Frame Semantics for Human-Robot Interaction

Mitchell Abrams* MITCHELL.ABRAMS@TUFTS.EDU
Chris Bao* CHRIS.BAO@TUFTS.EDU
Matthias Scheutz MATTHIAS.SCHEUTZ@TUFTS.EDU
Department of Computer Science, Tufts University, Medford, MA 02155 USA

Abstract

We propose a novel approach to semantic frame parsing for human-robot interaction (HRI), grounded in the FrameNet framework and integrated into the DIARC cognitive architecture. Our system improves natural language understanding by supporting dynamic interpretation of frame-evoking elements and their associated roles (frame elements) in real time. To expand lexical coverage and improve robustness in open-ended dialogue, we incorporate a large language model (LLM) to suggest additional lexical units for frame evocation and to assist in frame element filling. This hybrid neuro-symbolic method improves upon existing robotic frame parsers, such as RoboFrameNet, by enabling broader generalization and reflects the FrameNet ontology. We demonstrate that our system facilitates downstream reasoning, planning, and reference resolution in situated interaction scenarios. Our architecture-agnostic approach offers a flexible, modular parsing approach designed for focused HRI domains where frame coverage can be curated and extended.

1. Introduction

Consider the following brief exchange between a human and a household robot:

> **Human:** "Can you grab the mug from the table and bring it to me?"
>
> **Robot:** "Okay."

A shallow semantic parser might extract simple relational facts:

```
action(grab)
object(mug)
location(table)
recipient(human)
```

Although correct at a surface level, this representation does not capture the event structure that supports planning and inference. Using FrameNet (Baker et al., 1998), the utterance evokes the `Bringing` frame (a conceptual structure), which captures the movement of a Theme under control of an Agent from a Source to a Goal. A FrameNet-style parse makes these roles explicit:

0. *These authors contributed equally to this work.

```
Frame: Bringing
  Agent: robot
  Theme: mug
  Source: table
  Goal: human
  Carrier: robot_hand
  Constant_location: in_hand
  {...}
```

This frame-based structure supports downstream reasoning (e.g., validating that the mug is graspable, or that a collision-free path exists). It also allows for generalization to similar requests (e.g., *“Bring me the book from the shelf”*)[1], planning of intermediate actions (e.g., navigation around obstacles), and more robust reference resolution in situated contexts. FrameNet-based knowledge thus provides a principled way to structure the semantics of everyday language for natural, goal-directed human-robot interaction.

To our knowledge, there are no FrameNet-based parsers tailored to dialogue and human–robot interaction that account for embodiment. RoboFrameNet (Thomas & Jenkins, 2012) is a verb-centric frame parser for robotics, but it fails to capture broader semantic domains that span both general and specific concepts. These themes are as varied as chance, perception, communication, transaction, time, space, motion, life, social context, emotion, and cognition (Baker et al., 1998).

The few FrameNet parsers that exist are trained on text and apply to domains that do not handle situational dialogue or HRI settings (Kalyanpur et al., 2020; Swayamdipta et al., 2017; Das et al., 2014). FrameNet is a lexicographic resource built from manually annotated example sentences drawn largely from the British National Corpus (BNC) with additional full-text documents (Fillmore et al., 2002; Ruppenhofer et al., 2010).

In practice, most FrameNet annotations originate in the BNC, a balanced but predominantly written corpus (Davies, 2004). As a result, FrameNet-style parsers are typically trained on exemplar and full-text annotations from FN 1.3 or FN 1.5/1.7 (e.g., SEMAFOR, Open-SESAME), reflecting written prose rather than task-oriented dialogue with imperatives and questions (Das et al., 2010; Kalyanpur et al., 2020).

FrameNet parsers are not built for HRI-style situated commands where roles and entities may be implicit in a scene, and thus require augmentations to infer such information. (As an example, consider the implicit role of a speaker or listener in an instruction.) The required adaptation is not only to the language itself but also to the *situational* context that comes with embodiment—roles, constraints, and referents that are not explicit in text. FrameNet parsing for HRI therefore needs extra-linguistic input (perception, world state, norms) and tight coupling to a cognitive architecture to support goal-oriented, agentic behavior. We address this by building FrameNet-based semantic representations from both linguistic evidence and factual/perceptual input available to the agent, yielding grounded structures that can complement a shallow parser that supplies dialogue-act intent.

[1] When the verb lexicalizes *transfer of possession* (e.g., “hand/give X to Y”), we instead use the `Giving` frame, with *Recipient* as a core role. In this example, the phrasing “bring . . . to me” naturally instantiates `Bringing` with *Goal* = human.

We present a FrameNet-based parsing approach tailored to human–robot interaction and integrated into the DIARC architecture (Scheutz et al., 2018). Our contributions are: (1) *situated* FrameNet parsing for HRI that produces FrameNet-style structures (core and peripheral roles, frame relations) and grounds them to percepts and a knowledge base in real time; (2) a *hybrid neuro-symbolic* filling mechanism in which an LLM expands lexical units for frame evocation and performs targeted frame-element filling when linguistic evidence is sparse while preserving transparent symbolic representations; and (3) *tight architecture integration* that links frame interpretations to goal/plan generation in DIARC with feasibility checks from planning, with demonstrations in supermarket and kitchen scenarios showing improved planning and reference resolution over verb-centric and shallow semantic baselines. Taken together, these results indicate that a situated, LLM-assisted FrameNet yields semantic representations that are both interpretable and effective for embodied, goal-directed HRI.

