Exploiting Deep Semantics and Compositionality of Natural Language for Human-Robot-Interaction

Manfred Eppe[1], Sean Trott[1], Jerome Feldman[1]

Abstract— **We are developing a natural language interface for human robot interaction that implements reasoning about deep semantics in natural language. To realize the required deep analysis, we employ methods from cognitive linguistics, namely the modular and compositional framework of Embodied Construction Grammar (ECG) [18]. Using ECG, robots are able to solve fine-grained reference resolution problems and other issues related to deep semantics and compositionality of natural language. This also includes verbal interaction with humans to clarify commands and queries that are too ambiguous to be executed safely. We implement our NLU framework as a ROS package and present proof-of-concept scenarios with different robots, as well as a survey on the state of the art in knowledge-based language HRI.**

I. INTRODUCTION

Robots are becoming more and more our companions and co-workers. Hence, a lot of effort has been made to develop methodologies to make the interaction between humans and robots seamless, intuitive and uncomplicated. Surprisingly, only relatively shallow work has been conducted with one of the most natural of all interaction methods with robots, namely Natural Language.

Of course, there exist several robots and other virtual agents that are equipped with some kind of natural language interface. Examples of current commercial products are Apple's Siri, Microsoft's Corona, Google Now, and the social multimedia robot *Jibo* [9]. A problem with these products is that they rely mostly on learned or predefined keyword-based input templates that trigger certain actions. This limits the subset of language that these products can interpret.

Natural Language Processing (NLP) is the is broad general term covering all the goals and approaches to computer systems that treat language. There are three main approaches that can be called: fixed-template, statistical, and theory/knowledge based. Many of the best known systems (Siri, etc.) use fixed templates to deal with specific language and often then default to a search engine. A wide range of shallow statistical, machine learning (ML) techniques are being employed to provide analyses and other information to human users. When an NLP system is designed to take meaningful action without human intervention, we say that Natural Language Understanding (NLU) is required. We suggest that a robust NLU system requires a strong basis in linguistic and conceptual theory and knowledge, although templates and ML methods can play a role. For robots, automatic cars, and other autonomous physical systems, shallow template-based and statistical approaches usually not sufficiently reliable, and their traditional safety layers can not cover all possible consequences of misunderstanding.

Robotics offers many properties that are a good fit for knowledge-based NLU systems, mainly because robots usually work in a closed domain, such as assisted living, household, or disaster response. This can be exploited by using reliable knowledge based approaches. In this work, we present such a grammar based approach for NLU in HRI, and formulate our straight-forward hypothesis as follows:

Incorporating cognitive theories and linguistic expert knowledge leads to significant improvements in language understanding for Human-Robot-Interaction.

Surprisingly, this simple hypothesis has not yet been deeply addressed in robotics (see Sections II-C and IV-B on related work and evaluation). We believe that the lack of mature computational systems for deep NLU is a major reason for this shortcoming. In our approach, we use an *Embodied Construction Grammar* (ECG) [18] analyzer to capture deep semantics in natural language by pairing syntax with semantics and universal embodied concepts. An advantage of ECG vs. other cognitive linguistic frameworks is its modularity. There is a universal core set of constructions and conceptual schemas, developed by experts from cognitive linguistics. These capture modalities, spatial relations, temporal expressions, actions, etc. The core set is common across domain-specific grammars, while the domain specific grammars themselves are usually relatively small, requiring only little knowledge about linguistics. The conceptual schemas in this core set are neurologically and cognitively motivated, and therefore largely language independent [18], [28].

In this paper, we describe how we connect the ECG language analyzer to the Robot Operating System (ROS) [21] in an effort to make it accessible to a wide range of robotic platforms. We also review the core grammar set to fine-tune it for robotic applications. We do not claim to solve all linguistic problems that exist, but we do claim that our modular language understanding framework does better than the state of the art for an important subset of linguistics problems, which we illustrate in Table I. We also do not claim to have solved the full pipeline of the IROS 2016 theme, "Road to companionship with intelligent robots in everyday life and workspaces", which also includes speech recognition, vision, action planning and other problem solving. Instead, we leave these problems to the respective experts, and assume that they are sufficiently solved for our purposes. However, we do claim that our approach is an important part of that pipeline, and that it contributes to an key aspect of robot companionship, namely exploiting deep semantics in natural language to facilitate understanding.

