Givenness Hierarchy Theoretic Cognitive Status Filtering

Poulomi Pal (poulomipal@mines.edu), Lixiao Zhu, Andrea Golden-Lasher, Akshay Swaminathan, and Tom Williams
MIRRORLab, Colorado School of Mines
1600 Illinois Street, Golden, CO 80401 USA

Abstract

For language-capable interactive robots to be effectively introduced into human society, they must be able to naturally and efficiently communicate about the objects, locations, and people found in human environments. An important aspect of natural language communication is the use of pronouns. According to the linguistic theory of the *Givenness Hierarchy* (GH), humans use pronouns due to implicit assumptions about the *cognitive statuses* their referents have in the minds of their conversational partners. In previous work, Williams et al. presented the first computational implementation of the full GH for the purpose of robot language understanding, leveraging a set of rules informed by the GH literature. However, that approach was designed specifically for language understanding, oriented around GH-inspired memory structures used to assess what entities are candidate referents given a particular cognitive status. In contrast, language generation requires a model in which cognitive status can be assessed for a given entity. We present and compare two such models of cognitive status: a rule-based *Finite State Machine* model directly informed by the GH literature and a *Cognitive Status Filter* designed to more flexibly handle uncertainty. The models are demonstrated and evaluated using a silver-standard English subset of the OFAI Multimodal Task Description Corpus.

Keywords: Cognitive Status modeling, Natural language generation, Human-robot interaction

Introduction

As human-robot interaction becomes increasingly common, robots need to be able to talk about the objects, locations, and people in their environments in the same way humans do, to facilitate concise, easy, and unambiguous communication. To reap these benefits, just like humans, robots must be able to understand and use pronouns like *it*, *this*, and *that*. The linguistic theory of the *Givenness Hierarchy* (GH) (Gundel, Hedberg, & Zacharski, 1993) suggests that humans tend to use pronouns rather than longer referring expressions due to implicit assumptions about the *cognitive status* the referent has in the mind of their interlocutor. That is, the use of different referring forms is viewed as justified based on whether the referent is *In Focus*, *Activated*, *Familiar*, and so forth, within the current conversation. Thus, for robots to understand and generate human-like natural language they must be able to model this notion of cognitive status.

Previously, Williams and Scheutz (2019) (see also (Williams, Acharya, Schreitter, & Scheutz, 2016; Williams, Krause, Oosterveld, & Scheutz, 2018)) presented the first full computational implementation of the GH for the purpose of robotic natural language understanding, using a set of handcrafted rules informed by the GH literature. However, that approach was designed specifically for robotic natural language understanding, oriented around GH-inspired memory structures used to assess what entities are candidate referents given a particular cognitive status. In contrast, natural language generation requires a model in which cognitive status can be assessed for a given entity.

Such a model of cognitive status could either be developed as a rule-based model (not dissimilar from the rule-based approach to GH-theoretic language understanding taken by Williams and Scheutz (2019)), or could instead be developed as a statistical model which would attempt to learn to predict an entity's cognitive status from data. While in practice both rule-based and data-driven empirical models are useful (Bangalore & Rambow, 2005), data-driven models may be better able to handle unseen, uncertain situations (Bangalore & Johnston, 2003; Bangalore & Rambow, 2005).

In this paper, we thus propose (and compare to a rule-based Finite State Machine (FSM) model) the *Cognitive Status Filter (CSF)*: a data-driven probabilistic model of cognitive status, structured to be optimized for natural language generation rather than natural language understanding, trained and evaluated using a silver-standard[1] English subset of the OFAI Multimodal Task Description Corpus (Schreitter & Krenn, 2016). Specifically, the CSF seeks to predict the cognitive status for a given entity based on whether and how it has been referenced in natural language.

The remainder of this paper is organized as follows. After discussing related work on cognitive status and referring expressions, we formally define the concept of a Cognitive Status Filter. We then present the results of a crowdsourced human-subject experiment to gather the data necessary to train and evaluate this model, and compare the CSF model's performance to that of a rule-based Finite State Machine model. Finally, we discuss our results and conclude with possible directions for future work.

