Givenness Hierarchy Informed Optimal Document Planning for Situated Human-Robot Interaction

Kevin Spevak*, Zhao Han*, Tom Williams, and Neil T. Dantam

***Abstract*— Robots that use natural language in collaborative tasks must refer to objects in their environment. Recent work has shown the utility of the linguistic theory of the *Givenness Hierarchy* (GH) in generating appropriate referring forms. But before *referring expression generation*, collaborative robots must determine the content and structure of a sequence of utterances, a task known as *document planning* in the natural language generation community. This problem presents additional challenges for robots in situated contexts, where described objects change both physically and in the minds of their interlocutors. In this work, we consider how robots can "think ahead" about the objects they must refer to and how to refer to them, sequencing object references to form a coherent, easy to follow chain. Specifically, we leverage GH to enable robots to plan their utterances in a way that keeps objects at a high *cognitive status*, which enables use of concise, anaphoric referring forms. We encode these linguistic insights as a *mixed integer program* within a planning context, formulating constraints to concisely and efficiently capture GH-theoretic cognitive properties. We demonstrate that this GH-informed planner generates sequences of utterances with high inter-sentential coherence, which we argue should enable substantially more efficient and natural human-robot dialogue.**

I. Introduction

Robots in domains ranging from collaborative manufacturing to intelligent tutoring will need to use sequences of utterances to teach or otherwise provide information to human interlocutors. In collaborative manufacturing, for example, a robot may need to instruct a worker as to how to perform a complex task over several steps. In these types of domains, there is often substantial flexibility in the set of instructions that the robot can convey, and the order in which instructions are given. In a manufacturing task, for example, a robot may need to describe a set of multi-step procedures that can be executed in any order; the robot could in this case begin by describing the first step of each procedure, or could describe each subtask as a whole before moving on to the next subtask.

In the natural language generation (NLG) community, this task of determining the overall content and structure of generated language is referred to as *document planning*. As the name suggests, however, most previous approaches to document planning are designed for non-situated, purely textual domains. Situated dialogues, in contrast, require speakers to take into account their interlocutors' cognitive context. Specifically, speakers must generate utterances in a way that leverages the *cognitive status* of information for their listeners—e.g., referring to an object as "it" rather than "the N" if their interlocutor is already focused on the object. Speakers need to take into account changes in this cognitive context on the basis of their own utterances, or on the basis of interjections, questions, and non-verbal behaviors by their interlocutor. And, as we argue in this paper, they thus need to be able to perform document planning in a way that is sensitive to these dynamics, generating sequences of references that allow for meaningful inter-sentential coherence and the use of a wide variety of referring forms, to facilitate compact, easily followable utterance sequences.

To facilitate these capabilities, robots must thus maintain not only mental representations of the objects comprising their shared collaborative context, but must moreover maintain information about the cognitive status of those objects, for use in referring form selection and document planning. Approaches toward this goal can be grounded in conceptions of cognitive status from the linguistic literature, such as the Givenness Hierarchy (GH) [1]. According to this theory, when humans use anaphora ("it," "that", etc.) rather than only definite descriptions (e.g., "the blue box"), they signal subconscious assumptions about the cognitive status held by their target referent, either in the mind of their interlocutor or in the (metaphorical) mind of the conversation. This theory has been highly successful in the linguistics literature, validated across many disparate natural languages [2].

GH-based approaches have also seen significant success in the robotics community. Williams et al. [3] demonstrated how this approach can be used to facilitate the understanding of a wide variety of referring forms (see also the chapter on this topic in the Oxford Handbook of Reference [4]); and more recently, Pal et al. [5], [6] demonstrated how cognitive status can be estimated across the course of a conversation and used to guide the selection of referring forms. We argue that this theory could also be leveraged to enable advances in other stages of a robot's language generation process. Specifically, we argue that this theory could enable effective new robot document planning capabilities that are fundamentally well-suited for situated human-robot interaction (HRI) domains, thus enabling robots to generate sequences of utterances that are sensitive to the cognitive status dynamics of their interlocutors.

We present an NLG approach that incorporates the GH-theoretic cognitive status of a robot's human interlocutor to generate optimal document plans. We show that cognitive status can be encoded as constraints in *mixed integer programming* (MIP), which we integrate into a MIP formulation

*The first two authors contributed equally.

All authors are with the Department of Computer Science, Colorado School of Mines, 1500 Illinois St., Golden, CO 80401, USA. {kspevak,zhaohan,twilliams,ndantam}@mines.edu

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

of classical planning. We also present an example objective function that rewards high-cognitive-status referents. This approach enables generation of document plans with high inter-sentential coherence and facilitates effective use of anaphora over definite descriptions (e.g., "it" over "the N"). The presented system offers a proof-of-concept for the use of cognitive status as state variables in planning and optimization approaches for NLG.

