RECS: LSTM-Based Cognitive Status Estimation for Human-Robot Interaction

Mark Higger
mhigger@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Zhuoyi Wang
zhuoyiwang@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Polina Rygina
prygina@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Lara Ferreira Bezerra
lbezerra@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Logan Daigler
logandaigler@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Zane Aloia
Colorado School of Mines
Golden, Colorado, USA

Sheena Wu
Colorado School of Mines
Golden, Colorado, USA

Ishani Pandey
Colorado School of Mines
Golden, Colorado, USA

Amanda Chen
amchen5@wisc.edu
University of Wisconsin
Madison, Wisconsin, USA

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

Abstract

Effective communication is critical to the success of many types of human-robot interaction. A key capability for enabling effective communication is accurate modeling of the cognitive status that entities hold (e.g., modeling what objects interlocutors are currently thinking about, or are generally aware of). However, existing models of cognitive status estimation are not well suited for situated and embodied interactions, as they do not account for nonverbal cues, which are a key way in which humans moderate cognitive status. To address this gap, we make three primary contributions. First, we introduce the BOWTIE corpus of dialogues from a multi-modal open-world referential task, annotated with cognitive status, gesture type, linguistic roles, and grammatical roles of entities across utterances. Second, we introduce RECS, the first LSTM-based model of cognitive status, which we train on the BOWTIE corpus. Third, we present empirical evidence for the success of RECS. These contributions stand to accelerate the future use of cognitively informed algorithms for robot language understanding and generation.

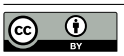
CCS Concepts

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; *Natural language generation*; Language resources; **Cognitive science**; **Cognitive robotics**; **Reasoning about belief and knowledge**; *Neural networks.*

Keywords

Human-Robot Interaction, Cognitive Modeling, Natural Language Generation, Natural Language Understanding, Gesture

ACM Reference Format:
Mark Higger, Zhuoyi Wang, Polina Rygina, Lara Ferreira Bezerra, Logan Daigler, Zane Aloia, Sheena Wu, Ishani Pandey, Amanda Chen, and Tom Williams. 2026. RECS: LSTM-Based Cognitive Status Estimation for Human-Robot Interaction . In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3757279.3785597



This work is licensed under a Creative Commons Attribution 4.0 International License.
HRI '26, Edinburgh, Scotland, UK
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3785597

1 Introduction

Generation and understanding of referential language is a key task for many types of human-robot interactions [2, 32]. One of the most important enabling capabilities for generating and understanding referential language is the ability to ground language used in an interaction to the real world. That is, for a robot to communicate about a task, it must be able to make explicit associations between real-world entities and the language that is used throughout the task. Importantly, this is true for both robots understanding references made by humans (that is, they must be able to identify what known or unknown entity is being described by humans), as well as generating references that humans can understand (that is, they must be able to describe entities in ways that are grounded in their own knowledge of those entities, and in ways that will enable their human interactants to effectively understand their descriptions).

In robots, language referring to locations, objects, people, and goals is typically grounded into formal, symbolic representations. While there has been some work in robotics that avoids these representations and instead favors end-to-end policies for translating language inputs to low-level actions and joint states [20], there are many good reasons why most roboticists continue to leverage formal representations [6]: (1) they are readily produced by modular components of robot architectures (e.g., topological SLAM systems, person recognition modules), (2) they provide clean hooks for explainability and safety reasoning, (3) they help reduce the search space over actions, and, critically, (4) they allow for clean, precise representation of the things in the world that are of most importance and interest to human tasks.

Recently, there has been some discussion in the robotics community as to whether LLMs and VLMs are heralding in a future in

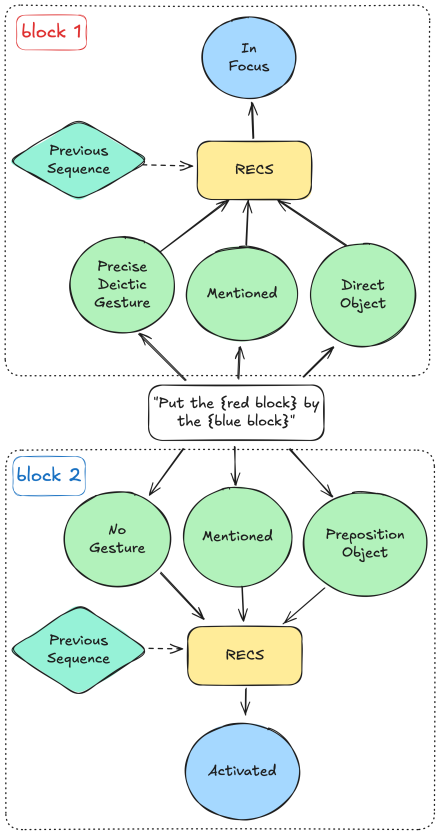


Figure 1: Overview of the cognitive status estimation pipeline. First, gestural, linguistic, and grammatical features are extracted for each entity from an utterance. For each entity, these features are then fed into RECS alongside the sequence of all previous features, and estimated cognitive statuses are input into the RECS estimator. Finally, the RECS estimator returns an estimate of the updated cognitive status.

which roboticists fundamentally shift away from symbolic representations, toward end-to-end policies. However, as Cohen et al. point out [6], the advent of high-powered foundation models need not signal a shift to *replace* symbolic representation, as they may instead *augment* symbolic representations by allowing grounding from free-form language to symbolic representation more easily than ever before. However, the use of foundation model driven language grounding may create new challenges, as the ability to flexibly ground language to a broader, more flexible, and open ended set of symbols fundamentally increases the space of possible symbols that language might be ground to, creating computational complexity challenges. This leads to a key research challenge: how can robots effectively restrict the space of symbolic representations under consideration during language understanding and generation tasks in large-scale, open-world domains?