2. Background

Ontologies and lexical-semantic resources provide complementary structures for language understanding in HRI. Type hierarchies (e.g., object and place taxonomies) capture *what* entities are, but they say little about the *events* those entities participate in or the roles they play. Frame semantics, operationalized by FrameNet (Baker et al., 1998), fills this gap by modeling conceptual situations (*frames*) together with their participants (*frame elements*, or FEs for short) and the words that evoke them (*lexical units*, or LUs). This structure captures higher-level situational information that can be leveraged for richer understanding and improved downstream reasoning for robots.

A *frame* names a schematic event, relation, or entity type (e.g., `Apply_heat`, `Placing`, or `Revenge`). Each frame specifies both *core* FEs that are essential to defining the themes or agents in an event and *non-core* or *peripheral* FEs that augment the event without redefining it. For example, the `Apply_heat` frame includes the Cook, Food, Heating_instrument, and Container core FEs and the Time, Manner, Place, Duration, and Purpose peripheral FEs.

LUs (verbs, nouns, adjectives, and some multiword expressions) *evoke* frames; `Apply_heat` is evoked by the verbs `fry.v`, `bake.v`, `boil.v`, and `broil.v`, among others. FrameNet also defines *frame–frame relations* that organize frames into a conceptual network, including *Inheritance* (is-a), *Subframe*, *Using*, *Perspective_on*, *Precedes*, and causative/inchoative links. Annotations pair LUs with sentence spans and transcribe, for each realized FE, its (a) name/role, (b) grammatical function (e.g., `External_Argument`, `Object`), and (c) phrase type (e.g., `NP`, `PP`); practically, this yields FE triplets per example. Although many evokers are verbs, FrameNet explicitly covers nominal and adjectival LUs (e.g., `retaliation.n` in `Revenge`, `asleep.a` in `Sleep`). The English resource has on the order of a thousand frames and has motivated parallel FrameNet efforts in other languages.

VerbNet (Kipper, 2005), another semantic ontology, groups verbs into classes with shared thematic roles and subcategorization patterns; it is *verb-class centric* and links predicate syntax to coarse-grained roles and selectional restrictions. *Abstract Meaning Representation (AMR)* (Banarescu et al., 2013) encodes sentence meaning as an acyclic directed graph whose predicates are typically PropBank framesets (Palmer et al., 2005) (e.g., `give-01`) with role labels like `ARG0`–

ARG5. AMR is thus *predicate-centric* and anchored in PropBank rather than in a frame inventory of event types. While both VerbNet and AMR are valuable, they lack FrameNet's explicit inventory of named frames with core/non-core FEs and frame–frame relations. For embodied HRI, FrameNet's situational roles (AGENT, THEME, SOURCE, GOAL, PATH, etc.) align more naturally with grounding, reference resolution, and planning constraints.

Building on these resources, Dialogue-AMR (Bonial et al., 2020) extends AMR with speaker intent (*illocutionary force*), tense, and aspect to better support human-robot dialogue. Like our work, Dialogue-AMR adapts an existing semantic framework to HRI. The key difference is that Dialogue-AMR introduces new annotation layers, while our approach focuses on automatically generating FrameNet parses that can complement simpler intent-based parsing (e.g., commands, questions, statements). Our goal is to make FrameNet practical for situated, embodied use by embedding it into a cognitive architecture.

RoboFrameNet (Thomas & Jenkins, 2012) likewise targets robotics applications, but its verb-centric semantics do not capture higher-level situational frames and events in the way FrameNet does. Thus, it is not directly equivalent to FrameNet as a general lexical ontology for HRI. A comparison between RoboFrameNet and our approach is detailed in section 5.

FrameNet parsing itself is challenging. It involves *target identification* (determining what words evoke a frame), *frame identification* (assigning the frame), and *argument labeling* (marking spans for core and non-core roles). Existing state-of-the-art parsers include SEMAFOR (Das et al., 2010), Open-SESAME (Swayamdipta et al., 2017), and the generative and multi-task parsers introduced by Kalyanpur et al. (2020). However, these approaches are not designed for embodied, multimodal contexts or HRI; they operate on text and cannot leverage perceptual grounding or contextual information that is crucial in HRI.

In general, semantic parsing refers to the task of mapping natural language utterances to structured meaning representations ranging from logical forms and robot actions to frame-semantic roles and arguments. For robotic applications, many recent systems leverage end-to-end neural semantic parsers, often powered by large language models (LLMs) to map natural language directly to structured outputs such as robot actions or logical forms. These approaches typically bypass intermediate symbolic structures, instead relying on sequence-to-sequence models to perform direct interpretation. While this strategy enables strong generalization and robustness to linguistic variation, it often sacrifices transparency and interpretability. Notably, such systems rarely produce structured intermediate representations (such as semantic frames or compositional meaning structures) that can be inspected and manipulated for reasoning and downstream modules.

SayCan (Ahn et al., 2022), for example, combines an LLM with affordance-based planning to interpret high-level instructions via interpretable affordance graphs, offering a transparent interface for grounding linguistic input in physical capabilities. While such methods move toward integrating symbolic knowledge in planning, the language understanding component remains opaque, lacking explicit symbolic decomposition.

Recent work has also investigated the use of LLMs beyond parsing, particularly for high-level planning and reasoning in interactive settings. For instance, TidyBot (Wu et al., 2023) uses an LLM to infer object categories, user preferences, and personalized clean-up strategies in home environments. While this approach showcases the capacity of LLMs for flexible, context-sensitive reasoning in HRI, it still lacks the structured semantic representations—such as frames or logical forms.