Related Work

The Givenness Hierarchy, originally presented by Gundel et al. (1993), consists of a nested hierarchy of six tiers of cog-

[1] This subset constitutes English transliteration of originally German dialogues.

nitive status: {*in focus* ⊆ *activated* ⊆ *familiar* ⊆ *uniquely identifiable* ⊆ *referential* ⊆ *type identifiable*}, each of which is associated with a set of referring (or pronominal) forms that can be used when referring to an entity with that status (Gundel et al., 2006; Hedberg, 2013). The hierarchical nesting here means that an entity with one status can also be said to have all other statuses lower in the hierarchy. If a target referent is *in focus*, for example, it can also be inferred to be activated, familiar, and so forth. Accordingly, a speaker's selection of a pronominal form depends on their assumptions as to the cognitive status of their target referent in the mind of their conversational partner. For example, if a speaker uses "it" to refer to an object, the listener can infer that the object being referenced must be one that is already *in focus*, whereas if a speaker uses "that $< NP >$", the speaker can only infer that the object is at least familiar (but may in fact be activated or even in focus).

The hierarchical structure of the GH is also important due to the way it parallels the hierarchical nesting of models of human memory, such as Cowan (1998)'s, in which the focus of attention is a subset of short-term memory (or working memory), which is in turn a subset of long-term memory.

The GH *coding protocol*, presented by Gundel et al. (2006), provides guidelines as to what features of linguistic and environmental context should dictate the cognitive status of a given entity. For example, this protocol suggests that an entity that is mentioned in a topic role in the preceding clause should be considered to be in focus, and that any entity that is mentioned at all should be considered to be at least activated (Gundel et al., 2006; Hedberg, 2013).

Due to the GH's popularity within the research literature, and its validation across a wide variety of languages beyond English (Gundel, Bassene, Gordon, Humnick, & Khalfaoui, 2010), many researchers have sought to computationally implement it in whole or in part, especially within the context of reference resolution algorithms. Kehler (2000), for example, use the GH to justify an approach in which elements of an interface that are highlighted are considered to be "in focus", and referring expressions that use pronominal forms are automatically resolved to those highlighted referents.

Building on this work, Chai, Hong, and Zhou (2004) proposed a probabilistic graph-matching algorithm for resolving referring expressions that are complex (involving multiple target referents) and ambiguous (involving gestures that could indicate multiple candidate referents) in multimodal user interfaces. Because this algorithm had high computational complexity, Chai, Prasov, and Qu (2006) demonstrated how the algorithm's performance could be improved using a greedy algorithm based on the theories of Conversational Implicature (Dale & Reiter, 1995; Grice, 1975) and the GH. Chai et al. combine these theories to create a reduced hierarchy: *Gesture* ⊆ *Focus* ⊆ *Visible* ⊆ *Others*, where Focus combines the "in focus" and "activated" tiers of the GH, and Visible combines its "familiar" and "uniquely identifiable" tiers. When a referring expression is processed, the relationship between referring form and status is then used to help resolve that referring expression.

Finally, while the approaches above focused on modeling of reduced versions of the GH, Williams et al. (2016); Williams and Scheutz (2019) instead presented an implementation of the full GH, through a set of rules that associated different referring forms with different sequences of actions involving all six tiers of the GH. This required, in part, four data structures corresponding to the top four tiers of cognitive statuses of the GH, while the last two tiers were instead associated with new "mnemonic actions" such as creating new mental representations (Williams & Scheutz, 2019).

In all of these previous approaches, the GH is used to justify a set of data structures used to store representations for entities that could be referred to, and to justify which of these data structures should be considered (and how) when a given referring form is used. However, while this is sensible during natural language understanding, it may not be appropriate for the purposes of natural language generation. During generation, the speaker already knows what object they wish to refer to, and do not need to search through these sorts of data structures. Instead, when a speaker decides what referring form to use to refer to a given object, we argue that they would instead start by determining the status of that object, and only then may they look through the data structure associated with that status, in order to determine what distractors must be ruled out. Critically, this requires the ability to quickly determine the cognitive status of a given entity. Accordingly, in the next section we propose an approach to this problem, which we term as *cognitive status modeling*.