One promising and intuitive approach is to track the entities (e.g., objects, people, and locations) likely to be of interest to humans at any given point in time, and attempt to restrict reasoning to those entities alone. For example, consider a robot working in a hospital. If a human asks the robot to get "the green linens on the shelf", the robot may know of thousands of individual pieces of linen it has seen on shelves throughout the hospital environment — yet if there is a nearby shelf, the robot may only need to even begin to consider linens residing on that shelf — especially if the robot and human had already been talking about that shelf. Similarly, if the robot wishes to talk about those linens, e.g., to ask a clarification request, it will need to describe them in a way that picks them out unambiguously — but that criteria of nonambiguity need only apply to that same set of relevant and interesting items. So, for example, a robot might respond "those green linens next to the blue ones?" even if it were aware of other linens in a similar configuration elsewhere in the hospital. The ability to establish this set of relevant and interesting entities, then, stands as a key challenge for enabling efficient natural language understanding and generation in task-based domains where language is grounded to objects, locations, and people distributed throughout the human and robots' shared environment. How, then, can roboticists best enable this capability?

Over the past several decades, a number of proposals have been made in linguistics that stand to answer this question, including Relevance Theory [37], Accessibility Theory [1], and the Givenness Hierarchy. Of these (for reasons we will describe later on), the Givenness Hierarchy has received by far the greatest attention in Human-Robot Interaction. Indeed, a wide range of language understanding and generation algorithms for task-oriented HRI have been premised on the ability to effectively model cognitive status. For example, Chai et al. [3, 4], Kehler [21], and Williams et al. [34, 36] have all used cognitive status to inform which entitites are most relevant for robot reference resolution algorithms; Pal et al. [26], Han et al. [14, 15], and Del Castillo et al. [8], have all used cognitive status to inform robot referring form selection algorithms; Spevak et al. [31] and Han et al. [13] have used cognitive status to inform more efficient robot instruction sequencing; and Higger et al. [17] has used cognitive status to inform the set of relevant entities used for referring expression generation. However, these algorithms all have a fatal flaw: they assume the ability to accurately assign a presumed cognitive status to entities, but do not actually leverage such a method. Instead, all of these previous approaches either assume oracle knowledge of cognitive status (e.g., as annotated in the dataset used for evaluation), or, when they do actually computationally endeavor to predict cognitive status, use oversimplistic models that are discrete, are reliant only on simple linguistic features, and which do not take into account non-verbal cues, despite their central influence on cognitive status estimation (e.g., the form "this <Noun Phrase>" is almost always (immediately) preceded by a pointing gesture in grounded, task-based contexts — and it is that pointing gesture that fundamentally enables the use of that ambiguous form). This leaves a clear research gap: Cognitive status is critical for task-based referential language, and there are many existing language generation and understanding algorithms which depend on having cognitive status information, yet there are no existing computational models that can actually predict cognitive status with meaningful accuracy.

In this work, we address this gap by making three key contributions. First, we present the BOWTIE (Block-based Open-World Task Instruction and Execution) Corpus of human-human interactions,

annotated with linguistic, gestural, and cognitive status features. Second, we introduce RECS, the first LSTM-based multimodal model of cognitive status, which we train on the BOWTIE corpus. Finally, we present empirical evidence for the success of RECS, as compared to several alternate models, and ablations thereof. Through our empirical evaluation of RECS on the BOWTIE corpus, we demonstrate that RECS is able to predict cognitive status with high accuracy. Our results thus represent a substantial step forward for the modeling of cognitive status, with significant potential across computational cognitive science, cognitive robotics, and human-robot interaction.

2 Related Work

2.1 Pragmatic Theories of Reference

For language generation and understanding in large-scale, open-world domains, robots must efficiently restrict the space of entities under consideration for both language understanding and generation. In order to understand how we might meet this challenge effectively, we can turn to psycholinguistic theories of how humans produce reference. One salient option is relevance theory [37], a Gricean [11] theory of pragmatics that asserts that language can be effectively decoded because speakers are able to make assumptions on the basis of Grice's cooperative principle and conversational maxims, with primacy given to Grice's maxim of relevance. Relevance theory is generally concerned with the interpretation and uptake of Speech Acts [30], on the basis of what is relevant to the "beliefs, doubts, hopes, wishes, plans, goals, intentions, [and] questions" that define a particular context [37], and in deriving an *overall interpretation* of an utterance this process (following the relevance-guided comprehension heuristic). This will ultimately entail resolving references on the basis of relevance, and for the purpose of identifying the sentence interpretation that best aligns with an expectation that the speaker intends to be maximally relevant.

