A Domain-Independent Model of Open-World Reference Resolution

Tom Williams and Matthias Scheutz
{williams,mscheutz}@cs.tufts.edu
Human-Robot Interaction Laboratory, 200 Boston Avenue
Medford, MA 02155

Abstract

The ability to ground conversational referents is a key requirement for human dialogue. This process, known as *reference resolution*, has received much attention from both psycholinguists seeking to understand how humans process language and computer scientists seeking to improve the performance of language-capable agents. However, the majority of previous research has focused on what we term *closed-world* reference resolution, in which the set of possible referents is assumed to be known *a priori*. In this paper we present a domain-independent model of *open-world* reference resolution which appropriately handles uncertain knowledge, and the results of an empirical human-subject experiment conducted to verify the model's predictions.

Keywords: computational modeling, natural language understanding, reference resolution

Introduction

The ability to ground conversational referents is a key requirement for human dialogue. This process, known as *reference resolution*, has received much attention from both psycholinguists seeking to understand how humans process language and computer scientists seeking to improve the performance of language-capable agents. However, the majority of previous research has focused on what we term *closed-world* reference resolution, in which the set of possible referents is assumed to be known *a priori*, and there has been little consideration of what we term *open-world* reference resolution, in which entities not known *a priori* may be referenced, leading to modification of existing knowledge and/or the creation of new representations. For example, consider an agent who, after entering a building for the first time, is told:

(1) I'll be in the office across the hall from the kitchen.

The agent will need to identify which portions of the utterance refer to entities he already may know about (i.e., "the hall"), and which portions refer to entities he does not yet know of (i.e., "the office" and "the kitchen"). The agent will then need to modify his internal representation of the building's structure using the information provided about the newly mentioned entities. New information may be communicated either intentionally (i.e., if an agent says "There is a kitchen across from the breakroom"), or unintentionally due to erroneous assumptions about common ground (i.e., If an agent says "I was in the kitchen across from the breakroom" without realizing that their interlocutor is not familiar with the kitchen and/or the breakroom). In short, the agent must be able to handle *incomplete knowledge*, one of three forms of *imperfect knowledge* we identify.

One way to empirically study such human capabilities is through computational cognitive modeling of those capabilities, i.e., the development of detailed mathematical or algorithmic models which can be implemented to simulate those capabilities(Sun, 2008). The development of such models allows one to make predictions as to the processes underlying human cognition, and what is more, actually allows one to *test* those predictions by comparing their model's behavior to human behavior. Thus one way to study human mechanisms for open-world reference resolution is to develop a computational model that provides such resolution capabilities to a machine.

A model of open-world reference resolution should be able to handle at least three types of imperfect knowledge. First, it should handle *incomplete knowledge* as previously discussed, in which all candidate referents are not known a priori. Second, it should handle *ambiguous knowledge*, in which multiple candidate referents are known (e.g., if in Example 1 the listener knows of two kitchens in the current hallway). This is a general capability of all models of reference resolution, and thus will receive limited attention in this paper. Third, a model of open-world reference resolution should handle *uncertain knowledge*, in which relations and properties are not known with absolute confidence (e.g, if in Example 1 the listener knows of a room which it thinks to be a kitchen, but is not entirely sure). As we will discuss, no previous model satisfactorily models every type of imperfect knowledge.

In addition, a model of open-world reference resolution should be domain-independent. The majority of computational models of reference resolution target a specific domain, such as descriptions of objects or locations. However, while different thought processes may be employed in understanding Examples 1 and, say, "Jim's uncle is a paleontologist", a single mechanism should be used for the tasks of acquiring and arbitrating between candidate referents in both examples.

In this work, we present a domain-independent computational model of open-world reference resolutions which meets all the aforementioned criteria. As a first step, we select a challenging but tractable set of natural language utterances to model: *complex, first-mention definite noun phrases*. *Definite* descriptions in general are one of the most common forms in natural language (Brown-Schmidt, Campana, & Tanenhaus, 2002), especially in domains with a limited (i.e., tractable) number of possible candidate referents (Hanna & Tanenhaus, 2004). *First-mention* definite noun phrases (which introduce new entities into the discourse) are attractive as they exhibit the open-world aspects we seek to capture, and are known to be particularly difficult to process (Fraurud,

1990). Finally, we seek to tackle both simple and *complex* first-mention definite noun phrases, in which referents are described in relation to one or more "anchors", thus providing additional decision boundaries between known and unknown knowledge.