In contrast, a broad class of robot language understanding systems has traditionally emphasized symbolic semantic parsing as a foundation for mapping language into executable actions or structured logical forms. For example, Tellex et al. (2011) stress the importance of interpretable intermediate representations for grounding natural language in robotic settings. Their model, Generalized Grounding Graphs (G3), uses the linguistic structure of commands to dynamically build probabilistic graphical models that link language constituents to groundings in the robot's environment—enabling compositional and transparent interpretation. Several frameworks leverage Combinatory Categorial Grammar (CCG) or logic-based formalisms to support this mapping: Artzi & Zettlemoyer (2013) use weak supervision to train CCG-based parsers that map instructions to grounded action sequences, while Matuszek et al. (2013) employ a probabilistic CCG model to translate route descriptions into a LISP-style Robot Control Language (RCL) capable of expressing procedural structures like loops and conditionals.

Together, these approaches demonstrate the value of transparent, structured semantic representations in enabling robust, interpretable, and grounded language understanding for robotic agents. Our work draws on this tradition while extending it with FrameNet-based situational semantics.

While FrameNet offers a rich, manually curated lexicon grounded in frame semantics, its utility as a general-purpose knowledge base for intelligent agents has been challenged. As discussed by McShane et al. (2024), attempts to leverage FrameNet for learning and interpretation in language-endowed intelligent agents (LEIAs) revealed some challenges: inconsistent frame granularity, metaphorical usages, ambiguous role labels. Additionally, FrameNet was not originally built for agents. These limitations suggest that FrameNet is most effective when used selectively: targeting frames that align well with an agent's internal representations and domain-specific goals. We acknowledge this constraint. Rather than treating FrameNet as a universal semantic solution, we use it as a structured representation for context, event structure, and role expectations for human–robot interaction tasks in constrained domains. When paired with perceptual and situational knowledge, even a small subset of frames (central to HRI) can support transparent situated interpretation and reasoning. This hybrid approach supports the claim in McShane et al. (2024) that semantic resources are most useful when tightly integrated into an intelligent system that understands how to use the frame information. In our system, frame knowledge is explicitly embedded within a broader cognitive architecture to support real-time HRI.

By coupling the FrameNet lexical database (to trigger frames from lexical units) with large language models (to assist with argument identification and labeling), we enable situated interpretation for HRI. Perceptual components and a knowledge base further contribute extra-linguistic information (e.g., settings, roles, and object properties) to frame interpretation. In the next section, we describe this architecture in detail.

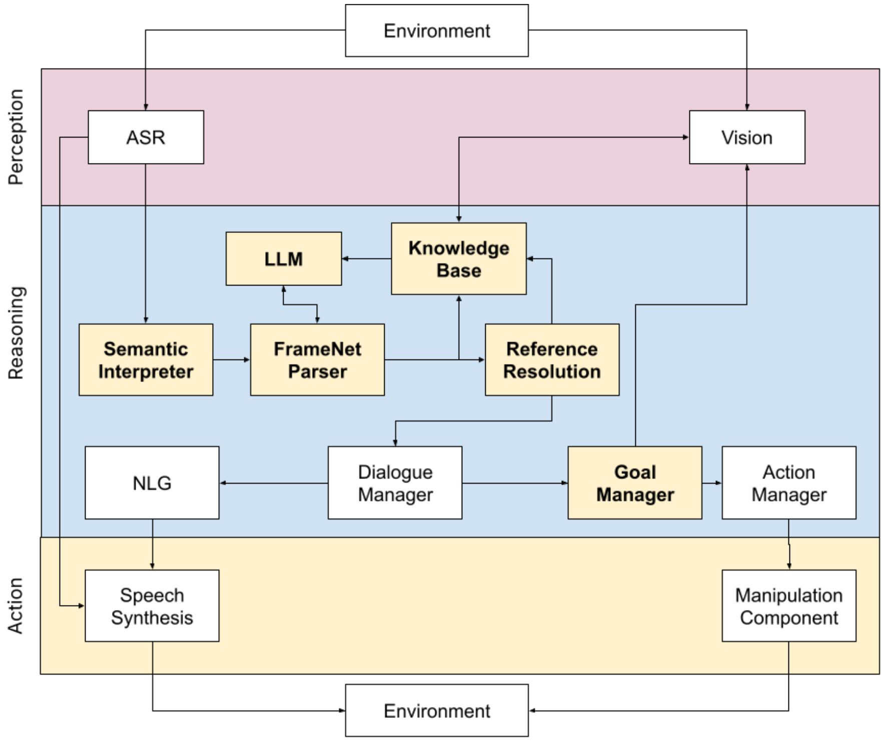

Figure 1: Cognitive architecture integrating FrameNet parsing, first-order logic representation, and an LLM component within the DIARC architecture. Utterances are incrementally interpreted into both FrameNet-style frame structures and logical semantic forms. The FrameNet Parser works in tandem with a semantic interpreter and LLM component, updating the knowledge base with frame-level knowledge including settings, roles, and contextual information. Reference resolution, goal management, and action execution are grounded in both perceptual inputs (from vision and speech) and internal beliefs, enabling situated understanding and behavior.

3. System Overview

3.1 Cognitive Architecture

Figure 1 shows our cognitive architecture, which consists of various interdependent components that support *perception*, *reasoning* (including language understanding), and *action*. We will focus on the language and reasoning components of this architecture, as they are core to the parsing approach.