Alternatively, we could turn to Accessibility Theory [1], which argues that references are not understood and generated *merely* on the basis of what is necessarily most relevant (or expected to be most relevant), but rather on the basis of what is most cognitively accessible (or expected to be most accessible). As such, Accessibility Theory is presented as a *complement* to relevance theory, with references likely to be resolved (or generated) in ways that account for how accessible information is, with accessibility determined on the basis of factors like linguistic distance, salience, and so forth.

The theory that has been most readily uptaken by roboticists over the past ten years, however, is Gundel's theory of the Givenness Hierarchy [12] — a theory that, like accessibility theory, assumes that language understanding and generation is guided by a notion of the cognitive status of the mental representations to which language might be grounded. But whereas accessibility theory posits a continuous (and largely linguistically determined) spectrum of degrees of accessibility, the Givenness hierarchy instead posits a discrete set of nested levels of cognitive status (In Focus, Activated, Familiar, Uniquely Identifiable, etc.), which roughly align with different mnemonic cognitive structures (the Focus of Attention, Working Memory, Long Term Memory, and so forth), as described in more detail in the next section. As such, the Givenness Hierarchy, when viewed from a robotics perspective, essentially asserts a hierarchical caching structure, with speakers and listeners generating and interpreting references in a way that only requires disambiguation with respect to the other entities stored in the same (or higher) level of cache as the target referent. Due to the Givenness Hierarchy's straightforward psychological interpretation and computationalizability (at least, as we will see, in terms of levels-of-cache), it has seen heavy uptake within the HRI community. However, as described above, no effective methods have been proposed for estimating cognitive status, especially from multimodal data. In the following sections, therefore, we will examine the specific nature of Givenness Hierarchy theoretic Cognitive Status, as well as the few prior attempts that have been made to computationally estimate it.

2.2 Cognitive Status

The concept of cognitive status originates in Gundel, Hedberg, and Zacharski's linguistic theory of the *Givenness Hierarchy* [12]. The Givenness Hierarchy [12] theorizes six hierarchically nested statuses that a referent may occupy:

(1) In Focus: The entity is at the center of attention.
(2) Activated: The entity is represented in working memory.
(3) Familiar: The entity is represented in long-term memory.
(4) Uniquely Identifiable: The entity can be identified uniquely (either by being located in memory or by creating a new mental representation).
(5) Referential: The entity has a unique representation in the speaker's memory.
(6) Type Identifiable: A type of entity can be identified.

According to the Givenness Hierarchy, each of these cognitive statuses is cued by a different set of referring forms: when a speaker uses *it*, they are (subconsciously) signaling that they believe their referent to be 'In Focus' for their interlocutor; when they use *this*, they similarly signal that they believe their referent to be 'Activated', and so forth. Critically, this signaling then allows the listener to more effectively identify their target referent by limiting their search to particular cognitive structures. When an 'In Focus'-cuing form is used, the listener can assume that the entity being referenced is something their interlocutor believes to be in their *Focus of Attention*; when an 'Activated'-cuing form is used, the listener can assume that the entity being referenced is something their interlocutor believes to be in their *Working Memory*; and when a 'Familiar'-cuing form is used, the listener can assume that the entity being referenced is something their interlocutor believes to be in their *Long Term Memory*. Similarly, a speaker may rely on their estimates of an entity's cognitive status to select referring forms that provide similar benefits to their interlocutor. Thus, the Givenness Heirarchy describes a set of hierarchical cognitive statuses that provide a direct mapping between the choice of referential language that interactants can use, and have used, to their understanding of each other's level of attention to the referents. Due to these benefits, the Givenness Hierarchy has been used to inform a variety of state-of-the-art robot language understanding and generation algorithms [14, 17, 26, 31, 35].

2.3 Cognitive Status Estimation

The only previous computational model for cognitive status estimation that we are aware of is the Bayesian Filter presented by Pal et al. [27]. This approach leveraged a transliterated version of

a German corpus [29], for which the authors annotated cognitive status and topic versus non-topic mentions of objects. While some promising results have been demonstrated through this approach, it nevertheless suffers from three major limitations.

The first major limitation is that the corpus used to train the approach included a very small number of objects, all of which were at least 'Familiar' in cognitive status. That is, the corpus did not include any objects that were not immediately visible and known to both conversational partners. This is a significant problem for robotics applications, as robots may frequently need to discuss entities that their interlocutors do not have prior knowledge of, or which they are likely to have forgotten. This suggests a need for a new corpus that includes entities that have a maximum cognitive status of 'Uniquely Identifiable', and for computational models that can handle this lower status.

The second major limitation is that the computational model used for estimation was an extremely simple model that was not able to capture nuanced relationships between features or long-distance temporal dependencies. Given the complex relationships between language and cognition, we believe that this simplicity drastically limited the effectiveness of the model. This motivates a need for a more flexible model, such as a Recurrent Neural Network (RNN), which is capable of accounting for these complexities.