[1]International Computer Science Institute, Berkeley, USA {eppe, seantrott, feldman} @icsi.berkeley.edu

II. PRELIMINARIES AND RELATED WORK

A. What Makes Natural Language Understanding Hard

Natural Language Understanding is generally still unsolved. However, we can exploit the fact that robotics usually requires only a finite, albeit large, domain description, so that grammar-based approaches with a controlled language subset are applicable. Under these circumstances, one can constrain the focus on the following specific solvable issues:[1]

a) Compositionality and the combinatorial explosion of form and meaning: Word meaning is usually heavily underspecified and contextual. As an example, consider a motion word like "move" which has a very different semantics in, e.g., the transitive expression "move the table" vs. intransitive "move to the table". Often, the appropriate meaning is determined by the construction, in this case transitive vs. intransitive, in which it appears [22]. This does not only lead to a combinatorial explosion of form and meaning, but it also allows one to invent novel and uncommon word meanings spontaneously. For example, the expression "to sneeze the napkin off the table" imposes a movement semantics on the verb "sneeze", which is rare but absolutely legitimate.

Under the condition that enough data is available, learning based approaches are attractive to solve the combinatorial problem [4]. However, they can not deal with the novel inventions depicted above. They also have no proper way of dealing with context. For example, "move" means something quite different in a chess game or in changing houses.

To understand a large language subset, including novel and previously unheard expressions, while also being robust to misinterpretation, one requires a more sophisticated and modular way to compose grammatical primitives. This is exactly what *construction grammars* do [22], as we describe in Section II-B.

b) Reference resolution and ambiguity: Reference resolution means to identify an object in the real world that is referred to by a pronoun, as e.g. in the sentence "If there is a cup on the dining table, please bring it to me". Cognitive linguistics tells us that clearly the "it" refers to the cup and not to the table, because the "cup" is the head noun and the dining table is the subordinate description [22]. In contrast, consider "If there is a table under the cup, please bring it to me". Here, the "it" clearly refers to the table. Note that in addition it carries the implicit information that the speaker expects the hearer to be capable of carrying the table. Such deep semantics can be exploited for HRI.

In general, reference resolution consists of two steps. Firstly, *anaphora resolution* identifies the noun to which a pronoun refers within a sentence. Secondly, *grounding* identifies the object in the real world to which the noun in the sentence refers. In cases where grounding can not disambiguate the reference, clarification dialogs are required. As an example, consider the scenario in Sec. IV-A, where the DARwin-OP needs to resolve an ambiguity to decide which object to pick up (in this case the blue marker and not the red one).

[1]There are also other issues with Language, such as temporal aspects, situatedness and other problems, as pointed out e.g. in [35]. Furthermore, Grounding and perception also play a big role for HRI, as described e.g. in [45]. For brevity reasons we do not consider these topics in this paper.

c) Conditionals: Conditionals are another crucial building block of verbal articulation, which usually follow an *if-then-else* pattern, with the *else* being optional. For illustration consider our scenario with the PR2 robot in Sec. IV-A

d) Erroneous input and ungrammatical sloppy language: Input text is often erroneous in different ways. For example, when given as a transcript from speech recognition systems, as in "Please bring me the grass" instead of the correct "Please bring me the glass." Furthermore, users tend to often use jargon and ungrammatical or sloppy language.

e) Disfluency analysis and repair: Humans often abandon sentences and words midway during the speaking, switching to other conversation segments or correcting sentences on the fly. This makes language disfluent and requires repair. For example, "There is a blue kit, erm, I mean a blue box, at the end of the table" can be simplified to "There is a blue box at the end of the table" [11].

f) Indirect assertions through relative clauses and appositions: Relative clauses and appositions are often used for indirect assertions of world properties. As an example consider: "The noodles, which are still in the pot on the stove, must be wasted by now; bring them to the trash!" This sentence encodes the information about the location of the noodles in an embedded relative clause, which needs to be interpreted correctly before the command can be executed.

g) Modalities: Modalities are used to express ability, knowledge, temporality and other mental attitudes. For example, "Are you able to order pizza?" is not a question about the pizza but about the ability to order pizza (imagine an assistance robot that is connected to an online pizza service).