We evaluate our work on a manipulation task in a human-robot collaborative tasking scenario, and show how sequences of instructions are generated by our approach in that domain. We first encode this scenario in the Planning Domain Definition Language (PDDL) [7] with actions representing possible instructional utterances. We then generate document plans for the robot's instructions to the human using both a classical planner and our GH-aware planner, showing that our formulation does enable the use of more concise referring forms.

We argue that such referring form usage should improve inter-sentential coherence in a complete NLG pipeline, and ultimately improve efficiency and usability in HRI applications. In particular, the highest cognitive statuses allow for the use of pronouns over definite noun phrases. There are numerous benefits to using these forms [8], as they make dialogue more efficient (and thus less costly to listen to) [9], more predictable (and thus cognitively easier to follow and more humanlike) [10], and conforming to Gricean conversational maxims of cooperative speech [11]. Additionally, the impact of these effects are magnified in situated contexts, in which the use of these forms facilitates—and is facilitated by—the use of deictic gesture to more effectively pick out objects at (and based on) varying distances [12]. We hypothesize that leveraging these benefits will improve task performance and user satisfaction, similar to the advantages gained through use of shorter object descriptions (Wallbridge et al. [13]).

Another potential benefit of our approach is in reducing the cognitive workload (cf. [14]) of participants in human-robot collaborative tasks. It is well established that human performance is degraded when cognitive workload is too high [15]. Human-like, context-dependent referring forms have been shown to reduce workload [16], and high-cognitive-status referents are conducive to the use of such forms. Furthermore, high working memory load slows spoken-word recognition time [17]. This strain on language processing could be ameliorated by the shorter instructions generated by our method.

Reduced cognitive workload, improved human performance on tasks, and increased user satisfaction are highly desirable traits for HRI applications. While this paper focuses on the technical approach for GH-aware document planning, our work thus suggests key HRI hypotheses to evaluate through interaction studies in future work.

II. Related Work

Our approach is fundamentally grounded in the linguistic theory of the Givenness Hierarchy (GH) [1]. The GH is comprised of a hierarchically nested set of six cognitive statuses: $\{$in focus $\subseteq$ activated $\subseteq$ familiar $\subseteq$ uniquely identifiable $\subseteq$ referential $\subseteq$ type identifiable$\}$. Each status is associated with one or more referring forms. For example, an object that is *in focus* can be referred to with the pronoun "it". Furthermore, a speaker that uses the word "it" implicitly signals a belief that the object is *in focus* in the mind of the listener. The GH theory posits that these six cognitive statuses are universal to human discourse, and has been validated across multiple languages [2].

There is a growing body of work in using computational models of the GH for NLG applications. Pal et al. [5] showed that the GH can be used successfully for referring form selection, building on the work of other researchers [18], [19], [3], [4] who have implemented portions of the GH (although not attempting to directly model cognitive status) for the purposes of reference resolution.

In this work we focus on using GH for a key NLG task not considered in prior work. Modular NLG pipelines [20] typically include modules for sentence planning (deciding how to communicate a sentiment), referring expression generation (selecting properties to use to refer to referents), and linguistic realization (ensuring grammatical correctness [21]). Above all these components sits the module of greatest interest to us, the document planner, which decides on an overarching sequence of sentiments to communicate a larger point or achieve a larger goal [22].

In this work, we seek to enable more effective document planning by leveraging cognitive status estimation. Our key insight is that document planning approaches that fail to account for cognitive status may exhibit decreased inter-sentential coherence. For instance, these approaches may introduce more referents than is strictly needed (or repeatedly re-introduce referents that are no longer activated), requiring full definite descriptions rather than shorter anaphoric phrases. In contrast, an approach that aims to use and continue referring to task-relevant entities that are already in focus or activated would lead to greater inter-sentential coherence, shorter and easier-to-follow dialogues, and perhaps even fundamentally simpler plans overall.

Towards this end, we propose to leverage the GH in document planning. Our modeling of the GH is similar to Pal et al.'s [5] Finite State Machine model, which they show to accurately predict cognitive status of referents in a preexisting monologue. We reproduce this model within planner state so that it can inform content determination and text structuring. We use these insights to guide the performance of a constraint-based task planning algorithm [23], [24], to plan actions that not only achieve desired goals, but also keep objects at higher tiers of cognitive status when possible. Constraint-based approaches for classical planning have typically used formulations based on Boolean satisfiability (SAT) [25], [26], [27]. We apply a formulation based on constrained optimization, specifically, mixed integer programming (MIP), to optimize based on cognitive status. The relationship between SAT and integer programs is long-established [28], and integer programming has been used in efficient planners [29]. Additionally, MIP's capabilities in real-valued optimization

makes our approach extensible to more complex models such as Pal et al.'s [5] probabilistic *Cognitive Status Filter*.