The final major limitation is that the feature space of this previous model does not take into account nonverbal cues such as gestures. Yet gestures are a vital component of both human-human communication [9, 22, 28], and human-robot communication [2, 32]. Gesture is not only used by humans to direct visual attention [24], but also affects how entities are represented in memory[10, 28, 33]. This critical link between memory, communication, and gesture means that to accurately estimate cognitive status, gesture must be taken into account. Accordingly, we argue that Pal's model is missing vital information that might drastically impact cognitive status estimation. This suggests a need for a new corpus that includes gestural information, and for computational models that can handle these gestures.

2.4 Hypotheses

H1: An RNN will be able to better predict cognitive status than the previous models used for cognitive status estimation.

H2: Accounting for gesture will increase the performance (i.e. F1, accuracy, precision, recall) of cognitive status prediction.

3 The BOWTIE Corpus

The BOWTIE (Block-based Open-World Task Instruction and Execution) Corpus consists of annotations for twelve videos collected by Han et al. [14], comprising 337 minutes of human-human interactions with an average of 28 minutes per video. These videos were collected in a task environment partitioned into four quadrants, each containing a variety of colored blocks (Fig. 2). Within this environment, pairs of participants (an instructor and a learner) performed a sequence of four building tasks in which the instructor instructed the learner in how to construct a specific building from the blocks. Notably, there is no robot present in this task, and all instructions are generated from a single human participant to a different human participant within each video collected. Each



Figure 2: Dataset collection environment seen in [14]

building task required some blocks that were unavailable in the current quadrant, such that the instructor was required to refer not only to blocks that were immediately visible, but also to blocks that had been seen in previous quadrants, and to blocks whose locations were as-yet-unknown. In this way, the data collected in this task represents a high-quality dataset of references to objects with diverse cognitive statuses within the Givenness Hierarchy. However, the data collected by Han et al. only consisted of raw video (not publicly released) and transcripts. In our work, we present the BOWTIE corpus, which augments that raw data with additional annotations. Specifically, this dataset does not, as originally released, contain cognitive status labels, nor the linguistic or gestural features needed to predict those labels. These features are critical to enabling a robot to estimate cognitive status. Thus, to create the BOWTIE corpus, we annotated Han et al.'s original dataset with these additional features and labels.

A 270-utterance subset of Han et al.'s dataset (9 videos with 30 utterances each) was selected for further annotation. Of this subset, 11% of the utterances were recoded by a pair of coders, with inter-rater reliability (IRR) quantified using Cohen's Kappa Coefficient [5], and the remainder were recoded by individual coders. Each utterance included the following information for each of the 72 distinct blocks contained within Han et al.'s task environment:

Cognitive Status – The cognitive status of each entity in each utterance was annotated manually using the following methodology: Each time the participants moved to a new quadrant (at the start of a new structure's construction), all blocks within that quadrant were assigned a status of 'Familiar'. All other blocks were assigned the status of 'Uniquely Identifiable'. For each utterance, a team of three coders manually coded each block that was either mentioned within that utterance or any block with an assigned status of at least 'Activated' in the previous utterance. If a block was referred to by any means, the object was assigned a status of at least 'Activated'. If the coders agreed that the target was at the center of attention the block was considered 'In Focus'. The coders removed the 'In-Focus' status from blocks when they were no longer at the center of attention. It was assumed that at most, a single entity (which could consist of multiple blocks grouped together) could be at the center of attention and thus have the 'In Focus' status for a given utterance. Blocks would have the 'Activated' status removed

based on a variety of multi-modal factors such as language choice, movement, time since the last reference, and number of blocks currently assigned a particular cognitive status. Cognitive statuses of 'Referential' and 'Type Identifiable' were never used, as all objects were presumed to be at least 'Uniquely Identifiable'.

Linguistic Status – The linguistic status of each entity in each utterance was manually annotated (IRR = 0.85) using the following methodology: The linguistic status of an entity was coded as either mentioned (*M*) or not mentioned (*N*). This status was found by two coders independently determining which block or blocks were referred to by the participants. For a given utterance, blocks that were referred to were coded as *M* and all other blocks were coded as *N*.

Grammatical Role – The grammatical role of each entity in each utterance was extracted using SpaCY2 Part of Speech (POS) tagging [19]. The Grammatical Role for mentioned entities was assigned one of 6 categories: nominal subject ('*nsubj*'), direct object ('*dobj*'), preposition object ('*pobj*'), conjunct ('*conj*'), or Not Identified ('*NI*') for entities whose grammatical role were not successfully extracted. Additionally, any entities which were not mentioned were assigned the 'None'('*N*') category.