h) Indirect speech acts: These are often used for politeness and sociality. For example, instead of saying "Open the window!", one often uses the more polite "Can you open the window?" While this is literally a modality question about the hearer's abilities, a robot should instead treat it as a command. It is an issue to decide when a robot should interpret a modality literally or as indirect speech [48].

i) Interlocutor feedback: During a dialog, humans constantly give feedback to each other. For example, a hearer (H) often uses "[o]kay" to notify the speaker (S) that he understood the last part of an instruction correctly. S: "so turn right" H:"kay" S: "and walk a little bit and turn right again"[11]. Feedback can also be nonverbal (gaze, pointing), but that is out of our work's scope.

j) Metaphor: Metaphor is much more than an instrument of poetry and arts. Humans frequently use metaphors in language, often even unconsciously, and there is significant evidence suggesting that metaphor plays a central role in abstract thought [27]. Therefore, metaphor is also crucial to understand language, and there has been extensive work on the cognitive linguistics of metaphor [15]. However, this has not yet been applied to human robot interaction.

k) Integration of Robots with NLU systems: Connecting a NLU system to a robotic platform is not trivial. Firstly, NLU is a problem which is completely independent from robotic problems like action and motion planning, so it should be independent from the robot that one wants to use. Therefore, one should integrate an NLU system into a

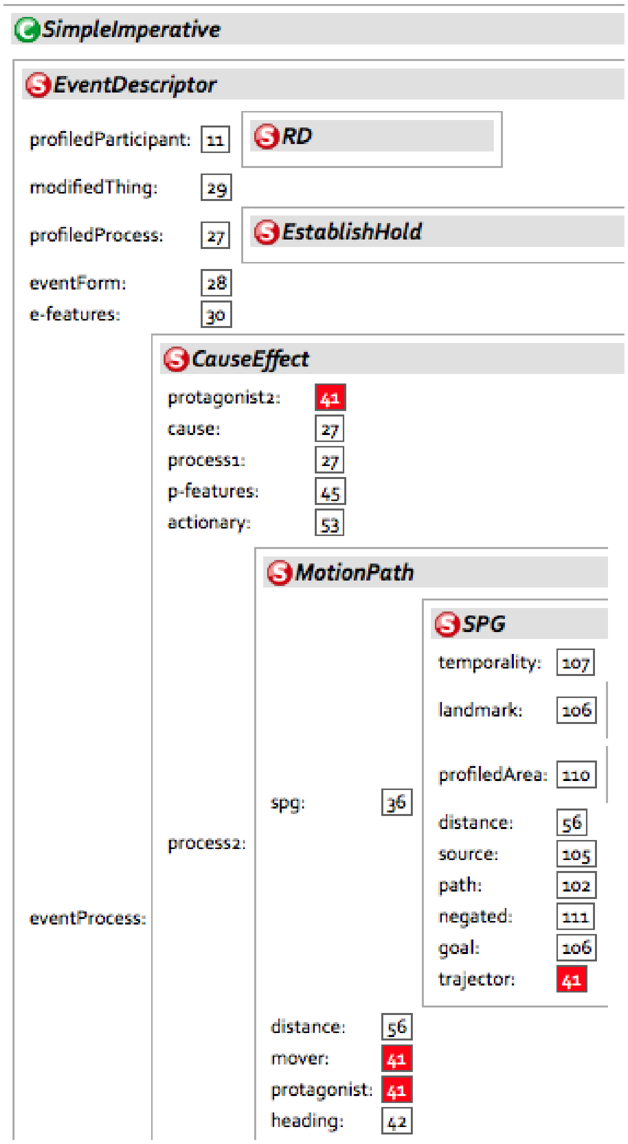

Fig. 1: Excerpt from SemSpec for utterance "PR2, bring the soda can to the dining table!" .

robotic framework that is modular, such as ROS. Secondly, it is problematic to interpret the semantic specification of the NLU analysis as robot commands. For example, after analyzing the sentence "if there is tea on the table, please bring it to me", one needs to find an appropriate representation that is suitable as input for a robot's problem solving mechanism. In the case of the above sentence, we have an epistemic planning problem with possibly incomplete world knowledge (see e.g. [17], [16]). The NLU system needs to communicate with the problem solver on an abstraction layer that is appropriate wrt. the underlying reasoning problems.