The remainder of the paper will proceed as follows: first, we will define our model at a *computational* level of analysis (Marr, 1982). We will then discuss the model with respect to our explicit modeling goals (i.e., handling of imperfect knowledge, and domain independence), as well as its relation to previous work. Next, we will present an empirical human-subject experiment conducted to verify our model's predictions. Finally, we will conclude with discussion of possible directions for future work.

Model Definition

In this section, we first define our model's parameters, and then define the model itself at a computational level.

Parameters

1. S: A *set of formulae* describing the semantic constraints imposed by a referential expression. For the resolution of Example 1, S might take a form such as $\{room(X) \wedge room(Y) \wedge hall(Z) \wedge office(X) \wedge kitchen(Y) \wedge across\_from(X,Y,Z)\}$.

2. M: A *world model* containing some number of atomic entities whose relationships can be described using formulae such as those contained in S. It is important to note, however, that we make no claims over the actual information storage or retrieval methods for M. For the resolution of Example 1, M might consist of a cognitive map whose atomic entities are various locations, and for which formulae such as $across\_from(X,Y,Z)$ can be assessed.

3. V: A *sequence of variables* used in S, for which each variable V_i (from i = 0 to $|V|-1$) is defined in reference to variable V_{i+1}, For the resolution of Example 1, this might be $\{X,Y,Z\}$ if it is determined that the office (X) is defined in reference to the kitchen (Y), and that the kitchen is in turn defined in reference to the hallway (Z).

Computational Model

We model the problem of open-world reference resolution as, given S, M and V, the problem of finding (1) the longest suffix Θ of sequence V for which there exists a probable mapping between variables in Θ and entities in M, and (2) the most probable mapping for Θ. A mapping is deemed *probable* if its probability (as assessed in M) is above some threshold τ. Intuitively, we wish to find the longest *suffix* because new information is typically defined relative to old information, and thus we would expect to be able to make a cut at some point in V that partitions it into two sub-sequences which contain new and old entities, respectively.

When seeking a probable mapping for sequence Θ, the model must consider the hypothesis space of possible bindings between the subset of variables from V that exist in suffix Θ and atomic entities in M. This hypothesis space is denoted as H^{Θ}. The model must evaluate the probability of each mapping in H^{Θ} on the basis of $P(s|h)$ for each s in the subset of formulae from S that use variables found in suffix Θ, denoted as S^{Θ}. For example, if no probable mapping is found when $V = \{X,Y\}$ and $S = \{room(X), hall(Y), across\_from(X,Y)\}$, the next step would be to check for a probable mapping when $\Theta = \{Y\}$ and $S^{\Theta} = \{hall(Y)\}$.

We are thus able to define our model using a set of four equations:

$$\Phi(H^{\Theta}, S^{\Theta}, M) = \underset{h \in H^{\Theta}}{\operatorname{argmax}} P(S^{\Theta}|H) \tag{1}$$

$$P(S^{\Theta}|h) = \prod_{s \in S^{\Theta}} P(s|h) \tag{2}$$

$$\Theta_j = \{V_j \circ \cdots \circ V_{|V|}\} \tag{3}$$

$$\begin{aligned} &resolve(V,S,M) = \Phi(H^{\Theta_i}, S^{\Theta_i}, M) \mid \\ &i = min\{j \mid P(S^{\Theta_j} \mid \Phi(H^{\Theta_j}, S^{\Theta_j}, M)) > \tau\} \end{aligned} \tag{4}$$

Here, Equations 1 and 2 indicate that the process Φ of selecting the best hypothesis h is equivalent to finding the hypothesis with the highest probability, which in turn is calculated by finding the sum of the probabilities of each formula in S being true under hypothesis h. The process of assessing these formula-level probabilities will differ depending on the domain of M and type of formula. For example, very different processes might be used for evaluating the probabilities of two locations being across a hall from each other and for evaluating the probabilities of two people being brothers.