Gesture Type - The type of gesture used to refer to each entity in each utterance was manually annotated (IRR = 0.76). Each time an entity was gestured to, the entity was annotated with a gesture type in accordance with the 5 gestures presented in Higger et. al. [16]. This taxonomy was chosen as it offers a comprehensive taxonomy for open-world referential gestures. The 'Diectic Precise'(*DP*) gesture type was assigned for entities where a precise pointing gesture was used to gesture to exclusively that entity. The 'Deictic Small Region' (*DS*) gesture type was assigned for the mentioned entities for which an ambiguous pointing gesture was used to gesture to the target entity, as well as other nearby entities. The Deictic Large Region (*DL*) gesture type was assigned for the mentioned entities for which a large sweeping gesture was used to gesture to many entities. The Iconic (*I*) gesture type was assigned for the mentioned entities for which an iconic gesture was used to mimic attributes (such as shape) of the entity. The Abstract (*A*) gesture type was assigned for the mentioned entities which a non-directed pointing gesture was used to refer the target entity.

The BOWTIE Corpus produced by this annotation procedure can be found at https://osf.io/xpujy/overview

In the next section, we describe the cognitively-informed algorithm trained on this corpus.

4 RECS: A Cognitive Status Estimator

The Estimator for Cognitive Status (RECS) is a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) that has been trained on the BOWTIE corpus. The RECS estimator differs substantially from previous cognitive status estimators in three major ways:

(1) RECS allows for the Uniquely Identifiable cognitive status, so that unfamiliar entities can be tracked;
(2) RECS is a non-Markovian estimator which can take into account memory decay;
(3) RECS takes into account the use of gestures, which heavily influence cognitive status.

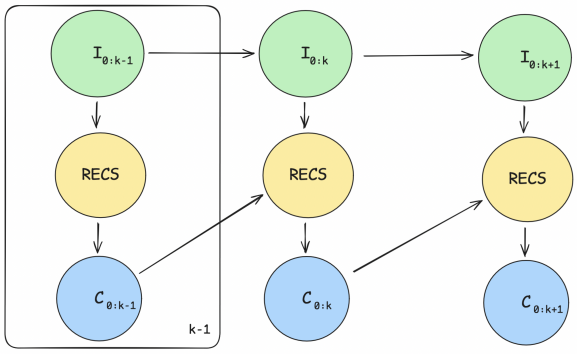


Figure 3: Plate diagram for cognitive status estimation of a single entity. At each iteration, a sequence of all previous inputs ($I_{0:k}$) and all previously estimated cognitive statuses ($C_{0:k-1}$) are used as inputs to the RECS model in order to estimate the next cognitive status (C_k).

Like other cognitive status estimators, RECS is used after each utterance in a human-robot dialogue to estimate the effect that utterance had on the cognitive status of each task-relevant entity a plate diagram for RECS is shown in figure 3. We can thus formally define the cognitive status estimation problem:

Problem Statement: Let $C_{e,0:N}$ be the cognitive status for entity e across all N utterances of an interaction. Then, let the linguistic status, grammatical role, and gesture be inputs ($I_{e,0:N}$) of the cognitive status of entity e across all utterances. We can then define the cognitive status estimation as an estimation problem, where at utterance k, the cognitive status of an entity at that utterance ($C_{e,k}$) is dependent on all the cognitive statuses of the entity at all previous utterances ($C_{e,0:k-1}$) and the current and previous inputs of that status ($I_{e,0:k}$). We can thus define cognitive status estimation as a function that maps $C_{e,0:k-1}$ and $I_{e,0:k}$ to a new cognitive status $C_{e,k}$. With that in mind, tracking all cognitive status estimates and all inputs over an unbounded amount of utterances is not computationally feasible. We thus limit the estimator to use, at most, the previous 14 cognitive status estimates and inputs, which is the maximum number of contiguous utterances in the BOWTIE corpus. The cognitive status estimation problem is then defined as the following:

$$C_{e,k} = f\left(C_{e,k-14:k-1}, I_{e,k-14:k}\right)$$

5 Experiments

5.1 Testing Dataset

To train RECS and our baseline models, we use the BOWTIE corpus, reserving 20% of the dataset for testing, 72% of the data for training and 8% of the data for validation. Notably, because entities with a status of 'Uniquely Identifiable' or 'Familiar' will only change status when either referred to or when participants move quadrants, the vast majority of entities will never change cognitive status until the building sub-task is complete. For this reason, extremely simplistic models (such as a model that always predicts current cognitive as equal to the previous cognitive status, regardless of inputs) will still achieve near-perfect accuracy when a large number of entities are

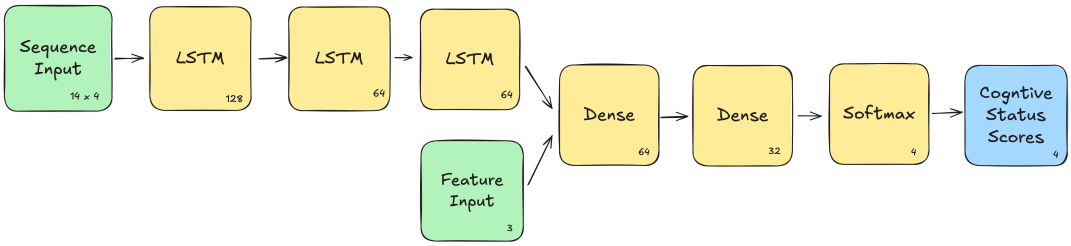


Figure 4: RECS model architecture

present. Thus to highlight differences between model performances for the testing dataset, we included only entities whose cognitive statuses changed within a contiguous set of utterances.