Existing planning languages impact the expressivity and capability of planning approaches [30] and do not easily represent the structure of GH-theoretic cognitive status. Instead, we incorporate cognitive status directly into the MIP formulation. As we will show, our MIP formulation succeeds in our goal of generating document plans with fewer and higher-cognitive-status referents. Finally, formulating document planning as MIP enables use of highly engineered solution procedures [31], [32], [33].

III. Problem Formulation

We propose a novel approach for situated document planning that optimizes GH-theoretic cognitive status using mixed integer programming.

We model the cognitive status for each object in the scene according to the GH coding protocol [34]. Our approach is designed to maximize use of linguistic forms associated with higher cognitive statuses—i.e., closer to *in focus* in the hierarchy.

A. Background on Constraint-based Planning

From a technical perspective, our approach is grounded in constraint-based planning techniques. Planning is an established technique for NLG. We briefly review the constraint-based planning formulation and refer the reader to [21], [24], [35] for more background.

Constraint-based planning: Classical approaches for constraint-based planning encode a scenario as a logical formula, then use a constraint solver—typically, a SAT solver [25], [26]—to find a satisfying variable assignment for that formula, corresponding to the plan [35, p69]. Variables in the formula represent the state and the action to take for a fixed number of steps. The formula itself describes valid plans, and we describe such a formula in equations (1)–(10). The planner progressively increases step count until the formula is satisfiable. The true action variables in the satisfying assignment encode the action to take for each step of the plan.

Mixed Integer Programs and Planning: Mixed integer programs (MIP) generalize SAT to include both real number variables and an objective function to optimize. We formulate situated document planning as MIP to enable optimal reference selection. While MIP is at least as hard as SAT [28], MIP formulations critically let us leverage highly engineered solution techniques [31], [32], [33].

We apply and extend the following formulation of planning as MIP.

Definition 1 (MIP-based planning): A planning instance is the tuple (P, A, O, I, G), where:

- P is the set of grounded predicates
- A is the set of grounded actions
- O is the set of objects
- $I \subseteq P$ is the set of true predicates in the initial state
- $G \subseteq P$ is the set of true predicates in the goal state

The MIP formula contains the following Boolean variables:

$$p_t, \quad \forall p \in P,\ t \in [0, T]$$
$$a_t, \quad \forall a \in A,\ t \in [1, T]$$

where p_t is true iff predicate p is true at time step t and a_t is true iff action a is taken at time step t. Then, the MIP contains the conjunction of the following constraints:

$$p_0 = 1, \quad \forall p \in I \tag{1}$$
$$p_0 = 0, \quad \forall p \notin I \tag{2}$$
$$p_T = 1, \quad \forall p \in G \tag{3}$$
$$a_t \leq p_{t-1}, \quad \forall a \in A, p \in \mathrm{pre}^+(a), t \in [1, T] \tag{4}$$
$$a_t \leq 1 - p_{t-1}, \quad \forall a \in A, p \in \mathrm{pre}^-(a), t \in [1, T] \tag{5}$$
$$a_t \leq p_t, \quad \forall a \in A, p \in \mathrm{eff}^+(a), t \in [1, T] \tag{6}$$
$$a_t \leq 1 - p_t, \quad \forall a \in A, p \in \mathrm{eff}^-(a), t \in [1, T] \tag{7}$$
$$\sum_{a \in A} a_t \leq 1, \quad \forall t \in [1, T] \tag{8}$$
$$p_t \leq p_{t-1} + \sum_{a \in \mathrm{add}(p)} a_t, \quad \forall p \in P, t \in [1, T] \tag{9}$$
$$p_t \geq p_{t-1} - \sum_{a \in \mathrm{del}(p)} a_t, \quad \forall p \in P, t \in [1, T] \tag{10}$$

where $\mathrm{pre}^+(a)$ and $\mathrm{pre}^-(a)$ are the sets of positive and negative preconditions of a, $\mathrm{eff}^+(a)$ and $\mathrm{eff}^-(a)$ are the sets of positive and negative effects of a, and $\mathrm{add}(p)$ and $\mathrm{del}(p)$ are the sets of actions with p as a positive/negative effect. Conditions (1) and (2) encode the initial state, (3) encodes the goal condition, (4) through (7) enforce consistency of action preconditions and effects, (8) encodes operator exclusion, and (9) and (10) are the frame axioms. Lastly, we define the objective function,

$$\min_{a_t} \sum_{a \in A, t \in [1,T]} a_t \ ,$$

which ensures we find the shortest valid plan.

B. Situated Document Planning

We consider a situated document planning problem where the robot must describe an embodied task to a human. Specifically, we generate the sequence of utterances describing the task; surface realization to generate the linear text is beyond the scope of this work, but established techniques [21] are applicable. In the planning problem, predicates describe the state of the situated world, and objects correspond to the interactable physical objects in the interlocutor's situated context. The actions represent utterance skeletons that instruct the interlocutor to manipulate physical objects. Action parameters correspond to physical objects. For example, the plan step (pick-up block-A) would correspond to an instruction to pick up the item designated as block-A.