Equation 3 simply serves as shorthand indicating that suffix Θ_j consists of the elements of V starting at element j. Finally, Equation 4, states that the best hypothesis overall is the best hypothesis for suffix Θ_i, where i is the smallest number such that the probability of that suffix's best hypothesis is greater than some threshold τ.

Discussion and Related Work

As stated in the Introduction, we are interested in creating a model of reference resolution which handles incomplete, uncertain and ambiguous knowledge, and which is domain-independent. We will first discuss the degree to which we have achieved each of these goals, and then discuss models which have sought to handle other aspects of reference resolution which we do not account for.

Incomplete Knowledge

Because we model open-world reference resolution as the problem of finding (1)the longest *suffix* of a variable sequence and (2) the most probable variable-to-entity mapping for that sequence, our model produces, as a side effect, a *prefix* sequence of variables which are not mapped to any entities. If an agent creates abstract representations for the unknown or hypothetical entities corresponding to these unmapped variables (using the formulae in S containing those variables), she will be able to discuss and reason about those entities

without having physically experienced them. This is a significant advancement from previous approaches, which either operate under an entirely closed-world assumption (i.e., that all possible candidates are known *a priori*) or which translate utterances directly to actions which must immediately be carried out before an agent is able to discuss or reason about the described entities (Matuszek, Herbst, Zettlemoyer, & Fox, 2012; Duvallet et al., 2014). To the best of our knowledge, this capability has only been previously achieved by (Williams, Cantrell, Briggs, Schermerhorn, & Scheutz, 2013). However, that approach makes a number of strong domain-dependent assumptions, and assumes full certainty of its knowledge.

Uncertain and Ambiguous Knowledge

Like most previous computational models (excepting, e.g., (Matuszek et al., 2012) and (Williams et al., 2013)), ours uses a probabilistic approach, and is thus able to resolve references in the face of *uncertain* knowledge. This allows a model to better arbitrate between multiple *ambiguous* candidates on the basis of certainty. However, as no existing approach to our knowledge explicitly represents an agent's *ignorance*, we believe that all current approaches fall short of the ideal. We believe that modeling of an agent's ignorance is critical, as it facilitates arbitration between exploration (e.g., through dialogue, focused attention, or physical exploration) and exploitation (i.e., choosing and acting on the most likely candidate referent). In future work, we hope to come closer to this ideal through the use of a Dempster-Shafer theoretic knowledge representation scheme. A Dempster-Shafer theoretic approach is attractive as it offers an elegant representation of uncertainty which differentiates between uncertainty from ambiguity and uncertainty from ignorance, in a way which does not require commitment to a particular probability distribution. This will also enable better integration with our Dempster-Shafer theoretic models of pragmatic analysis and generation (Williams et al., 2014; Williams, Briggs, Oosterveld, & Scheutz, 2015).

Domain Independence

Our model is not defined with respect to any particular domain. This is in contrast to most previous computational models of embodied reference resolution, which choose a particular domain to target, such as descriptions of routes (Matuszek et al., 2012; Fasola & Matarić, 2013; Kruijff, Janíček, & Zender, 2012; Duvallet et al., 2014), locations (Williams et al., 2013), interface elements (Chai, Prasov, Blaim, & Jin, 2005), or tabletop-objects (Scheutz, Krause, & Sadeghi, 2014; Kruijff, Kelleher, & Hawes, 2006).

An exception to this is the G^3 model used by (Kollar, Tellex, Roy, & Roy, 2014). This model is in principle domain independent, but must be trained on a particular chosen domain, and only models closed-world reference resolution. While Kollar et al. use beam-search for the most probable satisfaction of the variables contained in the model, we instead use best-first search, as a large number of viable candidates may exist at each step. When resolving a reference to some "room", for example, it would be imprudent to discard places that did not fall in the top ten most likely to be considered "rooms" since there may be hundreds of places that satisfy this constraint to a high degree.