Our architecture integrates a FrameNet-based parser (acting as our FrameNet component) alongside a first-order logic semantic interpreter and a large language model (LLM) component. This system parses utterances into dual semantic structures: a flat logical form (suitable for action execution and reasoning) and a structured FrameNet frame representation (useful for role assignment, context modeling, and conceptual inference). The FrameNet parser receives initial semantic input from the linguistic interpreter and collaborates with the LLM component to fill in missing frame elements when lexical gaps or underspecification occur. These frame-based representations are then stored in the knowledge base, where they can support downstream reference resolution, dialogue management, and goal inference.

In our system, FrameNet parsing operates in parallel with a CCG-style semantic interpreter that produces logical forms annotated with speaker intent (e.g., `INSTRUCT`, `QUESTION`, `STATEMENT`). This dual representation captures both the propositional content of an utterance and its illocutionary force—information crucial for dialogue management in instructional human–robot interaction. While FrameNet excels at modeling situational structure and participant roles, it does not explicitly encode communicative intent or dialogue function. By integrating both representations, our architecture enables robots to distinguish not just what action or event is described, but also whether the speaker is issuing a command, asking a question, or making an observation. This design parallels extensions like Dialogue-AMR (Bonial et al., 2020), which augment semantic representations with dialogue act information. In our implementation, FrameNet complements rather than replaces intent-based parsing.

The distinguishing feature of this architecture is its capacity to support FrameNet parsing in an embodied, situated context. Because the system is embedded in a broader cognitive architecture with access to perceptual components (e.g., vision and speech), frame role labeling can draw on real-world referents, visual search results, and context-specific goals. The knowledge base further enriches interpretation by contributing stored facts, beliefs, and persistent object knowledge. This situated integration enables FrameNet-style parsing to support grounded interpretation in structured human–robot interactions.

4. FrameNet Parsing Methodology

Our methodology addresses the three subtasks of FrameNet parsing: *target identification*, *frame identification*, and *argument labeling*. We describe how each step is adapted to the needs of situated human–robot interaction.

4.1 Target Identification

The system receives an utterance as input, such as: *"Hand it to me."* This utterance is passed to the FrameNet component, which performs target identification—the process of detecting lexical units (LUs) that potentially evoke frames. We implement a deterministic dictionary-based lookup, mapping observed words to their associated frames in FrameNet.

This step presented a unique challenge in HRI. While FrameNet's lexicon contains many LUs for frames (e.g., the `Containers` frame includes LUs such as `amphora.n`, `bag.n`, `barrel.n`, `basin.n`), these do not always align with the verbs and nouns commonly used in instructional or command-oriented dialogue. For instance, many everyday directives in HRI (*"hand"*, *"pass"*, *"give"*) evoke frames relevant to transfer or exchange but are not always robustly covered by the default lexical units. In other cases, evolution in modern language or gaps in the corpus may lead to some frames not being triggered. Addressing this mismatch required both extending FrameNet's LU coverage and introducing additional filtering downstream.

4.2 Frame Identification

The candidate frames generated during target identification provide the initial hypotheses. At this stage, the large language model (LLM) component (GPT-4) is invoked to refine the set. The LLM is prompted with the candidate frames, the original utterance, and symbolic world knowledge from the knowledge base (KB). Using this combined input, the LLM selects only those frames that are contextually appropriate.

For example, in *"hand it to me"*, both `Giving` and `Placing` may be triggered, but the KB (which encodes that the speaker is the intended recipient) and the linguistic context favor `Giving`. This step therefore operationalizes frame identification as a contextual disambiguation task, leveraging both linguistic and situational evidence.

4.3 Argument Labeling

Once the relevant frames have been identified, their frame elements (roles) must be instantiated. Here, we again rely on the LLM, which fills roles based on (a) the original utterance and (b) facts from the KB. The KB may contain dialogue history, perceptual beliefs, or contextual information such as speaker identity, listener identity, or current task goals.

Argument filling often requires incorporating implicit situational roles. For instance, in *"Hand it to me"*, the `Giving` frame requires an AGENT, THEME, and RECIPIENT. While only *"me"* (AGENT) and *"it"* (THEME) are explicit, the KB supplies the RECIPIENT (the robot, as addressee). This results in a more complete parse than what surface text alone would provide.

The final parse is a *composite frame*, which may include multiple relevant frames: a frame describing the essential "verb" of the situation (e.g., `Giving`) together with more specific ones that augment particular roles at play (e.g., `Container`). This multi-frame representation is useful for HRI, where utterances often evoke overlapping event structures. Frames are serialized into Prolog facts for symbolic reasoning (planning, action selection) and are also stored in the KB with unique frame IDs, allowing them to persist across dialogue turns and be updated during interaction.

4.4 Representation Format

Frame information is stored in structured JSON objects, enabling runtime access and modification of LUs, core roles, peripheral roles, and frame–frame relations. A shortened example representation is shown below:

```
{
  "frame": "Grasp",
  "lexical_units": ["grab.v", "grasp.v", "seize.v", "snatch.v"],
  "core_frame_elements": {
    "Agent": "The entity performing the grasping",
    "Item": "The object being grasped"
  },
  "peripheral_frame_elements": {
    "Instrument": "Tool used by the Agent to grasp",
    "Manner": "How the grasping is performed",
    "Place": "Location of the grasping event"
  },
  "frame_relations": {
    "inherits_from": ["Manipulation"],
    "inherited_by": ["Arrest"]
  }
}
```

This structured representation allows the parser to operate both symbolically (via KB integration) and neurally (via LLM-based filling), ensuring that frame-based semantics are transparent, interpretable, and suitable for downstream robotic reasoning.