5.2 RECS Architecture

Shown in Figure 4, the architecture consists of three LSTM layers and two fully connected layers using the ReLU activation function, with the input features encoded as integers. To generate class scores, the output layer uses a softmax function. In order to reduce over-fitting, Dropout layers were used after each primary computational unit (the Dense and LSTM layers). Loss was calculated using categorical cross entropy [38] and the network weights were updated using the ADAM optimizer [23].

5.3 Baselines

In this work, we evaluate the performance of RECS compared to two key baselines. The first is a Bayesian Filter (BF) similar to that used by Pal et al. [27], but using the same features and labels as used by RECS and re-trained on the BOWTIE corpus. This baseline gives us a direct comparison form the RECS model to previously existing models. Additionally, given the computational complexity of RECS, we use an additional baseline, which is model-free, but less computationally complex, to demonstrate that this computational complexity is warranted. For this baseline, we use a first-order Markovian Feed-Forward Neural Network (FNN), which uses the previous cognitive status in addition to the inputs at each utterance. The FNN uses four densely connected layers with identical activation, loss and optimization to RECS.

5.4 Approach

We performed a head-to-head comparison between our three models (RECS, BF, FNN) followed by a feature ablation analysis. Feature ablation is a technique used to analyze the importance of different features in a machine-learning model by systematically removing or modifying them and observing the impact on performance. In our feature ablation analysis we thus examine the performance of each of the 15 versions of RECS associated with the removal of each combination of features.

6 Results and Discussion

Confusion matrices for the three models can be found in Figure 5, and statistics on F1 score, accuracy, precision, and recall are presented in Table 1. The results of the ablation analysis are shown in Table 2, which also includes the base model for RECS with no ablated features. We will now step through these results through the lens of our experimental hypotheses.

Model	Accuracy	Precision	Recall	F1
RECS	**0.885**	**0.871**	**0.884**	**0.875**
FNN	0.839	0.751	0.740	0.739
BF	0.727	0.806	0.617	0.629

Table 1: Model Comparison

6.1 Assessment of Hypotheses

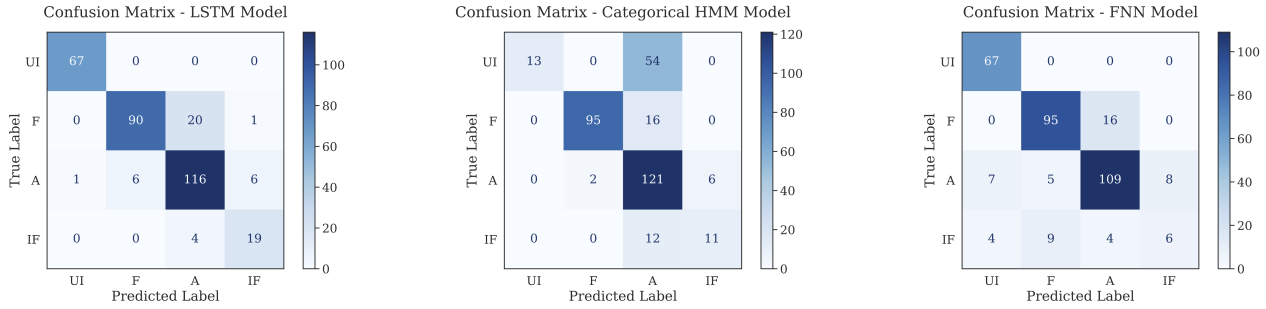
H1: Our first hypothesis (H1) was that RECS would more accurately predict cognitive status than previous cognitive estimators. RECS significantly outperformed the Bayes Classifier in terms of F1 score (0.875 vs 0.629), accuracy (0.885 vs 0.727), precision (0.871 vs 0.806), and recall (0.884 vs 0.617), supporting H1. We believe that this effect is due to the ability for RECS to account for more complex underlying relationships of cognition.

Additionally, we tested RECS against an FNN. While the accuracy was similar (0.885 vs 0.839), RECS had a significantly improved F1 score (0.875 vs 0.739), Precision (0.871 vs 0.751), and recall (0.884 vs 0.740). This performance difference is unsurprising given that RNNs have the capability to model the long-term dependencies which are required for accurate state estimation [7, 18, 25].

H2: Our second hypothesis (H2) was that accounting for gestures would increase the accuracy of the cognitive status predictions. The results of our ablation study (found in Table 2) distinctly demonstrate the important role of gesture in our model, supporting H2. Specifically, the F1 score of the full model (0.88) is nearly 20% higher than the model without gestures (0.71). However, gesture does not seem to be substantially more important than any of the other features used by our model.