Both our model and the G^3 model have the shortcoming of only handling a single domain at a time, however: G^3 must be trained on a target domain, and our model currently assumes that all entities referenced in an utterance are members of the same domain, as indicated by the use of a single world model M. The ideal model of reference resolution, on the other hand, would be able to interpret expressions such as "the man we had lunch with in that little cafe last week", which refers to entities from multiple domains. This capability is the focus of our ongoing work.

Of course, domain-independence and handling of imperfect knowledge are not the only important aspects of reference resolution which must be modeled: we will next examine models of reference resolution from the psycholinguistics literature, which have mainly focused on concerns such as incrementality.

Related Psycholinguistic Work

Among relevant psycholinguistic models of reference resolution, our work is most similar to that of (Schlangen, Baumann, & Atterer, 2009), which presents a Bayesian model of reference resolution. Under this model, a default a decision of "undecided" is maintained until a candidate with posterior probability above some *adaptive threshold* is found. In contrast, when the best hypothesis our model can find is below the threshold τ, it is treated not as "undecided", but rather as an indication that the referent of the utterance should be considered to be "new", and that the supposedly "given" portion of the noun-phrase (i.e., the referent's anchor) should be examined to determine if it too should be considered to be "new". (Here we use the "given/new" dichotomy traditionally employed at the sentence level (e.g., (Haviland & Clark, 1974; Clark, 1975))). However, (Schlangen et al., 2009), exploit the benefits provided by an incremental approach, while we do not. Much research has demonstrated the incremental nature of human language understanding(Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995), and shown how incremental language understanding facilitates fast processing and disambiguation of statements in, e.g., visual search tasks (Spivey, Tyler, Eberhard, & Tanenhaus, 2001; Krause, Cantrell, Potapova, Zillich, & Scheutz, 2013). While the incremental aspects of language processing were not the focus of our model, we aim to adapt an incremental, parallel approach like that seen in (Scheutz et al., 2014) in the future. In that work, Scheutz et al. used an incremental, parallel model of language-guided visual search. By effecting a similar approach, we could extend our model to handle the incremental aspects of natural language, increase performance through parallelization, and overall better model the cognitive processes in which we are interested.

Experiment

In the previous sections, we presented a model for domain-independent, probabilistic, open-world reference resolution of complex, first-mention definite noun-phrases, and discussed our model with respect to our modeling goals, and our model's relation to previous work. In this section, we present an empirical human-subject experiment to verify our model's predictions.

For this experiment, participants were recruited using Amazon Mechanical Turk. The pool of subjects who finished the task consisted of 40 participants (18 Male, 22 Female) with mean age 34.75. Participants were paid $2.00 to perform the task. Each participant was asked to consider three sets of referential statements. For each of the three sets of statements, they were provided with the corresponding third of the following knowledge base shown in Table 1.

ID	Name	Description
1	Jim Nelson	Doctor (**pretty sure**). Friends with Sam Greene.
2	Sam Greene	friends with Jim Nelson. **Probably** male.
3	Jim Cruz	?
4	Mary Greene	Sister of Sam Greene.
5	Frank Roberts	Jon says he's a painter, **but** Craig says he's an author **...** **?** Lives next door to Nicolas.
6	Martin Francis	Painter, lives next door to Heidi.
7	Kristy Roberts	**Might** be the daughter of Frank Roberts. **Unsure**.
8	Heidi Wilkerson	Chemist, lives next door to Martin.
9	Nicolas Morris	Chemist, lives next door to Frank.
10	Craig Horton	Chemist, **might** work with Heidi? **Probably** doesn't work with Nicolas, **but who knows**.
11	Ted Wells	Baker. **Possibly** brothers with Phillip and/or Troy.
12	Phillip Wells	Brewer. **Possibly** brothers with Ted and/or Troy.
13	Troy Wells	Byron's friend. **Possibly** brothers with Phillip and/or Ted.
14	Laurie Rodgers	Byron's friend. Girlfriend of **one of** the Wells brothers.
15	Sally Owens	Teacher. Sibling of Willie Owens. Laurie's neighbor.
16	Willie Owens	Customs officer. **Possibly** female. Sibling of Sally Owens.
17	Byron Todd	**Could be** a podiatrist **... or maybe** a pediatrician.

Table 1: Knowledge Base provided to participants. In bold are words indicating uncertain information.