Problem Formulation

We formulate cognitive status modeling as a Bayesian filtering problem. Let a dialogue D consist of a set of utterances $U_0, ..., U_n$. For object o, let $S_o^t \in \{I, A, F\}$ denote the cognitive status of o at a particular timestep t after utterance U_t (either In Focus, Activated or Familiar), and let $L_o^t \in \{N, M, T\}$ denote the linguistic status of o in utterance U_t (e.g., either not mentioned in the utterance, mentioned in the utterance in a non-topic role, or mentioned in the utterance in a topic role). Using this formalism, our goal is to recursively estimate, for a given object, the probability distribution over cognitive statuses for object o at time t:

$$p(S_o^t) = p(S_o^{t-1})p(L_o^t)p(S_o^t \mid S_o^{t-1}, L_o^t) \quad (1)$$

We define a Bayesian filter of this form as a *Cognitive Status Filter* (CSF) for a given object o. Given a set of known objects, $O = \{o_1, ..., o_n\}$, our goal is then to estimate this distribution for each $o \in O$ at each time step. To do so, we use a Cognitive Status Modeling Engine C, consisting of a set of CSFs $\{c_0, ..., c_1\}$, one for each object believed to be of a status familiar or higher within the conversation. Here, we make the simplifying assumption that the same set of objects are known to both the robot and its conversational partner, meaning that the set of all objects

with status *Uniquely Identifiable* or higher is simply the set of objects O. We assume that it is straightforward to determine whether one of these objects is or is not *Familiar* based on whether or not it has appeared in the current conversation. This allows us to model whether or not an object is of status Familiar or higher based on whether or not a CSF $c \in C$ exists for that object, and to model *which* of those statuses the object likely has, using its associated CSF.

Data Collection

The core component of our CSF model that must be learned ahead of time is the conditional probability $p(S_o^t \mid S_o^{t-1}, L_o^t)$. To learn this, we trained our model using a silver-standard English translation of the German OFAI Multimodal Task Description corpus (Schreitter & Krenn, 2016). The corpus represents a collection of human-human and human-robot interactions where the human teacher shows and explains to a human or robot learner how to connect two separate parts of a tube and then how to mount the tube onto a box with holders, as shown in Figure 1 by actually moving around the objects and performing the task while explaining it to the learner. The average length of a sentence that is used in this corpus has 8-9 words. As the name suggests, since the corpus is "multimodal", the corpus contains both verbal and non-verbal cues such as speech, gaze, and gestures. Realistic multimodal HRI scenarios require the use of such non-verbal cues; however as our first step we begin in this work by looking only at our model's ability to handle the same kind of linguistic factors that are handled by the GH, leaving the ability to model other linguistic factors for future work.

While the OFAI MTD corpus contains data from four task scenarios, we only use the data from one particular task scenario (Task 3). The original dataset for this task consists of 16 monologues each having approximately 4 to 5 utterances. As a first step, in this work we begin by evaluating our model on a small subset of the original dataset, consisting of 4 of these monologues, each of which is comprised of just 4 utterances, to control for monologue length. As shown in Figure 1, this task context contains 8 objects, including the learner and teacher.

Task 3 was selected because it includes a larger number of objects than the other tasks in a dyadic instruction context, and contains data from both human-human and human-robot dyads. Specifically, Task 1 involved a human teacher explaining and performing a task in front of the camera without the presence of a learner in the scenario; Task 2 involved a human teacher and a human learner jointly performing the task of moving an object; and Task 4 is a pure "navigation task" involving both human-human and human-robot dyads (Schreitter & Krenn, 2016).

Appearance Feature Annotation

To collect linguistic status information L, three annotators independently annotated the OFAI Multimodal Task Description Corpus (Schreitter & Krenn, 2016) according to the following annotation procedure. Each annotator was provided a printed copy of all 16 monologues to annotate. For each sentence in each monologue, the annotator was instructed to underline any piece of the text that could refer to some object in the scene. For each of these underlined pieces of text, the annotator was instructed to indicate the correspondence between the underlined sentence fragment and the object in the scene it referred to. Finally, the annotator was required to circle the fragment-object mapping they believed to be the topic of the sentence. There were a few cases in which annotators circled multiple objects as the topic of the sentence; in these cases, both objects were recorded as being equally probable topic referents[2].

Cognitive Status Annotation

Ground-truth cognitive status information was then collected through a crowdsourced human-subject experiment. 160 US participants were recruited from Amazon Mechanical Turk. Two participants answered an attention check question incorrectly and were dropped from our analysis, leaving 158 participants (71 female, 85 male, 2 N/A). Participant ages ranged from 19 to 70 years (M = 35.03, SD = 11.36). Each participant was paid $0.25 for completing the study.