While not performed in this work, these forms could easily be translated into the types of utterance representations typically used in cognitive robotic architectures like DIARC [36]. Specifically, each plan step, when translated into a predicate p (e.g. pick-up(block-A)) can be assumed to be

TABLE I: Classical Approach to Situated Document Planning

Instruction	Example	Planning	Example
physical object	block A	object	block-A
utterance skeleton	"Pick up [ref]."	action	(pick-up ?o)
utterance	"Pick up that block"	grounded action	(pick-up block-A)

TABLE II: Cognitive Status Criteria

Cognitive Status	Condition
In Focus (I)	Topic of previous utterance
Activated (A)	Referenced in previous two utterances
Familiar (F)	Referenced in any previous utterance
Uniquely identifiable (U)	Always

part of an utterance of form $Inst(r, h, p)$, i.e. an Instruction from the speaker (robot r) to the hearer h instructing them to perform action p. However, this could be formulated in other ways, based on pragmatic inference, to achieve various communicative goals such as politeness [37].

In Section III-C, we extend the state space to include the cognitive status of action parameters (i.e. referents). The initial state of the planning problem corresponds to no instructions having been given, and the goal state is that each step in the manipulation task has been described. Table I summarizes this formulation.

C. MIP Formulation of Cognitive Status

We present a formulation of MIP variables and constraints to track the GH-theoretic cognitive status of each object in the scene at each planning time step. In particular, we capture the structure of GH-theoretic cognitive status to enable efficient reasoning. Classical planning languages and techniques often pose challenges for representing structure [30]. Instead, we incorporate GH-theoretic structure directly into the MIP.

Our formulation uses a simplified version of the criteria specified in the GH coding protocol [34], including only what is applicable in the instruction-giving task described in Section III-B. The exact criteria we use are shown in Table II. Note that "topic" in the above table encompasses the linguistically distinct terms *subject*, *syntactic topic*, and *syntactic focus*. We leave the specification of an utterance's *topic* to the planning domain (see Section IV-A) so the planning formulation is independent of these distinctions. Additionally, we assume that every object is at least *uniquely identifiable*, which forms the implicit default if none of the above criteria are met. This assumption holds true in the situated context of our sample problem: a workspace with a collection of distinct objects. The presented formulation can easily be extended to include the omitted statuses for applications where they are applicable.

The key challenge lies in grounding actions that affect cognitive status, i.e., converting from a first order to a Boolean (or integer) representation. The size of the grounded representation significantly impacts running time [24], [25], [27]. A naïve SAT-based encoding of cognitive status results in a large number of constraints in the grounded representation. The number of grounded actions is exponential in action arity: a single action requires $T|O|^n$ grounded variables where T is the number of time steps, O is the set of objects, and n is the action's arity. Typical planning problems have low-arity actions that only affect a small number of objects. However, GH-theoretic cognitive status is different: the cognitive status of an object may change at each time step even if it is not referred to at that time step. For example, if object A is *in focus*, then an utterance referencing only object B is used, the cognitive status of A changes to *activated*. Thus, accounting for cognitive status requires each grounded action to update a number of state variables proportional to the number of objects, resulting in an additional $\mathcal{O}\left(T|O|^{|O|}\right)$ constraints per action.

We address the challenge of concisely modeling cognitive status updates with a formulation of MIP constraints that capture GH-theoretic structure. Our formulation introduces only a $\mathcal{O}\left(T|O|\right)$ new constraints, instead of the exponential number of constraints required for a naïve encoding. We extend the formulation in Section III-A with the following Boolean variables,

$$I_{o,t},\ A_{o,t},\ F_{o,t}, \quad \forall o \in O,\ t \in [0, T]\ ,$$

where $I_{o,t}$ is true iff object o is *in focus* at time step t, $A_{o,t}$ is true iff object o is *activated* at time step t, and $F_{o,t}$ is true iff object o is *familiar* at time step t. Then, we introduce the following constraints:

$$I_{o,t} = \sum_{\text{topic}(o)} a_t, \quad \forall o \in O,\ t \in [1, T] \quad (11)$$

$$A_{o,t} \leq \sum_{\text{ref}(o)} a_t + a_{t-1}, \quad \forall o \in O,\ t \in [1, T] \quad (12)$$

$$A_{o,t} \geq \frac{1}{2} \sum_{\text{ref}(o)} a_t + a_{t-1}, \quad \forall o \in O,\ [1, T] \quad (13)$$

$$F_{o,t} \leq F_{o,t-1} + \sum_{\text{ref}(o)} a_t, \quad \forall o \in O,\ t \in [1, T] \quad (14)$$

