, P., Deits, R., Simeonov, D., Kollar, T., & Roy, N. (2012). *Toward a Probabilistic Approach to Acquiring Information from Human Partners Using Language* (Tech. Rep.). MIT.

Van Deemter, K. (2016). *Computational Models of Referring: A Study in Cognitive Science*. Cambridge, Massachusetts: MIT Press.

Williams, T., Acharya, S., Schreitter, S., & Scheutz, M. (2016). Situated Open-World Reference Resolution for Human-Robot Dialogue. In *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction* (pp. 311–318).

Williams, T., & Scheutz, M. (2015a). A Domain-Independent Model of

Open-World Reference Resolution. In *Proceedings of the 37th meeting of the Cognitive Science Society* (pp. 2667–2672).

Williams, T., & Scheutz, M. (2015b). POWER: A Domain-Independent Algorithm for Probabilistic, Open-World Entity Resolution. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1230–1235).

Williams, T., & Scheutz, M. (2016). A Framework for Resolving Open-World Referential Expressions in Distributed Heterogeneous Knowledge Bases. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 3598–3964).

Winograd, T. (1971). *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language* (Tech. Rep.). DTIC Document.

Zender, H., Kruijff, G.-J. M., & Kruijff-Korbayová, I. (2009). Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants. In *Proceedings of the 21st International Joint Conference on Artifical intelligence* (pp. 1604–1609).

Notes

[1]In fact, this form, which is only used colloquially, is the only form in English that overtly cues the referential status.

[2]C.f. work in open-world directive grounding (Matuszek, Herbst, Zettlemoyer, & Fox, 2012; MacMahon, Stankiewicz, & Kuipers, 2006), in which natural language utterances are translated directly into *action sequences*, bypassing the need to ground constituent noun phrases.

[3]A co-reference resolution procedure attempts to identify whether a referring expression refers to the same referent as a previous referring expression. If so, the new RE is added to the previous RE's *co-reference cluster*; a unique identifier is used to identify the presumed referent of each co-reference cluster during subsequent language processing steps.

Biographies

Tom Williams is a Cognitive and Computer Science Ph.D. candidate in the Department of Computer Science at Tufts University. He earned a B.A. in Computer Science from Hamilton College in 2011 and an M.S. in Computer Science from Tufts University in 2013. His current research focuses on enabling natural language capabilities for intelligent agents operating in uncertain and open worlds.

Matthias Scheutz is a Professor in Cognitive and Computer Science in the Department of Computer Science at Tufts University. He earned a Ph.D. in Philosophy from the University of Vienna in 1995 and a Joint Ph.D. in Cognitive Science and Computer Science from Indiana University Bloomington in 1999. His current research focuses on complex cognitive robots with natural language capabilities.