Procedure: At the beginning of the experiment, each participant is shown the scene depicted in Figure 1, and is instructed to remember the objects and their labelings in order to performing their upcoming task. Participants were then shown the same scene without labels while listening to a portion of one of the experiment's four monologues, as read by the experimenters. Specifically, participants were randomly assigned to hear a random prefix of a randomly selected monologue (i.e., either only the first utterance of that monologue, the first two, the first three, or all four).

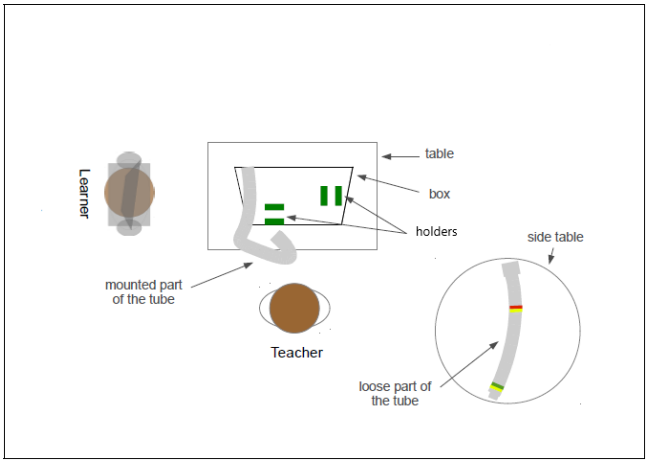

Figure 1: Scene (labeled)

At the end of this monologue excerpt, participants were

[2]The inter-annotator agreement score as measured through Fleiss' Kappa was $\kappa_n = 0.37$, indicating fair agreement between annotators. It will be important in future work to adapt the annotation protocol to increase rate of agreement.

asked to answer two questions, presented in a randomized order, with the second question becoming available after the first question was answered. The two questions are as follows:

- **Q1**: *Click on the object in the scene that you think the speaker would most likely be referring to if the speaker would have said "look at it" at the end of the monologue.*
- **Q2**: *Click on all the objects in the scene that you think the speaker would most likely be referring to if the speaker would have said "look at that" at the end of the monologue.*

Two of the monologues used in our experiment are shown below.

Monologue 1:

U1: You must take the tube with your right hand.

U2: And insert it in at the yellow-green connection here.

U3: Put it on the tube.

U4: Again, with your right hand insert it here in the holder.

Monologue 2:

U1: With the right hand stick the two tubes together.

U2: You put that together here with the yellow-green mark.

U3: It is okay that it is not holding firmly.

U4: Now lead the one tube through here.

These questions allowed us to probe the user's implicit beliefs as to the cognitive status of the objects in the scene. From a GH-theoretic perspective, if a participant implicitly believed a given object to be in focus, they should click on that particular object for both **Q1** and **Q2**, whereas if they believed the object to be activated, they should click on that object for **Q2** but not for **Q1**. Because the context is narrowly defined and participants were given time to examine each object in the scene, we assume that all objects in the scene are familiar or higher. Thus, if a participant believed the object to be familiar or lower, they should not click on the object at all. After completing the task, participants completed a check question (cf. Schreitter and Krenn (2016)) requiring users to identify the scene they had viewed from among several distractors. This allowed us to ignore data from participants who did not pay sufficient attention while completing the task.

Using this coding procedure, we are thus able to determine the perceived cognitive status of each object in the scene for each participant after the completion of the monologue excerpt they were exposed to. When paired with the linguistic status annotations, this allowed us to train our CSF model, using the procedure described in the following section.

Training and Evaluation

Training

After collecting this dataset, our CSF was trained in the following way: First, we initialized a 9x3 matrix whose rows correspond to the nine cognitive/linguistic status pairs an object could have at time $t-1$ ((I_{t-1},N_t), (I_{t-1},M_t), (I_{t-1},T_t), (A_{t-1},N_t), (A_{t-1},M_t), (A_{t-1},T_t), (F_{t-1},N_t), (F_{t-1},M_t), (F_{t-1},T_t)), and whose columns correspond to the three cognitive statuses that object could have at time t (I_t, A_t, F_t).