B. Embodied Construction Grammar and Compositionality

In linguistics, there are several kinds of compositionality. The most simple one is lexical compositionality, and it is the key feature of Chomsky's context-free grammars, which most existing NLU system for robotics use [33]. A problem with this approach is that it does not have adequate coverage, as described in Sec. II-A, item a). Construction grammar involves the pairing of form and meaning and allows one to model higher-level linguistic concepts, such as transitivity ("move the table") and intransitivity ("move to the table"). This is done in an abstract way that is independent from the actual verb in the respective construction.

Because such higher level constructions can be commonly used across all domains, the grammars remain compact, highly generalizable and productive. Hence, one can easily increase the size of the controlled language subset for the particular domain without much effort, while still having the advantage of being more reliable than systems based on learning. Note also that higher-level linguistic concepts like transitivity are often similar or even identical for many languages, which makes it easier to develop a grammar for a new language by starting from another language.

Embodied Construction Grammar (ECG) [18] goes one step further and not only provides a way to compose constructions, but also concepts. These concepts are neurologically and cognitively motivated in accordance with the established theory about *image schemas* [28].

As an example, consider the Semantic Specification (SemSpec) in Figure 1, a reduced semantic analysis of the utterance "PR2, bring the soda can to the dining table". The motion verb "bring" triggers the *EstablishHold* schema, which suggests that some *protagonist* must grasp an *actedUpon* object. The transitive construction ("bring the soda can") followed by a prepositional phrase ("to the dining table") triggers the *CauseEffect* schema, in which the *affectedProcess* is *MotionPath*. This in turn evokes the *Source Path Goal* (SPG) schema, which implies that there must exist some trajectory from a source location via a path to a goal location, and there must exist a trajector that follows this path. Conceptual compositionality is illustrated through co-indexing of roles both within and across schemas. Note that the *mover* role in the *MotionPath* schema is co-indexed and identified with the id "41" in the *trajector* role in the SPG schema, as well as the *affectedEntity* role in the *CauseEffect* schema. The SemSpec represents the conceptual network involved in understanding this utterance, and the relations between each component of this conceptual network. A crucial feature of this approach is that schemas are language independent. For example, the concept of *Source Path Goal* is universal in all languages and cognitively motivated [28].

A computational framework that can analyze language according to the ECG theory has been developed over the last decade [10], [19], [37], [24], [46]. The framework has been used in understanding and deep meaning of a wide range of complex constraints in multiple languages [24], [46], [29] beyond those described in this article.

C. A Brief Survey on NLU for Robotics

In the following, we present a survey on related work on knowledge-based approaches for NLP and NLU in robotics. For an introduction to statistical semantic parsing, consider e.g. the survey in [13]. A summary of our findings can be found in the evaluation in Section IV-B, Table I.

The work in [11] is probably the most relevant for our approach. The authors describe the NLU module for the *DIARC* HRI-System [34]. They perform Wizard-of-Oz experiments

[36] to identify several issues that are crucial for *natural* language understanding for HRI, especially when using speech input. The authors build a system that can deal with a considerable part of the linguistic problems described in Sec. II-A. Specifically, their architecture is capable of speech recognition, semantic analysis, disfluency analysis and reference resolution. For the semantic analysis, the authors use Combinatory Categorical Grammar (CCG) and lambda conversions [40]. A probabilistic extension, where the authors use a statistical Dempster-Shafer-theoretic approach to deal with natural language for robotics is presented in [48]. This approach focuses on intention detection and indirect speech acts. It does not explicitly mention modalities, but modalities are obviously often used in indirect speech acts, e.g., in "Do you know what time it is?". However, it remains unclear whether the authors' system can analyze modalities in the case where they are not used indirectly. Conditional statements are also not mentioned in the articles. Grounding for reference resolution is possible, albeit it remains unclear how the anaphora resolution is performed.