Participants were told that their siblings were planning a party, and that the aforementioned list was a list of people their sister had invited. Each participant was then given a second list corresponding to each third of the second column of Table 2, and were told that each description in this list represented a description given by their brother of someone *he* wanted invited to the party, that anyone mentioned in a description needed to be invited as well, and that it was their job to determine, for each person mentioned in one of their brother's descriptions, whether or not that person already appeared on their sister's list and if so who that person was.

The sixteen referential expressions used in this evaluation specifically probed 16 conditions we will now describe.

We delineate four categories of uncertainty that can apply to the resolution of a given entity: **0**: No valid referent can be found (requiring modeling of incomplete knowledge), **0.5**: One valid but tenuous referent can be determined, and it is thus unclear whether the correct referent has been found or whether the correct referent is yet unknown (requiring modeling of uncertain knowledge), **1**: Exactly one valid referent can be found, and **2**: Multiple valid referents can be found (requiring modeling of ambiguous knowledge).

In the resolution of a referential description, these categories can apply either to the *target* (i.e., the intended referent) of a referential description or to one or more of its *anchors*. For example, in the referential description "The uncle of the doctor's brother", *the uncle* is the target, and *the doctor* and *the doctor's brother* are the anchors. Similarly, when considering the subclause *the doctor's brother*, *the brother* is the target, and *the doctor* is the anchor.

Sixteen classes of uncertainty are created by classifying referential descriptions into four classes T0, T0.5, T1, T2 based on the uncertainty status of the referential description's *target*, crossed by four classes A0, A0.5, A1, A2 based on the uncertainty status of the referential description's *anchors*.

The sixteen referential expressions we used to probe these sixteen classes of uncertainty are listed, along with their uncertainty class, in columns 1 and 2 of Table 2. For each expression, our model was provided the same knowledge encoded in logical form, with confidences attached to each statement indicative of any uncertainty associated with that statement. For example, the agent was told that Kristy was the daughter of Frank with probability 0.5. All terms used to effect these probability values are highlighted in Table 1. Our model was then provided with the same referential descriptions as were given to participants, encoded into logical form, with hand-annotated variable orderings.

Results

The results of this experiment are summarized in Columns 3-5 of Table 2. Here, Column 3 shows the most frequent human response given for each referential expression, and the result or set of equally-likely results returned by the model are shown in Column 4. In both cases, referents deemed not already on the guest-list are denoted "?". For those referents, the model added new entries to the knowledge base and updated existing entries appropriately.

Column 5 of Table 2 shows the percentage of participants whose response aligned with each model responses, with conditions in which the most frequent human response matched a model responses displayed in bold in Column 1.

The results show that in 13 of the 16 conditions (81%), the model gave the response that was most frequent among the human participants. In these cases, human responses aligned with a model-given response 71% of the time. It is important

Condition	Description given to participant	Most Frequent Human Response	Model Responses	%
A1:T1	The doctor's friend's sister	(Sister:4, Friend:2, Doctor:1)	(Sister:4, Friend:2, Doctor:1)	80.0
A2:T1	Jim's friend	(Friend:2, Jim:1)	(Friend:2, Jim:1)	60.0
A2:T0	Jim's daughter	(Daughter:?, Jim:1)	(Daughter:?, Jim:1)	47.5
			(Daughter:?, Jim:3)	37.5
A0:T0	Tabitha's mother	(Mother:?, Tabitha:?)	(Mother:?, Tabitha:?)	90.0
A2:T2	The chemist's neighbor	(Neighbor:6, Chemist:8)	(Neighbor:6, Chemist:8)	22.5
			(Neighbor:5, Chemist:9)	15.0
A0.5:T0	Craig's coworker's neighbor's son	(Son:?,Nei.:6,Co.:8,Craig:10)	(Son:?,Nei.:6,Cow.:8,Craig:10)	65.0
A0:T1	Marion's daughter Kristy	(Kristy:7,Marion:?)	(Kristy:?,Marion:?)	18.5
A0.5:T0.5	Craig's coworker's neighbor's daughter	(Daug.:?,Nei.:6,Co..:8,Craig:10)	(Daug.:?,Nei.:6,Co.:8,Craig:10)	50.0
A1:T0.5	Troy's girlfriend	(Girlfriend:14,Troy:13)	(Girlfriend:14,Troy:13)	55.0
A1:T2	The baker's brother	(Brother:12,Baker:11)	(Brother:12,Baker:11)	70.0
			(Brother:13,Baker:11)	5.0
A0:T2	The chemist, Billie's father	(Father:?,Billie:?)	(Father:?,Billie:?)	97.5
A0:T0.5	Michelle's daughter, Willie	(Willie:16,Michelle:?)	(Willie:?,Michelle:?)	5.0
A1:T0	Sally's wife	(Wife:?,Sally:15)	(Wife:?,Sally:15)	95.0
A2:T0.5	The Wells boy's girlfriend	(Girlfriend:14,Wells boy:13)	(Girlfriend:?,Wells boy:11)	5.0
			(Girlfriend:?,Wells boy:12)	2.5
			(Girlfriend:?,Wells boy:13)	2.5
A0.5:T1	Troy Wells, the podiatrist's friend	(Troy Wells:13,Podiatrist:17)	(Troy Wells:13,Podiatrist:17)	85.0
A0.5:T2	The podiatrist's friend	(Friend:13,Podiatrist:17)	(Friend:13,Podiatrist:17)	27.5
			(Friend:14,Podiatrist:17)	20.0