For each pair of adjacent utterances in each monologue (U_{t-1},U_t), we consider the data from all participants (for all objects) who provided data immediately following utterance U_{t-1}, and from all participants who provided data immediately following utterance U_t. For each resulting pair of datapoints, we identify and increment the correct cell in this matrix. For example, for the combination of a datapoint from a participant who heard some utterance and subsequently viewed that object as in focus, and a datapoint from a participant who heard the next utterance in the same monologue, containing object 1 in a non-topic role, and at that point viewed the object as being activated, we would increment the cell $((I_{t-1},N_t),A_t)$. Once all data has been considered, we normalize each row of this table to produce a conditional probability table.

Evaluation

To evaluate our CSF model, we then considered each object o and each monologue M, and retrained our model using all data except that which was collected for object o or monologue M (for example, while testing for object o_1 in monologue M_1, we retrain our model with all the data except that concerned with M_1 and/or o_1), and used this model (along with a prior distribution over cognitive statuses for that object as described below) to simulate what status would be predicted for that object at each point in that monologue. After each of these utterances, we evaluated the model's prediction by comparing it to the majority opinion from participants who *had* provided data for that object at that point in that monologue. Combining these prediction results for all eight objects in all four utterances in all four monologues produced a 128-element prediction vector for the model.

Specifically, we computed these prediction vectors for each of two CSF models, each of which used a different prior distribution $p(S_o^{t-1})$ over cognitive statuses:

U-Model: an *uninformed* prior in which each cognitive status was assigned a prior probability of 0.33.

I-Model: a (weakly) *informed* prior, in which the three cognitive statuses were assigned prior probabilities I= 0.05, A= 0.1, F= 0.85. These probabilities reflect the fact that objects are a priori far more likely to be familiar than activated, and among the set of things that are currently activated it is more likely for a given object to be activated than in focus. While in theory this distribution could be learned from data, in a realistic environment it may be the case that hundreds or thousands of objects are familiar and only one is in focus, yielding an extremely unbalanced distribution. This weakly informed prior thus represents an *optimistic* belief state in which the prior probability of any given object being in focus is artificially boosted.

In addition to these two prediction vectors produced by different parameterizations of our CSF model, we also computed prediction vectors for two baseline models:

Finite State Machine: First, we computed the decisions made by a rule-based FSM model, which formalized a set of heuristics from the GH coding protocol (the same heuristics previously used in the work of Williams and Scheutz (2019)). In this FSM, the states correspond to cognitive statuses, and transitions are triggered based on linguistic statuses observed in incoming utterances. For example, for an FSM dedicated to some object, if that object is mentioned in a topic role, this will deterministically trigger a state transition to *in focus*.

Random Baseline: Second, we computed the decisions made by a random baseline (RB) model, which predicted cognitive statuses at random.

Results

The overall accuracy of each model (i.e., the proportion of correct entries in each model's prediction vector) is shown in Table 1. This demonstrates that our U-model had the highest accuracy, and that our I-model and the theoretical FSM model had the same accuracy, slightly less than the U-model. The accuracy measure of the FSM model suggests that the heuristics encoded in the GH coding protocol are a good representation of the patterns that can be learned from the data we collected, given our choice of data annotations. The similarity of the CSF model's accuracy to that of the FSM similarly demonstrates that the CSF did a good job of automatically learning these patterns from our data. The slightly higher accuracy of the U-model over the I-model suggests that the uniformly distributed prior probabilities may have been more helpful than the weakly informed prior distribution. Finally, the performance advantage of all of these models over the RB model provides a good baseline measurement of success.

Table 1: Accuracy measure of each model

model	accuracy
U-model	82.03
I-model	81.25
FSM	81.25
RB	32.81

To validate these intuitive assessments, we formally compared our four models using six pairwise McNemar's Tests (Bostanci & Bostanci, 2013; McNemar, 1947), whose results are shown in Tables 2 and 4.

Table 2: Contingency Table entries for model pairs

$model_1$	$model_2$	N_{ss}	N_{sf}	N_{fs}	N_{ff}
U-model	I-model	104	1	0	23
U-model	FSM	89	16	15	8
U-model	RB	34	71	8	15
I-model	FSM	89	15	15	9
I-model	RB	33	71	9	15
FSM	RB	34	70	8	16

Table 2 (see also Figure 2) shows the contingency table values used by McNemar's test for each pairwise comparison, where the four N counts refer to the contingency table cells shown in Table 3. That table layout simply depicts a general 2x2 contingency table (Clark & Clark, 1999; Liddell, 1976) comparing the performance of two models A and B. Here, N_{ff} and N_{ss} respectively denote the number of instances where both models failed and succeeded. N_{fs} and N_{sf} respectively denote the instances where one model failed and the other succeeded.