The authors of [26], [25] focus on dialog. They represent the semantics of an utterance in a categorical modal-logical form, based on Combinatory Categorical Grammar (CCG) [40]. Their system can perform reference resolution and starts a clarification dialog if the reference is too ambiguous. For disfluency analysis, the authors use contextual knowledge to prime utterances. Verbal feedback with words like "okay" or "fine", can also be handled. The authors present an implementation which they use to perform experiments with impressive results, but do not demonstrate how the system is connected to real or simulated robots in a modular manner.

Another construction grammar which is also often used in Robotics is Fluid Construction Grammar (FCG) [43], [42], [38]. For the most part, the goal of existing work around FCG in robotics is not Human-Robot-Interaction, but to provide a model for the evolution of language in robotic communities through so-called language games [41]. However, the authors also present the *Talking Heads Experiment*, which shows that the emergence of lexico-semantic models in robots can also be combined with human interaction, such that human language is acquired naturally by robots [42]. The work on FCG has been extended in [7] to cope with erroneous input, disfluencies and repair. Another nice approach [39] demonstrates how logical reasoning about spatial relations makes language processing more robust.

The approach of [8] uses Discourse Representation Theory (DRT) [23] to capture semantics in natural language. An important aspect of DRT is that it can be compiled into first-order logical Discourse Representation Structure (DRS), which allows one to perform problem solving and higher level reasoning. DRS is compositional, and it also allows for reference resolution and related problems that result from the deep semantics of natural language. The authors "[...] used a generic domain-independent, but linguistically motivated grammar as a starting point" [8], but it remains unclear how rich the grammar is and in how far it supports linguistic problems like conditionals, transitivity, and other advanced linguistic argument structure for reasoning.

The information-theoretic probabilistic approach by [44], [45], [14], [47] seeks to minimize uncertainty of commands by asking questions that maximize the expected disambiguation. Hence, the focus is on dialog and clarification. The system is based on learning and so-called Spatial Description Clauses and Generalized Grounding Graphs [44]. The constructional and conceptual compositionality aspect of language is not mentioned in the paper, and linguistic problems like conditionals, appositions, relative clauses, etc. are not specifically addressed.

Recent work presented in [12] focuses on the problem that humans and robots do not have a common perceptual ground, which is important for dialog. The problem is that the background knowledge and the object recognition capabilities of robots are far behind human level, so that grounding objects in the real world is difficult. In this context, the authors investigate the additional effort that is required to establish a common perceptual ground between robot and human. The authors demonstrate their approach with an implementation and experiments on humanoid NAO robots.

Impressive recent work by [3] presents a supervised learning method to let a a robot learn word meaning for navigation purposes and spatial relations. The authors emphasize their use of continuous domains, instead of using symbolic primitives like "Drive to Location 1". The language model the authors use is based on a simple non-compositional grammar that can deal with relative sentences and some intermediate grammatical structure.

[4] describe a method to interpret natural language using frame semantics [20], [2], [32]. Their framework allows them to correctly parse simple sentences like "take the book on the table" by mapping them to a form they call *Abstract Meaning Representation* (AMR), which in turn can be mapped to robot commands. The grammar they use is context-free. The authors elaborate their work in [6], where they compare a grammar-based NLU system with a statistical learning one, and combine the two to obtain a hybrid system. The authors do not demonstrate how they analyze conditionals, relative sentences, modalities and other rich linguistic structure. They handle disfluencies by adding wildcards for "please" or "erm" to their grammar. A notable contribution is also the integration of different corpora, namely the RoboCup@Home [49], the Speaky4Robot [1] corpora, and another grammar generated corpus, which can be used for the evaluation of NLU systems [5]. However, a problem is that these corpora contain mostly very clean and simple sentences without conditionals, anaphora resolution or other problems that we describe in Section II-A

Another approach that focuses on navigation is presented in [30]. The authors use a corpus of natural language text to follow instructions, and show that their agent follows navigation commands nearly as precise as humans. However, the system is restricted to navigational language and does not use deeper semantics or compositional grammar. Accordingly, no deep semantic expressions like conditionals or metaphors in language is accounted for, and reference resolution is also not mentioned. The authors perform experiments in a simulated