Table 2: Evaluation Results. (1) Each condition, (2) the expression used to probe that condition, (3) the most frequent human response for that description, (4) the model responses for that description (with multiple rows used when multiple responses were returned), and (5) the percentage of human participants who provided the same answer as the model for each model response. Cases in which the most frequent human response matched a model response are bolded in Column 1.

to note that any low percentages of agreement in these categories are not indicative of model shortcomings, but rather of diversity of human response. In addition, in four of the five cases in which the model produced multiple responses deemed equally likely, the percentages of human responses aligning with each of those responses differed by at most 10%. We discuss the fifth case below.

Overall, these results suggest that our model was successful at modeling reference resolution. We will now turn our attention, towards those few cases where human and model responses did not align: A0:T1, A0:T0.5, and A2:T0.5. All three are examples of *false negatives*, in which the model failed to find a match it thought sufficiently probable. These are strictly better than false positives in which the model is overconfident in an incorrect match.

In the first two cases, participants' answers suggested that they were willing to overlook the fact that their "sibling's" directions erroneously referenced an anchor they were not familiar with because the reference's target was uniquely identifiable by a fairly unique label. Future investigation will be needed to determine if this response was due to the use of proper nouns or due to a reliance on prior probabilities.

Finally, we discuss condition A2:T0.5, in which the most frequent human response was that "the girlfriend" referred to entry 14 (Laurie Rodgers) and that the "Wells boy" referred to entry 13 (Troy Wells), while the model instead produced three hypotheses it considered equally likely; one for each known male with the surname Wells, with "the girlfriend" considered unknown in each hypothesis. We believe that this discrepancy is due to an unintentional connection between survey questions: our guess is that readers assumed that since Laurie Rodgers was likely referenced in question 9, that she was also the referent of question 14 given the similarity between the two questions. Our model, on the other hand, performs each resolution in isolation, and thus found Laurie to be too unlikely a candidate in question 14. A similar explanation can be given for the wide difference in percentage of humans aligning with the two responses provided for condition A1:T2, which the model deemed equally likely; since Troy Wells (entry 13) was already chosen by the majority of participants as the referent for the previous question, he may have seemed to be a less likely choice. If this explanation is correct, our model's performance might improve if integrated into an embodied model able to account for environment- and dialogue-related contextual factors, as also suggested by previous psycholinguistic work (e.g., (Brown-Schmidt et al., 2002; Hanna & Tanenhaus, 2004)).

Conclusion

We have presented a domain-independent model for open-world reference resolution of complex, first-mention definite noun-phrases. We discussed our model's ability to handle uncertain, incomplete and ambiguous knowledge, and how this relates to previous models. We then demonstrated our model's ability to model the majority of a comprehensive set of resolution test cases, yielding behavior comparable to human participants.