5. Comparison to Existing Semantic Frame Representation

While both our system and RoboFrameNet (Thomas & Jenkins, 2012) adopt a FrameNet-inspired approach to semantic parsing, these architectures and underlying assumptions diverge substantially. RoboFrameNet focuses on *verb-centric* representations, where each action verb (e.g., *turn*, *grab*, *go*) is mapped to a predefined semantic frame with core roles extracted from syntactic dependency parsing. Despite some theoretical and implementation details that differ from the FrameNet representation, they still consider the main concepts of a semantic frame: including *semantic frames*, *lexical units*, *core* and *non-core roles*, and *frame elements*. This approach supports rapid semantic parsing in robot middleware, but is limited to surface-level verb frames and rigid mappings, neglecting peripheral roles, frame inheritance information, or contextual grounding.

The semantic frame representation in RoboFrameNet resourcefully leverages existing techniques in natural language processing (dependency parsing) to determine grammatical relations from natural language input. However, a key insight of the FrameNet representation is that it captures frame elements that may diverge from grammatical (syntactic) roles. Consider this example:

These two examples illustrate key limitations of dependency-based parsing in capturing event semantics. In Table 1, FrameNet distinguishes intentional roles like AVENGER, OFFENDER, and INJURY, even when they are realized syntactically as prepositional phrases modifying the same verb. Dependency parsing, by contrast, offers no semantic differentiation; both *"with you"* and *"for this"* are treated as generic modifiers of *"get."*

Table 2 presents a more grounded, instruction that is more fitting of the robotics domain. The phrase *"from the red toolbox"* is ambiguous from a syntactic standpoint; it could modify either *"hand"* or *"screwdriver."* Dependency parses cannot resolve this ambiguity or determine the intended semantic relation. FrameNet, however, interprets it within the `Giving` frame, labeling *"me"* as RECIPIENT, *"screwdriver"* as THEME, and *"red toolbox"* as SOURCE. These are conceptual roles critical for robotic planning and object grounding.

These examples highlight how FrameNet semantic roles more closely align with the reasoning needs of embodied agents than traditional syntactic roles. Our framework builds on this insight by incorporating large language models (LLMs) as semantic fillers for frame elements. Because

Dependency Parse	**FrameNet Semantic Parse**
Sentence: *I'll get even with you for this.*	**Frame:** `Revenge`
Verb: get → root	**Lexical Unit:** `get_even.v`
Subject: I → `nsubj(get)`	**Avenger (Core FE):** I
Prepositional Phrase: with you → `prep(get)`	**Offender (Core FE):** with you
Prepositional Phrase: for this → `prep(get)`	**Injury (Core FE):** for this
Interpretation: The two PPs "with you" and "for this" are structurally indistinct — both are modifiers of the verb "get." Dependency structure provides no semantic differentiation.	**Interpretation:** FrameNet reveals event-level roles: "with you" names the Offender; "for this" identifies the Injury. These roles are crucial to understanding the intent behind the action.

Table 1: Comparison of dependency parsing and FrameNet semantic role labeling for the sentence *"I'll get even with you for this."* FrameNet enables conceptual distinctions between participants (Avenger, Offender, Injury) that are indistinguishable syntactically.

Dependency Parse	**FrameNet Semantic Parse**
Sentence: *Hand me the screwdriver from the red toolbox.*	**Frame:** `Giving`
Verb: hand → root	**Lexical Unit:** `hand.v`
Indirect Object: me → `iobj(hand)`	**Recipient (Core FE):** me
Direct Object: screwdriver → `dobj(hand)`	**Theme (Core FE):** screwdriver
Prepositional Modifier: from the red toolbox → `prep(hand)` or `prep(screwdriver)`	**Source (Core FE):** red toolbox
Interpretation: Structural roles (subject, objects, modifiers), but attachment is ambiguous: Is "from the red toolbox" describing where to fetch or which screwdriver?	**Interpretation:** Semantically explicit roles: Agent gives Theme to Recipient, with Source specified. Attachment ambiguity resolved by frame structure.

Table 2: Comparison between dependency parsing and FrameNet semantic role labeling for a robot-directed instruction. FrameNet assigns conceptually meaningful roles (Theme, Source, Recipient) that may not align directly with syntactic roles or resolve from grammar alone.

LLMs are not bound to grammatical dependencies, they can infer frame structures from broader context, including perceptual and situational cues—enabling robust interpretation and disambiguation in situated environments. This flexibility allows us to recover richly structured FrameNet-style representations in real time, suitable for grounded HRI and symbolic reasoning.

We offer an overall comparison of RobotFrameNet semantic frame and our semantic frame representation to highlight the richness of our representation with more frames and roles. On the right of Table 3 is a fuller representation for an HRI setting where there are multiple frames (we refer to this as a *composite frame*) triggered from the utterance *"Dempster, bring the mug to Evan's office tomorrow."*

This comparison underscores the strength of our approach. Unlike RoboFrameNet's verb-centric parsing, which maps a single lexical item to a single frame, our method constructs *com-*

RoboFrameNet Representation	**FrameNet Representation**
Frame: Bringing **Lexical Unit:** bring.v **Core FEs:** Agent (dempster), Theme (mug), Goal (Evan's office) **No peripheral roles modeled** **No additional frames triggered** **No frame relations or inheritance modeled** **Context:** Unspecified — no link to world model or perceptual systems. **Use:** Maps directly to a robot action like `bring(Agent, Theme, Goal)`.	**Frames Triggered:** **Bringing** (via *bring.v*) *Agent*: dempster *Theme*: the mug *Goal*: Evan's office *Carrier*: null (unspecified) *Time (peripheral)*: tomorrow **Containers** (inferred from noun *mug*) *Container*: mug *Contents*: null **Locative_relation** (supporting spatial reasoning) *Figure*: mug *Ground*: Evan's office **Calendric_unit** (via *tomorrow*) *Relative_time*: tomorrow **Frame Relations:** *Bringing* uses *Cause_motion* and *Motion* *Containers* inherits from *Object_properties* **Contextual Integration:** Roles linked to referents in the knowledge base and visual scene Frame-to-Plan grounding supports constraint validation (e.g., is "mug" graspable?)