6.2 Implications for Robotics Community

Multimodal understanding of common ground is critical for language-capable robots[4]. Notably, much HRI research predicates itself on having accurate models for estimating common ground (specifically, the Givenness Hierarchy cognitive status)[14, 17, 31, 35], for which existing methods are not sufficient. This leaves a clear gap within the HRI community. RECS represents an accurate cognitive status estimator that could be readily used to provide the cognitive status required as input to cognitive-status informed algorithms for robot language understanding and generation, which have not



(a) RECS confusion matrix **(b) BF confusion matrix** **(c) FNN confusion matrix**

Figure 5: Confusion matrices for RECS and baseline models. Any entities whose cognitive status did not change within the corpus is excluded.

C_{0-k}	Gest.	Ment.	Gram.	All	IF	A	F	UI
No Features Removed								
				0.88	0.78	0.86	0.87	0.99
One Feature Removed								
			X	**0.74**	**0.54**	**0.74**	0.67	**0.99**
		X		0.65	0.33	0.57	0.73	**0.99**
	X			0.71	0.40	0.70	**0.77**	**0.99**
X				0.61	0.47	0.64	0.56	0.78
Two Features Removed								
		X	*X*	0.68	0.54	0.61	0.68	0.91
	X		*X*	0.76	0.61	0.70	0.74	**0.99**
X			*X*	0.68	0.64	0.73	0.58	0.78
	X	*X*		**0.78**	0.59	0.76	**0.76**	**0.99**
X		*X*		0.72	0.60	0.82	0.66	0.78
X	*X*			0.77	**0.76**	**0.88**	0.67	0.78
Three Features Removed								
	X	*X*	*X*	0.43	0.23	0.35	0.53	0.61
X		*X*	*X*	0.48	0.44	0.38	0.50	0.61
X	*X*		*X*	0.40	0.23	0.43	0.42	0.50
X	*X*	*X*		**0.64**	**0.53**	**0.68**	**0.57**	**0.78**

Table 2: Ablation Results: F1 scores for each of the ablation models. The F1 score is shown for each cognitive status, as well as all cognitive statuses combined. '*X*' indicates features ablated for that model. Macro F1 scores are used due to imbalances in the dataset. The best performing scores for each quantity of ablated features are bolded.

been previously run on-line [14, 17, 35]. Importantly, RECS can be informed by features that are either perceived by – or generated by – the robot. That is, a robot using RECS may estimate the cognitive status based on its assumptions of what objects have been mentioned (as indicated by reference resolution algorithms), or it may do so based on its ground truth knowledge of the entities that it itself has objectively mentioned. Future work should demonstrate the effects of RECS when deployed on interactive robots for both language understanding and generation in this way. Beyond the contributions represented in RECS itself, our work also makes key contributions through our introduction of the BOWTIE corpus. To the best of our knowledge, BOWTIE is the only corpus annotated with verbal, nonverbal, and cognitive status features.

6.3 Limitations and Future Work

The primary limitation of our work is that the BOWTIE corpus is very small in comparison to the usual corpora used for training deep learning algorithms, comprised of (the transcripts of) 9 videos of 30 utterances each, annotated with a relatively small (albeit novel) set of features. Future research will be needed to expand BOWTIE to encompass a larger fraction of Han et al.'s original dataset, and to include a wider range of features. Similarly, while we expect our cognitive status estimation results to replicate for tasks different from the one used in the BOWTIE corpus, the lack of additional corpora make this impossible to verify. Future research is needed to

develop similar corpora in other task environments, and to replicate our analyses in those new domains.

A second key limitation of RECS is that it does not differentiate between speaker-mentions and robot-mentions, nor between speaker-gestures and robot-gestures. It is likely (but uncertain) that these could differently impact cognitive status. Future research is needed to extend BOWTIE (and thus RECS) to differently account for these behaviors, and to examine whether and to what extent they differently impact cognitive status.

Finally, the third limitation is that we assume the robot has knowledge of each of the input features for RECS. While this assumption is not limiting for inputs from the robot's own utterances, where the robot has oracle knowledge of its own language and gestures, this means that future research is needed to extract and ground the language and gestures produced by human interactants.

7 Conclusion

In this research, we significantly advance the state of the art of cognitive robotics through three key contributions. First, we introduce the BOWTIE corpus of dialogues from a multi-modal open-world referential task, annotated with cognitive status, gesture type, linguistic roles and grammatical roles of entities across utterances. Second, we introduce RECS, the first LSTM-based model of cognitive status, trained on the BOWTIE corpus, that can be used to enable a wide range of cognitively-informed algorithms that are critical for HRI. Third, we present empirical evidence for the success of RECS. Overall, we believe that RECS will play a vital role in the future of human-robot communication by enabling cognitively-informed interactions.

Acknowledgments

This work has been supported in part by the Office of Naval Research grant N00014-21-1-2418.