There are several ways we hope to improve our model in the immediate future. First, we must investigate the test cases in which our model's behavior did not align with human behavior. Second, we plan to examine the performance of our model when a Dempster-Shafer-theoretic approach to knowledge representation is used, as it has proven to be an effective way to represent an agent's own ignorance. Third,

the model should be modified to simultaneously use multiple world models. This is a modification that is underway but is not has yet been fully evaluated. Finally, as previously mentioned there are a variety of suggestions from the psycholinguistic literature which would improve the performance of our model, such as a parallel, incremental examination of the semantic constraints imposed by referential expressions, and the ability to use environment- and dialogue-related context to arbitrate between candidates produced by the model.

Acknowledgments

This work was in part funded by grants N00014-11-1-0493 and N00014-14-1-0149 from the US Office of Naval Research.

References

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2002). Reference resolution in the wild: Online circumscription of referential domains in a natural interactive problem-solving task. In *Proceedings of the 24th annual meeting of the cognitive science society.*

Chai, J. Y., Prasov, Z., Blaim, J., & Jin, R. (2005). Linguistic theories in efficient multimodal reference resolution: An empirical investigation. In *Proceedings of the 10th international conference on intelligent user interfaces.*

Clark, H. H. (1975). Bridging. In *Proceedings of the 1975 workshop on theoretical issues in natural language processing.*

Duvallet, F., Walter, M. R., Howard, T., Hemachandra, S., Oh, J., Teller, S., ... Stentz, A. (2014). Inferring maps and behaviors from natural language instructions. In *ISER.*

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).

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 (IROS).*

Fraurud, K. (1990). Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, *7*(4).

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*(1).

Haviland, S. E., & Clark, H. H. (1974). What's new? acquiring new information as a process in comprehension. *Journal of verbal learning and verbal behavior*, *13*(5).

Kollar, T., Tellex, S., Roy, D., & Roy, N. (2014). Grounding verbs of motion in natural language commands to robots. In *Experimental robotics.* Springer.

Krause, E., Cantrell, R., Potapova, E., Zillich, M., & Scheutz, M. (2013). Incrementally biasing visual search using natural language input. In *Proceedings of the 2013 international conference on autonomous agents and multi-agent systems* (pp. 31–38).

Kruijff, G.-J. M., Janíček, M., & Zender, H. (2012). Situated communication for joint activity in human-robot teams. *IEEE Intelligent Systems*, *27*(2), 0027–35.

Kruijff, G.-J. M., Kelleher, J. D., & Hawes, N. (2006). Information fusion for visual reference resolution in dynamic situated dialogue. In *Perception and interactive technologies.* Springer.

Marr, D. (1982). *Vision: a computational investigation.*

Matuszek, C., Herbst, E., Zettlemoyer, L., & Fox, D. (2012). Learning to parse natural language commands to a robot control system. In *Proc. of the 13th int'l symposium on experimental robotics (ISER).*

Scheutz, M., Krause, E., & Sadeghi, S. (2014). An embodied real-time model of language-guided incremental visual search. In *Proceedings of the 36th annual meeting of the cognitive science society.*

Schlangen, D., Baumann, T., & Atterer, M. (2009). Incremental reference resolution: The task, metrics for evaluation, and a bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the 10th annual meeting of the special interest group on discourse and dialogue.*

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological Science*, *12*(4), 282–286.

Sun, R. (2008). Introduction to computational cognitive modeling. *Cambridge handbook of computational psychology*, 3–19.

Williams, T., Briggs, G., Oosterveld, B., & Scheutz, M. (2015). Going beyond command-based instructions: Extending robotic natural language interaction capabilities. In *Proceedings of twenty-ninth aaai conference on artificial intelligence.*

Williams, T., Cantrell, R., Briggs, G., Schermerhorn, P., & Scheutz, M. (2013). Grounding natural language references to unvisited and hypothetical locations. In *Proceedings of the 27th AAAI conference on artificial intelligence.*

Williams, T., Núñez, R. C., Briggs, G., Scheutz, M., Premaratne, K., & Murthi, M. N. (2014). A dempster-shafer theoretic approach to understanding indirect speech acts. *Advances in Artificial Intelligence*.