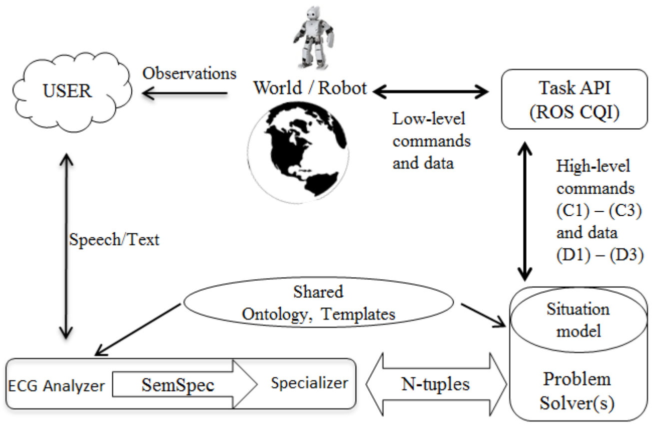

Fig. 2: ECG analysis architecture.

environment.

The work presented in [31] describe the learning of a parser to infer robot commands from natural language. Towards this, the authors define a Robot Control Language (RCL) that is used to ground natural language in robot actions (specifically navigation actions). They learn a probabilistic version of Combinatory Categorial Grammar (CCG) and map it to RCL. CCG features constructional compositionality, which would make it easy to also deal with conditionals or other rich structure in language, but we have not found the use of such structure in the paper.

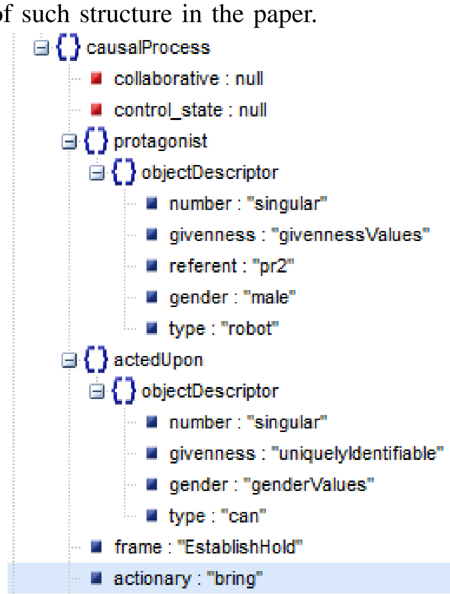

Fig. 3: N-tuple excerpt for input utterance: "PR2, bring the soda can to the dining table!", Here, "PR2" is automatically identified as protagonist in a causal process that executes a "bring" action upon an object of type "can".

III. SYSTEM OVERVIEW

This section describes our ECG analyzer and its integration with ROS. The overall system consists of two nested parts. Firstly, there is the ECG analyzer which takes natural language as input and gives high-level robot commands as output. Secondly, we describe how the ECG analyzer is nested and embedded within ROS. We provide an additional illustration of the system in the supplementary video.

A. The ECG semantic analysis framework

The semantic analysis is illustrated in Figure 2. Firstly, the ECG Analyzer [10] uses an ECG grammar to parses a sentence and perform a best-fit analysis which produces so-called *Semantic Specifications*, or *SemSpecs*. We depict a partial example of a SemSepc in Fig. 1. To transform the semantic specifications into actual robot commands, several intermediate steps are necessary. The second step is performed by the *Specializer*, which crawls the SemSpec and identifies task-relevant information that is shared between the Specializer and the *Problem Solver*. The output is a data structure that we call an *n-tuple*. An n-tuple consists of nested key-value pairs that are used as a communication language between the Specializer and the Problem Solver. N-tuples can i) specify commands, ii) represent a query, or iii) assert embedded knowledge about the world. An example excerpt of a command-type n-tuple is provided in Figure 3. N-tuples are generated using templates that are shared among the ECG Analyzer, Specializer and Problem Solver, and are aligned to a shared ontology. The sharing of structure is crucial for the interoperation between these three parts, e.g. when it comes to generating a clarification dialog.

Commands are to be directly executed by a robot, such as the command: *"PR2, bring the soda can to the dining table!"*.

Queries are user questions to retrieve information about the surrounding or the robot state. For example: *"Which marker is blue?"*.

Assertions assign a value to a world property that is unknown to the Situation Model. For example: *"The marker is under the table."* Since we focus only on language understanding in this work, we assume that all assertions made by a user are correct, and neglect belief revision and related epistemic issues.