Table 3: Comparison between RoboFrameNet's verb-centric frame parsing and our multi-frame situated FrameNet interpretation for the utterance *"Dempster, bring the mug to Evan's office tomorrow."*

posite frames by triggering multiple frames from distinct lexical units in the utterance. This enables the system to capture both the primary action (`Bringing`) and additional situationally relevant structures such as containers, spatial relations, and temporal modifiers.

Importantly, the overlapping lexical spans across frames do not introduce redundancy, but rather highlight different conceptual facets of the same expression. For example, *"mug"* simultaneously evokes the role of THEME in the `Bringing` frame and the role of CONTAINER in the `Containers` frame. Similarly, *"tomorrow"* activates both a peripheral TIME element and the `Calendric_unit` frame. Although these overlaps share surface spans, the semantic roles they fill are frame-specific, providing non-interchangeable information that is crucial for downstream reasoning.

By allowing multiple frames and frame elements to co-exist, our representation produces a fuller interpretation of utterances from multiple perspectives. This directly supports situational awareness for embodied agents: the robot is not only informed that an action of `Bringing` must occur, but also that the object to be manipulated is a container, that the goal location is grounded in a spatial relation, and that the event is temporally constrained. Such richly structured semantics support tighter coupling between natural language understanding, perception, and planning, and can improve task execution in structured HRI environments.

6. Demonstrations

This work is novel, and our implementation of the pipeline is in progress. We have seen success using the LLM to extend the dictionary of frame-evoking units, as well as identifying and filtering frames from given utterances. The following demonstrations, specifically tables 5 and 7, include some of the preliminary results of our pipeline. The model is able to accurately identify key context-defining frames and integrate knowledge base data to attempt frame filling. In addition, we evaluate potential downstream applications of frame knowledge on tasks such as norm-based reasoning and planning (so far unimplemented).

6.1 Supermarket example: cashier

The retail environment, specifically a supermarket, is commonly referenced in HRI. In our example, the robot takes the role of the cashier, taking the human customer through the checkout process. The robot follows a predefined script of actions, including retrieving items from the conveyor belt, cart, or basket, scanning each product, and bagging them. This scenario presents a rich and detailed environment that contains a diverse array of objects and social norms to track.

As a representative case, we focus on one particular moment in the execution of this script: returning the checked-out items to the customer after they have been paid for. Such a command might read, *"Return the items to the customer."* At this instant, the robot is aware of several objects in the environment, summarized in Table 4. In addition to the intended object of the script — the bag containing the purchased groceries — there are also three distractor objects: an empty grocery bag, a cash register, and the money inside the register. The task of the robot is to determine which object to return and, subsequently, which concrete action to take.

Object	Description	Properties
`Person0`	The human customer the robot interacts with	`object, person, customer`
`Bag0`	An empty grocery bag	`object, bag, grasp`
`Bag1`	A full grocery bag	`object, bag, grasp, contains`
`Groceries0`	Merchandise in the bag	`object, grasp, contained`
`Register0`	The cash register	`object, register, grasp, contains`
`Money0`	The cash inside the register	`object, money, grasp, contained`

Table 4: Supermarket environment objects and their properties.

In the semantic frame parsing workflow we propose, the robot would have access to a number of sources of information. In the scene leading up to this particular moment, utterances and actions, in addition to *a priori* knowledge, will have built up a knowledge base about the objects (Belief, tabularly represented by Table 4) and about the context itself (FrameDatabase). Table 5 contains a snapshot of the frames currently active and their relations.

Frames triggered
Commerce_goods-transfer
Buyer: `Person0`
Seller: `self`
Goods: `Groceries0`
Money: `Money0`
Placing
Agent: `self`
Goal: `Person0`
Theme: `Groceries0`
Containers
Container: `Bag1`
Contents: `Groceries0`

Table 5: Supermarket environment frame relations (FrameNet frames).

From the semantic and frame information, reference resolution can then solve for *“the items"* which are to be returned: As `self` is the seller, the goods — `Groceries0` — must be the desired object. However, the groceries are contained inside `Bag1`. Therefore, the planner reasons that the bag must be placed near the consumer. Indeed, the bag is graspable, so the robot moves the bag, ending the transaction.

6.2 Kitchen example: assisting in food preparation

The household kitchen is another common domain in HRI, involving many manipulable objects, spatial relations, and practical constraints about cleanliness and task ordering. In our example, the robot assists a human in preparing a meal. The robot follows a collaborative plan that includes fetching utensils and ingredients and performing light preparation tasks.

As a representative case, the human is cooking soup and requests, *“Pass me the spoon.”* At this instant, the robot’s perceptual system detects multiple candidate spoons along with other nearby items, summarized in Table 6. In addition to the intended clean spoon for stirring, there is a dirty spoon on the counter and other distractors.