Table 3: A 2X2 Contingency Table

	model A success	model A fail
model B success	N_{ss}	N_{sf}
model B fail	N_{fs}	N_{ff}

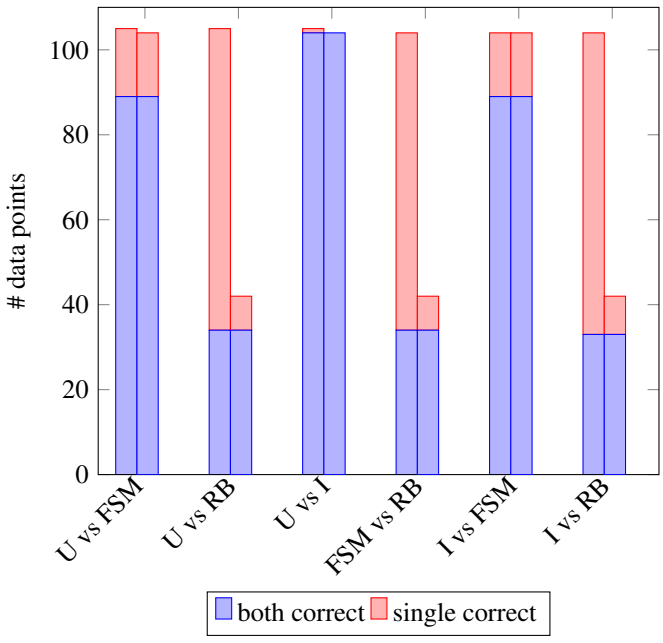

Figure 2: Comparison between models

The McNemar's Test statistics χ^2 (with 1 degree of freedom) and p-values (Bostanci & Bostanci, 2013; Fay, 2011; Liddell, 1976) are calculated for each pair of models as shown

in Table 4. By looking at the McNemar's Test results the following deductions can be formally made:

1. The U-model and I-model show similar performance ($\chi^2 \approx 0$ and p-value = 1).
2. The U-model and FSM also show similar performance ($\chi^2 \approx 0$ and p-value = 1).
3. The FSM and RB models show significant difference in their performance ($\chi^2 = 47.705$ and p-value = 0.0001).
4. The CSF model and RB model differ significantly in performance regardless of model parametrization.
5. The performance difference between the CSF model and the FSM model is not statistically significant.

Table 4: McNemar's Test statistic (χ^2) and p-values

	χ^2	p-value
U-model, I-model	0.000	1.000
U-model, FSM	0.000	1.000
U-model, RB	48.658	<0.0001
I-model, FSM	0.033	0.8551
I-model, RB	46.513	<0.0001
FSM, RB	47.705	<0.0001

Conclusion and Future Work

In this paper we present the notion of a *Cognitive Status Filter*: a statistical model for estimating the cognitive status of some entity that may be referenced in conversation. We then described a Mechanical Turk experiment used to gather ground truth data to train this model, and demonstrate how the accuracy of this model compares to a rule-based FSM model and a random baseline.

The overall accuracy of our CSF model in predicting the cognitive status of an object was slightly better than that achieved by a FSM. This simultaneously speaks in favor of the heuristics encoded in the GH coding protocol, while also demonstrating that those heuristics are learnable from data. However, there are a number of directions for future work that may significantly improve the potential performance of the statistical CSF model over the rule-based FSM model.

First, this experiment used a relatively small corpus collected in a single task; given the fact that our model works on this small dataset one follow up step would be to collect a larger dataset from a broader set of HRI scenarios (preferably a gold-standard English corpus), as that could yield a model with better generalizability. Second, our model currently only uses linguistic status information that is already explicitly called for by the subset of the GH coding protocol used to design the FSM model. However, the CSF model could straightforwardly be extended to include additional non-linguistic cues like gaze and gesture which are critical in both human-human and human-robot communication (e.g., for establishing joint attention (Moore, Dunham, & Dunham, 2014; Peeters, Azar, & Özyürek, 2014)), which although not well described in the GH coding protocol would clearly play a role in informing notions of cognitive status. Similarly, we considered only three simple linguistic features (topic mentioned, mentioned, and not mentioned) given by the GH coding protocol, whereas more complex and varied linguistic features could improve performance. Finally, one of the theoretical advantages of the CSF model is its ability to handle uncertainty. This will be critical for integrating gaze and gesture, which are inherently ambiguous and uncertain cues.