$$F_{o,t} \geq \frac{1}{2} \left(F_{o,t-1} + \sum_{\text{ref}(o)} a_t \right), \quad \forall o \in O,\ t \in [1, T] \quad (15)$$

$$I_{o,0}, A_{o,0}, F_{o,0} = 0, \quad \forall o \in O \quad (16)$$

where $\text{topic}(o)$ is the set of actions with o as their *topic* and $\text{ref}(o)$ is the set of actions that reference o. Constraints 11 through 15 encode the cognitive status criteria, and 16 encodes the initial condition. Finally, we replace the objective function with:

$$\min_{a_t} \sum_{t \in [1,T]} \sum_{o \in O} \sum_{a \in \text{ref}(o)} a_t(8 - 4F_{o,t} - 2A_{o,t} - I_{o,t}) \quad (17)$$

which encodes a cost for each object reference depending on its cognitive status at the time of the reference. The costs are shown in table Table III. As this work is meant to serve as a proof-of-concept, we chose the simple rule that a reference at any cognitive status is equivalent in cost to two references at the next higher status. We leave the development of a more informed objective function, which could be informed by linguistic literature on, e.g., cost-of-comprehension or other psycholinguistically relevant criteria [8], to future work.

TABLE III: Cognitive status costs

Status	*In Focus (I)*	*Activated (A)*	*Familiar (F)*	*Uniquely Identifiable (U)*
Cost	1	2	4	8

IV. Evaluation

We evaluate our approach on a sample problem to identify the impact of GH-theoretic cognitive status. We apply both a classical formulation and our GH-theoretic optimizing approach to construct MIPs, which we solve using the Gurobi Optimizer [32]. We compare the resulting plans, which show that GH-theoretic optimization does produce plans that use higher cognitive status referents.

A. Sample Problem

We develop the sample problem *gadgets* (see Figure 1), which involves generating instructions to assemble gadgets from parts. This problem is similar to the tower construction task used in Robotics [38], and has potential utility in key HRI domains like collaborative manufacturing. However, this gadgets problem is not meant to directly represent a realistic scenario, but rather an example to compare the plans generated by a classical planner and our optimizing planner. In the *gadgets* domain, there are parts, tools, and boxes. Items in boxes must be taken out before they can be used. There are three types of tools: screwdrivers, wrenches, and grippers. Certain parts can be attached to another part using a screwdriver or using a wrench, and parts can be wired using a gripping tool.

B. Results

For the *gadgets* problem given in Figure 1, the classical encoding generated the plan in Table IV, and our GH-aware encoding generated the plan in Table V.

The GH-aware plan makes several changes that result in an increased use of high-cognitive-status referents, as summarized below:

- *Object reuse:* The classical plan uses three separate tools, while the optimized plan uses only the multi-tool. This avoids the low-cognitive-status references required in switching tools.
- *Planning ahead for object reuse:* In order to access the multi-tool, the new planner must first take it out of the box. This shows how the GH-aware planning approach goes beyond greedily reusing objects.
- *Accepting longer plans:* The new planner gives a plan that is longer than the classical plan, due to the step of taking the multi-tool out. This demonstrates that our system is able to handle a trade-off between brevity and complexity. Note that the GH-aware plans are not always longer, and that this trade-off can be tuned with the objective function.
- *Separating sub-tasks:* The classical plan switches back and forth between working on the motor gadget and the chip gadget. The new planner completes work on one before starting the other to keep objects at maximum cognitive status.

Although this work focuses on the document planning level, Table V middle shows possible surface realizations for each step of the GH-aware plan from Table V left, using referring forms informed by cognitive status as the most prominent feature [6]. We include this natural language example to facilitate comparison to the classical plan given in Table IV and highlight the summarized changes.

Figure 2 shows the total references used in each plan by cognitive status. The total costs of the classical and optimized plans according to our objective function are 100 and 90 respectively.

We compare running times of our proof of concept implementation for the classical and GH-aware encodings in Table VI. The GH-aware encoding does increase running time over the classical encoding. There are a number of potential causes for the difference, including the quadratic objective function of the GH-aware encoding and the existence of many of plans that satisfy the goal conditions. The GH-aware encoding must find not only a shortest, satisficing plan but one that optimizes the objective function (17). In our sample problem, the shortest, satisficing plan takes 7 steps while the optimal plan takes 8 steps. There are several potential refinements that would improve running time, including object typing and parallel actions [24]. Further heuristics to prune satisficing but nonoptimal solutions may also yield improvements.