References

[1] Mira Ariel. 2001. Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects* 8 (2001), 29–87.
[2] Rehj Cantrell, Paul Schermerhorn, and Matthias Scheutz. 2011. Learning actions from human-robot dialogues. In *2011 RO-MAN*. IEEE, 125–130.
[3] Joyce Yue Chai, Zahar Prasov, and Shaolin Qu. 2006. Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research* 27 (2006), 55–83.
[4] Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *2014 HRI*. 33–40.
[5] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[6] Vanya Cohen, Jason Xinyu Liu, Raymond Mooney, Stefanie Tellex, and David Watkins. 2024. A survey of robotic language grounding: tradeoffs between symbols and embeddings. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 7999–8009.
[7] Jerome T Connor, R Douglas Martin, and Les E Atlas. 1994. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks* 5, 2 (1994), 240–254.
[8] Gabriel Del Castillo, Grace Clark, Zhao Han, and Tom Williams. 2023. Exploring the Naturalness of Cognitive Status-Informed Referring Form Selection Models. In *Proceedings of the 16th International Natural Language Generation Conference*. 269–278.
[9] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences* 3, 11 (1999), 419–429.
[10] Susan Goldin-Meadow. 2005. *Hearing gesture: How our hands help us think*. Harvard University Press.
[11] Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics* 3 (1975), 43–58.
[12] Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* (1993), 274–307.
[13] Zhao Han, Daniel Hammer, Kevin Spevak, Mark Higger, Aaron Fanganello, Neil Thomas Dantam, and Tom Williams. 2025. Givenness Hierarchy Theoretic Sequencing of Robot Task Instructions. *Frontiers in Robotics and AI* 12 (2025), 1640535.
[14] Zhao Han, Polina Rygina, and Thomas Williams. 2022. Evaluating Referring Form Selection Models in Partially-Known Environments. In *Proceedings of the 15th Int'l Conf. on Natural Language Generation*. 1–14.
[15] Zhao Han and Tom Williams. 2023. Evaluating cognitive status-informed referring form selection for human-robot interactions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 45.
[16] Mark Higger, Polina Rygina, Logan Daigler, Lara Ferreira Bezerra, Zhao Han, and Tom Williams. 2023. Toward Open-World Human-Robot Interaction: What Types of Gestures Are Used in Task-Based Open-World Referential Communication?. In *The 27th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.
[17] Mark Higger and Tom Williams. 2024. GAIA: A Givenness Hierarchy Theoretic Model of Situated Referring Expression Generation. In *Proceedings of the annual meeting of the cognitive science society*.
[18] Salah Hihi and Yoshua Bengio. 1995. Hierarchical recurrent neural networks for long-term dependencies. *Advances in neural information processing systems* 8 (1995).
[19] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.
[20] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. 2022. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094* 2, 3 (2022), 6.
[21] Andrew Kehler. 2000. Cognitive status and form of reference in multimodal human-computer interaction. In *AAAI/IAAI*. 685–690.
[22] Adam Kendon. 1997. Gesture. *Annual review of anthropology* 26, 1 (1997), 109–128.
[23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[24] Sotaro Kita. 2003. *Pointing: Where language, culture, and cognition meet*. Psychology Press, Chapter Pointing: A Foundational Building Block of Human Communication, 1 – 9.
[25] Larry R Medsker, Lakhmi Jain, et al. 2001. Recurrent neural networks. *Design and applications* 5, 64-67 (2001), 2.
[26] Poulomi Pal, Grace Clark, and Tom Williams. 2021. Givenness Hierarchy Theoretic Referential Choice in Situated Contexts. In *CogSci*.
[27] Poulomi Pal, Lixiao Zhu, Andrea Golden-Lasher, Akshay Swaminathan, and Tom Williams. 2020. Givenness hierarchy theoretic cognitive status filtering. *arXiv preprint arXiv:2005.11267*.
[28] Raedy Ping and Susan Goldin-Meadow. 2010. Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science* 34, 4 (2010), 602–619.
[29] Stephanie Schreitter and Brigitte Krenn. 2016. The OFAI multi-modal task description corpus. In *Proceedings of the Tenth Int'l Conference on Language Resources and Evaluation (LREC'16)*. 1408–1414.
[30] John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.
[31] Kevin Spevak, Zhao Han, Tom Williams, and Neil T Dantam. 2022. Givenness Hierarchy Informed Optimal Document Planning for Situated Human-Robot Interaction. In *2022 IROS*.
[32] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems* 3, 1 (2020).
[33] Elizabeth Wakefield, Miriam A Novack, Eliza L Congdon, Steven Franconeri, and Susan Goldin-Meadow. 2018. Gesture helps learners learn, but not merely by guiding their visual attention. *Developmental science* 21, 6 (2018), e12664.
[34] Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. 2016. Situated Open World Reference Resolution for Human-Robot Dialogue. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.
[35] T Williams and M Scheutz. 2019. A givenness hierarchy theoretic approach. *The Oxford handbook of reference* (2019), 457.
[36] Tom Williams and Matthias Scheutz. 2019. Reference Resolution in Robotics: A Givenness Hierarchy Theoretic Approach. In *The Oxford Handbook of Reference*, Jeanette Gundel and Barbara Abbott (Eds.). Oxford University Press.
[37] Deirdre Wilson and Dan Sperber. 2002. Relevance theory. In *Handbook of pragmatics*, L. Horn and G. Ward (Eds.).
[38] Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems* 31 (2018).

Received 2025-09-30; accepted 2025-12-01