In addition, one limitation of our experimental paradigm is that users may have been coerced into selecting an object in the scene as a candidate referent for "it" (question Q1, i.e., as opposed to selecting nothing at all) even when they believed that no felicitous referent existed. This could be addressed in future work by modifying the question asked to participants in order to allow them to not select any present object if they did not believe them to be sufficiently likely candidates.

Finally, in future work, we intend to leverage our CSF model to implement a GH-theoretic anaphora generation model that uses an object's cognitive status when selecting a referring form during natural language generation. We further plan to integrate this model into the DIARC cognitive robotic architecture (Scheutz et al., 2019) and demonstrate its use in realistic HRI scenarios.

Data Availability

Our experimental data can be found at https://osf.io/qse7y/, along with our analysis scripts, experimental materials, and model outputs.

References

Bangalore, S., & Johnston, M. (2003). Balancing data-driven and rule-based approaches in the context of a multimodal conversational system. In *Proceedings of the 2003 IEEE workshop on automatic speech recognition and understanding.*

Bangalore, S., & Rambow, O. C. (2005). *Probabilistic model for natural language generation.* Google Patents. (US Patent 6,947,885)

Bostanci, B., & Bostanci, E. (2013). An evaluation of classification algorithms using McNemar's test. In *Bio-inspired computing: Theories and app. (BIC-TA).*

Chai, J. Y., Hong, P., & Zhou, M. X. (2004). A probabilistic approach to reference resolution in multimodal user interfaces. In *Intelligent user interfaces (IUI).*

Chai, J. Y., Prasov, Z., & Qu, S. (2006). Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research*, *27*, 55–83.

Clark, A. F., & Clark, C. (1999). Performance characterization in computer vision: a tutorial.

Cowan, N. (1998). *Attention and memory: An integrated framework* (Vol. 26). Oxford University Press.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, *19*(2), 233–263.

Fay, M. P. (2011). *Exact McNemar's test and matching confidence intervals.*

Grice, H. P. (1975). Logic and conversation. In *Syntax and semantics 3: Speech acts* (pp. 41–58).

Gundel, J. K., Bassene, M., Gordon, B., Humnick, L., & Khalfaoui, A. (2010). Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics*, *42*(7), 1770–1785.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Gundel, J. K., Hedberg, N., Zacharski, R., Mulkern, A., Custis, T., Swierzbin, B., ... Bassene, M. (2006). Coding protocol for statuses on the givenness hierarchy. *Unpublished manuscript (1993/2006). http://www. sfu. ca/hedberg/Coding_ for_Cognitive_Status. pdf*.

Hedberg, N. (2013). Applying the givenness hierarchy framework: Methodological issues. *International workshop on information structure of Austronesian languages*.

Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI/IAAI.*

Liddell, D. (1976). Practical tests of 2×2 contingency tables. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *25*(4), 295–304.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157.

Moore, C., Dunham, P. J., & Dunham, P. (2014). *Joint attention: Its origins and role in development*. Psychology Press.

Peeters, D., Azar, Z., & Özyürek, A. (2014). The interplay between joint attention, physical proximity, and pointing gesture in demonstrative choice. In *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 1144–1149).

Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., & Frasca, T. (2019). An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive architectures* (pp. 165–193). Springer.

Schreitter, S., & Krenn, B. (2016). The OFAI multi-modal task description corpus. In *Proceedings of the international conference on language resources and evaluation (LREC)* (pp. 1408–1414).

Williams, T., Acharya, S., Schreitter, S., & Scheutz, M. (2016). Situated open world reference resolution for human-robot dialogue. In *Proceedings of the international conference on human-robot interaction (HRI)* (pp. 311–318).

Williams, T., Krause, E., Oosterveld, B., & Scheutz, M. (2018). Towards givenness and relevance-theoretic open world reference resolution. In *RSS workshop on models and representations for natural human-robot communication.*

Williams, T., & Scheutz, M. (2019). Reference in robotics: A givenness hierarchy theoretic approach. *The Oxford Handbook of Reference*.