V. Discussion and Future work

The plans generated by our GH-informed planner differ substantially in structure from those generated with a classical approach. These differences result in more references to high-cognitive-status objects while still accomplishing the same communicative goal. Since cognitive status is closely tied to referring form choice [6], plans generated with our approach will result in shorter language after text realization. The language may sound more natural as well, since humans tend to avoid forming complex referring expressions when possible [39]. We believe our results demonstrate the utility of the modeling cognitive status at the document planning level. There has been recent interest in computational models of cognitive status and GH-based referring form selection has been demonstrated with good results [6]. However, previous work has focused on tasks that fall under text realization. Our results are a proof of concept of how document planning can benefit from these works as well.

For this work, we considered only cognitive status; but how speakers choose to refer to objects depends on a myriad of other situational values that could also be modeled. Referring form selection, for example, also depends on physical distance and distractors [6]. These values could be derived from planner state and combined with cognitive status to create a more sophisticated objective function for the planner.

The use of the multi-tool in the sample problem suggests another potential line of investigation. The planner uses the tool so consistently that it remains *activated* from the first time it is referenced to the end of the plan. In such a situation, a human instructor would likely stop referring to it altogether,

```
(define (domain gadgets)
  (:predicates (screwdriver ?o) (wrench ?o) (part ?o) (box ?o)
               (gripper ?o) (out ?o) (in ?o ?b) (attached ?o1 ?o2)
               (wired ?o) (screwable ?o) (boltable ?o))
  (:action take-out
     :parameters (?topic ?b)
     :precondition (and (in ?topic ?b) (box ?b))
     :effect (and (out ?topic) (not (in ?topic ?b))))
  (:action screw-in
     :parameters (?topic ?p ?s)
     :precondition (and (out ?topic) (out ?p) (out ?s) (part ?topic)
                        (part ?p) (screwdriver ?s) (screwable ?topic))
     :effect (and (attached ?topic ?p)))
  (:action bolt-in
     :parameters (?topic ?p ?w)
     :precondition (and (out ?topic) (out ?p) (out ?w) (part ?topic)
                        (part ?p) (wrench ?w) (boltable ?topic))
     :effect (and (attached ?topic ?p)))
  (:action wire
     :parameters (?topic ?g)
     :precondition (and (out ?topic) (out ?g)
                        (part ?topic) (gripper ?g))
     :effect (wired ?topic)))
```

```
(define (problem assemble)
  (:domain gadgets)
  (:objects toolbox partbox multitool allen
            phillips motor axle gear board
            chip led pliers)
  (:init (box toolbox) (box partbox) (part motor)
         (part axle) (part gear) (part board)
         (part chip) (part led) (gripper pliers)
         (screwdriver phillips) (wrench allen)
         (gripper multitool) (wrench multitool)
         (screwdriver multitool) (screwable axle)
         (screwable chip) (boltable gear)
         (boltable led) (out allen) (out phillips)
         (out motor) (out axle) (out board)
         (out pliers) (out gear) (in chip partbox)
         (in led partbox) (in multitool toolbox))
  (:goal (and (attached gear axle)
              (attached axle motor)
              (attached chip board)
              (attached led board)
              (wired board))))
```

Fig. 1: Gadgets domain and problem definitions in PDDL

TABLE IV: Classical Encoding Plan

Planner Output	Possible Surface Realization	Cognitive Status Costs
`(take-out led partbox)`	*"Take the LED out of the box of parts"*	8 (U), 8 (U)
`(take-out chip partbox)`	*"Take the chip out of that"*	8 (U), 2 (A)
`(screw-in axle motor phillips)`	*"Screw the axle into the motor with the phillips screwdriver"*	8 (U), 8 (U), 8 (U)
`(bolt-in gear axle allen)`	*"Bolt the gear onto it with the allen wrench"*	8 (U), 1 (I), 8 (U)
`(screw-in chip board phillips)`	*"Screw that chip into the breadboard with that"*	4 (F), 8 (U), 2 (A)
`(wire board pliers)`	*"Wire that with the pliers"*	2 (A), 8 (U)
`(bolt-in led board allen)`	*"Bolt that LED onto it with that allen wrench"*	4 (F), 1 (I), 4 (F)

TABLE V: GH-aware Encoding Plan

Planner Output	Possible Surface Realization	Cognitive Status Costs
`(take-out multitool toolbox)`	*"Take the multi-tool out of the toolbox"*	8 (U), 8 (U)
`(screw-in axle motor multitool)`	*"Screw the axle into the motor with it"*	8 (U), 8 (U), 1 (I)
`(bolt-in gear axle multitool)`	*"Bolt the gear onto it with this tool"*	8 (U), 1 (I), 2 (A)
`(wire board multitool)`	*"Wire the breadboard with that"*	8 (U), 2 (A)
`(take-out chip partbox)`	*"Take the chip out of the box of parts"*	8 (U), 8 (U)
`(screw-in chip board multitool)`	*"Screw it into this board with that"*	1 (I), 2 (A), 2 (A)
`(take-out led partbox)`	*"Take the LED out of this box"*	8 (U), 2 (A)
`(bolt-in led board multitool)`	*"Bolt it onto this board with that"*	1 (I), 2 (A), 2 (A)