In our semantic frame parsing workflow, the robot’s knowledge base includes both object properties and currently active frames relevant to the situation. Table 7 shows a snapshot of these frames. We rely on FrameNet frames (e.g., `Giving`, `Containers`, `Apply_heat`), while practical cleanliness preferences are modeled as rules in the KB.

Object	Description	Properties
`Person0`	The human cook the robot is assisting	`object, person, cook`
`Spoon0`	A clean metal spoon on the counter	`object, utensil, spoon, grasp, clean`
`Spoon1`	A dirty spoon with sauce residue	`object, utensil, spoon, grasp, dirty`
`Pot0`	A pot containing soup on the stove	`object, container, pot, contains`
`Soup0`	Soup inside the pot	`object, food, contained`
`Stove0`	A stovetop that can apply heat	`object, appliance, stove, heat_source`

Table 6: Kitchen environment objects and their properties.

Frames triggered

Giving

Agent: `self`
Recipient: `Person0`
Theme: `Spoon?` (to be resolved)

Containers

Container: `Pot0`
Contents: `Soup0`

Apply_heat

Container: `Pot0`
Cook: `Person0`
Food: `Soup0`
Heating_instrument: `Stove0`

Table 7: Kitchen environment frame relations (FrameNet frames). Practical cleanliness preferences are modeled as KB rules, not as frames.

Contextual KB rules

```
prefer_clean_utensil(U) :- utensil(U), clean(U).
avoid_dirty_utensil(U) :- utensil(U), dirty(U).
```

From the frame information and KB rules, reference resolution selects `Spoon0` as the intended referent for *"the spoon"*. Although both `Spoon0` and `Spoon1` satisfy the lexical category `spoon`, the KB rule prefers clean utensils for immediate food preparation, ruling out `Spoon1`. The planner then executes: grasp `Spoon0` and hand it to `Person0`, completing the `Giving` frame.

Beyond reference resolution, active frames provide additional inferences about the environment. For example, knowing that `Soup0` is the Contents of `Pot0` within a `Containers` frame, and that `Pot0` participates in `Apply_heat`, the robot can anticipate ongoing cooking actions and constraints. For example, it may reason that removing the soup from the heat implies moving the pot that contains it, and not grasping the soup itself; or it may reason that the stove's heat may be adjusted if the recipe calls for it. These facts are provided by the `Containers` and `Apply_heat` frames, respectively. This overlapping frame knowledge supports more robust interpretation, allowing the robot to act meaningfully even when perceptual information is partial or underspecified.

7. Discussion & Conclusion

Our hybrid parsing approach demonstrates both promise and limitations for situated human–robot interaction. On the one hand, it enables transparent and verifiable semantic representations that can be directly inspected before being committed to the robot's knowledge base. The transparency of using the existing FrameNet lexical database allows our parse to be faithful to the FrameNet ontology, even though we heavily integrate an LLM component for lexical unit expansion, frame pruning, and role labelling and filling. In this way, our approach is more transparent than using an LLM end-to-end in creating a final composite frame. This transparency is particularly valuable in HRI, where safety and trust depend on interpretable reasoning.

At the same time, the system does not yet capture the full richness of dialogue phenomena (e.g., disfluencies, ellipses, corrective language), and the LLM component can potentially misclassify frames. Moreover, frame persistence remains an open challenge (robustly linking frames referenced across different turns in dialogue), such that a frame evoked early can be consistently re-identified when it is referenced later. Further, it may prove difficult to detect when particular frames should be dropped from consideration after being evoked. Future work should address temporal frame tracking to ensure continuity in frame-based dialogue understanding.

Beyond its immediate implementation, the approach is designed for focused HRI domains where frame coverage can be curated and extended, rather than for unrestricted open-domain language understanding. We argue that the compromise between rule- and LLM-based pipelines is particularly appropriate in these domains: rules provide speed, transparency, and symbolic grounding, while LLMs supply flexibility and coverage in open-ended dialogue.

In addition to the parser itself, the choice of FrameNet as a representational backbone has broader benefits for HRI. Frame-based representations provide a structured semantic layer that supports conceptual inference and symbolic reasoning. While our system is not designed for open-domain language understanding, it enables generalization across related interactions within focused domains by leveraging shared or related frame structures. This capability is made possible and understandable by FrameNet's hierarchical frame structure. As a simple example, a robot that has knowledge of the `Commerce_goods-transfer` frame (with roles such as BUYER, SELLER, and GOODS) may encounter a novel scenario that does not have these precise components but which overlaps with this frame in its broad structure. By mapping the observed interaction onto the more general `Transfer_scenario` or `Exchange` frames, the robot can still reason about participant roles and actions, leaving room to adapt or specialize once a more specific frame is introduced or learned. This ability to scaffold understanding with existing frames makes FrameNet a valuable resource for enabling flexible and adaptive interpretation in dynamic environments. Furthermore, our framework is flexible enough to support *learning*: frames can be extended or refined for new domains either manually or through natural language interaction, enabling incremental adaptation over time.