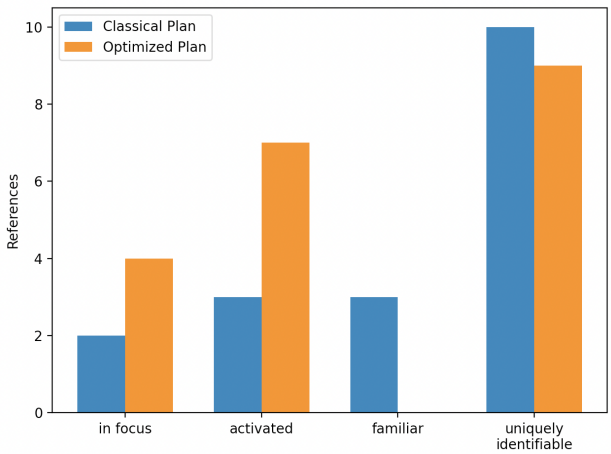

Fig. 2: References made by cognitive status

TABLE VI: Running Times

Classical encoding	GH-aware Encoding
0.59 seconds	14.12 seconds

assuming that the instruction follower will continue using the same tool if unspecified. This leads to the notion of leveraging referential assumptions at the document planning level.

While this work focuses on the technical approach and contribution, we plan to evaluate our work in a human-subjects study with both objective and subjective measures. Objective metrics include accuracy, whether a participant's resulting action reaches the desired goal state. Subjective measures, to be administered after each utterance, focus on the perceived understandability of the planner output; tentative metrics include comprehensiveness, simplicity, and mental workload (e.g., through NASA Task Load Index [40]). A baseline approach with definite descriptions will be compared with our approach where referring forms are used. We believe our approach will have higher scores in these metrics.

Once the planner output is evaluated, we plan to add gesturing capability and evaluate with a robot to understand how the perception of a robot is affected. A physically embodied robot presents further interesting challenges, such as how the physical distance between participants, the robot, and objects affect document planning.

VI. Conclusion

In this paper, we propose modeling GH-theoretic cognitive status within document planning for situated human-robot interaction. We present a proof of concept by encoding the GH coding criteria in a MIP-based planner and demonstrate the solution to a sample instruction-giving problem. Our MIP encoding captures the structure of GH-theoretic cognitive status to form a concise set of constraints. Our resulting plans indicate the utility of our approach and motivate further investigation in GH-aware, situated document planning.

Key areas for future research include human-subjects studies on the generated language, extending the GH-theoretic model, and experiments on physical robot platforms. This work could serve as a framework for human-robot collaboration systems that exhibit complex decision-making behavior that maximizes human understanding of natural language prompts and reduces human cognitive load.

References

[1] J. K. Gundel, N. Hedberg, and R. Zacharski, "Cognitive status and the form of referring expressions in discourse," *Language*, pp. 274–307, 1993.

[2] J. K. Gundel, M. Bassene, B. Gordon, L. Humnick, and A. Khalfaoui, "Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation," *Journal of Pragmatics*, vol. 42, no. 7, pp. 1770–1785, 2010.

[3] T. Williams, S. Acharya, S. Schreitter, and M. Scheutz, "Situated open world reference resolution for human-robot dialogue," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 311–318.

[4] T. Williams, "A givenness hierarchy theoretic approach," *The Oxford handbook of reference*, p. 457, 2019.

[5] P. Pal, L. Zhu, A. Golden-Lasher, A. Swaminathan, and T. Williams, "Givenness hierarchy theoretic cognitive status filtering," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2020.

[6] P. Pal, G. Clark, and T. Williams, "Givenness hierarchy theoretic referential choice in situated contexts," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, no. 43, 2021.

[7] D. McDermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins, *PDDL – the planning domain definition language*, AIPS-98 Planning Competition Committee, 1998.

[8] J. E. Arnold and S. A. Zerkle, "Why do people produce pronouns? pragmatic selection vs. rational models," *Language, Cognition and Neuroscience*, vol. 34, no. 9, pp. 1152–1175, 2019.

[9] H. Tily and S. Piantadosi, "Refer efficiently: Use less informative expressions for more predictable meanings," in *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*. Citeseer, 2009.

[10] E. Williams and J. Arnold, "Priming discourse structure guides pronoun comprehension," in *Poster, CUNY conference on human sentence processing, University of Colorado*, 2019.

[11] H. P. Grice, "Logic and conversation," in *Speech acts*. Brill, 1975, pp. 41–58.

[12] S. C. Levinson, "Deixis," in *The handbook of pragmatics*. Blackwell, 2004, pp. 97–121.

[13] C. D. Wallbridge, A. Smith, M. Giuliani, C. Melhuish, T. Belpaeme, and S. Lemaignan, "The effectiveness of dynamically processed incremental descriptions in human robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 1, pp. 1–24, 2021.

[14] Z. Han, A. Norton, E. McCann, L. Baraniecki, W. Ober, D. Shane, A. Skinner, and H. A. Yanco, "Investigation of multiple resource theory design principles on robot teleoperation and workload management," in *ICRA*.

[15] B. Xie and G. Salvendy, "Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments," *Work & stress*, vol. 14, no. 1, pp. 74–99, 2000.

[16] E. Campana, M. K. Tanenhaus, J. F. Allen, and R. Remington, "Natural discourse reference generation reduces cognitive load in spoken systems," *Natural Language Engineering*, vol. 17, no. 3, pp. 311–329, 2011.

[17] B. Hadar, J. E. Skrzypek, A. Wingfield, and B. M. Ben-David, "Working memory load affects processing time in spoken word recognition: Evidence from eye-movements," *Frontiers in neuroscience*, vol. 10, p. 221, 2016.

[18] A. Kehler, "Cognitive status and form of reference in multimodal human-computer interaction," in *AAAI/IAAI*, 2000, pp. 685–690.

[19] J. Y. Chai, P. Hong, and M. X. Zhou, "A probabilistic approach to reference resolution in multimodal user interfaces," in *Proceedings of the 9th international conference on Intelligent user interfaces*, 2004, pp. 70–77.

[20] E. Reiter, "Pipelines and size constraints," *Computational Linguistics*, vol. 26, no. 2, pp. 251–259, 2000.

[21] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[22] D. D. McDonald, "Issues in the choice of a source for natural language generation," *Computational Linguistics*, vol. 19, no. 1, pp. 191–197, 1993.

[23] H. A. Kautz and B. Selman, "Planning as satisfiability," in *ECAI*, vol. 92, 1992, pp. 359–363.

[24] J. Rintanen, "Engineering efficient planners with SAT," in *ECAI*, 2012, pp. 684–689.

[25] H. A. Kautz and B. Selman, "Blackbox: A new approach to the application of theorem proving to problem solving," in *AIPS98 Workshop on Planning as Combinatorial Search*, 1998, pp. 58–60.

[26] J. Rintanen, "Madagascar: Scalable planning with SAT," in *8th International Planning Competition (IPC-2014)*, 2014, pp. 66–70.

[27] N. T. Dantam, Z. K. Kingston, S. Chaudhuri, and L. E. Kavraki, "An incremental constraint-based framework for task and motion planning," *IJRR*, vol. 37, no. 10, pp. 1134–1151, 2018.

[28] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of computer computations*. Springer, 1972, pp. 85–103.

[29] M. Van Den Briel, T. Vossen, and S. Kambhampati, "Reviving integer programming approaches for ai planning: A branch-and-cut framework." in *ICAPS*, 2005, pp. 310–319.

[30] J. Rintanen, "Impact of modeling languages on the theory and practice in planning research," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[31] N. Bjørner, A.-D. Phan, and L. Fleckenstein, "νz-an optimizing SMT solver," in *TACAS*, vol. 15, 2015, pp. 194–199.

[32] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2022. [Online]. Available: http://www.gurobi.com

[33] L. Perron, "Operations research and constraint programming at Google," in *International Conference on Principles and Practice of Constraint Programming*. Springer, 2011, pp. 2–2.

[34] J. K. Gundel, N. Hedberg, R. Zacharski, A. Mulkern, T. Custis, B. Swierzbin, A. Khalfoui, L. Humnick, B. Gordon, M. Bassene *et al.*, "Coding protocol for statuses on the givenness hierarchy," *Unpublished manuscript (1993/2006). http://www. sfu. ca/hedberg/Coding_for_Cognitive_Status. pdf*, 2006.

[35] S. M. LaValle, *Planning algorithms*. Cambridge university press, 2006.

[36] M. Scheutz, T. Williams, E. Krause, B. Oosterveld, V. Sarathy, and T. Frasca, "An overview of the distributed integrated cognition affect and reflection DIARC architecture," in *Cog. Arch*, 2019.

[37] T. Williams, G. Briggs, B. Oosterveld, and M. Scheutz, "Going beyond command-based instructions: Extending robotic natural language interaction capabilities," in *Proc. AAAI*, 2015.

[38] M. F. Jung, D. DiFranzo, S. Shen, B. Stoll, H. Claure, and A. Lawrence, "Robot-assisted tower construction—a method to study the impact of a robot's allocation behavior on interpersonal dynamics and collaboration in groups," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 1, pp. 1–23, 2020.

[39] D. Byron, A. Koller, J. Oberlander, L. Stoia, and K. Striegnitz, "Generating instructions in virtual environments (give): A challenge and an evaluation testbed for nlg," 2007.

[40] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.