In contrast, there are certainly arguments for end-to-end foundation model approaches to goal-oriented robotics, and we do not aim to reject them out of hand. However, we argue it is important to acknowledge the benefits that a hybrid neuro-symbolic approach can provide in terms of flexibility and robustness. Real-world expectations of robots, especially in team-oriented scenarios, hinge on the reliability and transparency of reasoning agents (Hancock et al., 2011; Ososky et al., 2013). The addition of deterministic frame evocation and the restriction of output to semantic frames decrease variance in output and allow for increased understandability of the frame parser as a whole. Other considerations include the prevalent issue of hallucination in foundation models (Rawte et al., 2023), as well as the maintainability of a contextual knowledge base that adapts to a changing environment. We believe that requiring a frame-parsing component to act within symbolic constraints will allow for more predictable results and a more extensible scaffold to build upon.

Beyond the immediate implementation details, our framework also offers three broader advantages. First, by supporting *composite frames*, it goes beyond single-frame mappings: multiple frames may be triggered by overlapping lexical spans, and these overlaps enrich rather than duplicate information, leading to greater situational awareness for the agent. Second, frame semantics naturally connects to normative and contextual reasoning. Unlike representations anchored in syntax or dependency parsing—or even other semantic approaches such as Dialogue-AMR—our frame-based approach highlights roles and expectations that map directly onto social norms, task constraints, and pragmatic reasoning. These representations are not mutually exclusive: they can be employed in parallel with other semantic formalisms (e.g., a lambda-calculus or CCG-style semantic parse for direct compositional semantics), yielding complementary layers of meaning. Third, while our approach does not aim for unbounded scalability, frames generalize across related events within curated domains, supporting flexible adaptation to complex tasks, multimodal integration, and transfer across structured interaction settings. Together, these properties make frame-based parsing a promising foundation for situated, interpretable language understanding in HRI.

In conclusion, we advocate for semantic representations that go beyond surface-level propositions, incorporating frame structures, conceptual roles, and frame-to-frame relationships that enable deeper situational understanding in HRI. While the presented approach is an initial step, it illustrates how combining symbolic transparency with neural flexibility can yield interpretable yet robust language understanding for embodied agents. Crucially, it is the surrounding cognitive architecture that enables this style of frame parsing to function effectively: by situating the parser within a broader reasoning and planning framework, the agent can ground frames in perception, update them dynamically, and use them to guide action. Future extensions will involve updating frame information through dialogues, expanding coverage of dialogue phenomena, and integrating online learning mechanisms. Taken together, these directions highlight a broader research agenda: building robots that understand not only words, but the frames of meaning that structure human communication.

References

Ahn, M., et al. (2022). Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Artzi, Y., & Zettlemoyer, L. (2013). Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the association for computational linguistics*, *1*, 49–62.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Banarescu, L., et al. (2013). Abstract meaning representation for sembanking. *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse* (pp. 178–186).

Bonial, C., Donatelli, L., Abrams, M., Lukin, S. M., Tratz, S., Marge, M., Artstein, R., Traum, D., & Voss, C. (2020). Dialogue-AMR: Abstract Meaning Representation for dialogue. *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 684–695). Marseille, France: European Language Resources Association. From `https://aclanthology.org/2020.lrec-1.86/`.

Das, D., Chen, D., Martins, A. F., Schneider, N., & Smith, N. A. (2014). Frame-semantic parsing. *Computational linguistics*, *40*, 9–56.

Das, D., Schneider, N., Chen, D., & Smith, N. A. (2010). *Semafor 1.0: A probabilistic frame-semantic parser*. Technical report, Technical Report CMU-LTI-10-001, Carnegie Mellon University.

Davies, M. (2004). British national corpus (from oxford university press). `https://www.english-corpora.org/bnc/`. Online interface.

Fillmore, C. J., Baker, C. F., & Sato, H. (2002). The framenet database and software tools. *LREC*.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, *53*, 517–527.

Kalyanpur, A., Biran, O., Breloff, T., Chu-Carroll, J., Diertani, A., Rambow, O., & Sammons, M. (2020). Open-domain frame semantic parsing using transformers. *arXiv preprint arXiv:2010.10998*.

Kipper, K. (2005). *A broad-coverage, comprehensive verb lexicon*. Doctoral dissertation, Doctoral dissertation at University of Pennsylvania.

Matuszek, C., Herbst, E., Zettlemoyer, L., & Fox, D. (2013). Learning to parse natural language commands to a robot control system. *Experimental robotics: the 13th international symposium on experimental robotics* (pp. 403–415). Springer.

McShane, M., Nirenburg, S., & English, J. (2024). *Agents in the long game of ai: Computational cognitive modeling for trustworthy, hybrid ai*. MIT Press.

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013). Building appropriate trust in human-robot teams. *AAAI spring symposium: trust and autonomous systems* (pp. 60–65).

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, *31*, 71–106.

Rawte, V., Sheth, A., & Das, A. (2023). A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., & Scheffczyk, J. (2010). Framenet ii: Extended theory and practice.

Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., & Frasca, T. (2018). An overview of the distributed integrated cognition affect and reflection diarc architecture. *Cognitive architectures*, (pp. 165–193).

Swayamdipta, S., Thomson, S., Dyer, C., & Smith, N. A. (2017). Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*.

Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S., & Roy, N. (2011). Understanding natural language commands for robotic navigation and mobile manipulation. *Proceedings of the AAAI conference on artificial intelligence* (pp. 1507–1514).

Thomas, B. J., & Jenkins, O. C. (2012). Roboframenet: Verb-centric semantics for actions in robot middleware. *2012 IEEE International Conference on Robotics and Automation* (pp. 4750–4755). IEEE.

Wu, J., Antonova, R., Kan, A., Lepert, M., Zeng, A., Song, S., Bohg, J., Rusinkiewicz, S., & Funkhouser, T. (2023). Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, *47*, 1087–1102.