Next, the *Problem Solver* receives the n-tuple and determines what course of action to take. If the user poses a question, the Problem Solver answers it; if the user orders the robot to carry out a task, the Problem Solver uses reasoning to successfully fulfill the user's request. This requires access to information about the world, which the Problem Solver represents as a Situation Model (see Figure 2). The Situation Model ranges in complexity from the locations of objects, to spatial relations and properties of these objects.

B. Integration of the ECG analyzer in ROS

To integrate the ECG analyzer software with ROS, we build two ROS packages. One is the ECG wrapper package which embeds the ECG analyzer to translate natural language into high-level robot commands. The other one is the Command and Query Interface (CQI) which provides a robot-specific abstraction layer to translate high-level commands to low-level motor control.

ECG wrapper. The ECG analyzer and solver are accessed by a ROS wrapper package which is responsible for publishing high-level commands to the CQI via topic `/cqi/command/`, and which receives feedback and high-level sensor data from the CQI via topic `/cqi/data/`. For now, we focus mainly on object manipulation. Hence,

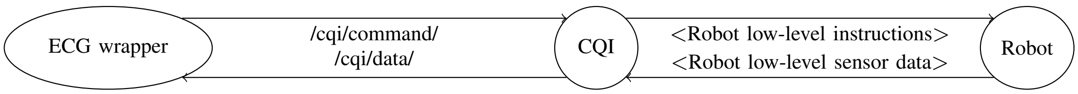

Fig. 4: ECG-ROS communication flow

the set of commands that we currently support are moving (C1), grasping (C2) and releasing objects (C3). This set could easily be extended.

(C1) `move_to_pose(x, y, θ)`
(C2) `grasp_object(object_label)`
(C3) `release()`

It is the responsibility of the ECG problem solver to ensure that moving from the robot's current pose to the destination pose is possible, i.e., that there exists a feasible trajectory between start and destination pose. The solver also has to assure that the object label of an object is known, and that, before grasping, the robot is in a pose which makes the grasping possible (i.e., not too far away from the object and oriented towards it, and not already holding another object). The latter can be relaxed when the grasp action is implemented robustly, such that the robot internally fine-tunes its pose before performing the actual grasp.

After a command is executed, ECG's problem solver expects feedback to determine whether an action was successfully finished. This feedback is received via messages of the form (D1) – (D3) on ROS topic `cqi/data/` from the CQI, as described in the following.

Command and Query Interface (CQI). The CQI is a modular interface between the low-level motor commands required by the robot hardware and the high-level action commands given by ECG's Problem Solver. For example, the CQI is supposed to translate a high-level `move_to_pose` command to an appropriate low-level behavior that involves locomotion, navigation and obstacle avoidance. So far, we have built two simple CQI modules, one for the DARwin-OP, and one for the PR2 robot, with ad-hoc solutions for navigation, object recognition and motion planning. The fixed high-level input interface (C1) – (C3) allows one to use the ECG framework with any robot for which a CQI module exists. Implementing a new CQI module is relatively simple, given the various open-source packages for moving, grasping, localization, etc. that are available for ROS. However, our focus here is more on the interface than on the implementation of the low-level behavior, and we leave a more sophisticated implementation to the respective robotics experts. The interface is provided as a base class in Python, from which robot-specific CQI's inherit. The communication flow is implemented as usual with the ROS-internal topic-based communication paradigm, and depicted in Fig. 4.

The possible data messages that the CQI can currently publish are as follows:

(D1) `at_pose(x,y,θ)` This is a dedicated data message to communicate the robot's current position and rotation in 2d. It is published continuously, and therefore inherently encodes success and failure of a `move_to_pose` action.

(D2) `holding(object)` The message communicates which object the robot is currently holding, returning `none` if no object is currently held, e.g., when grasping was unsuccessful.

(D3) `has_property(object, property, value)` This is a more general data message that can be used to talk about object properties in general. Examples are color, shape and location of objects in the environment, that the robot determines with its vision system.

The data interface can be extended as required by a specific scenario, to meet the capabilities of specific robots. Note that the robot pose and holding information could in principle also be encoded by (D3), but we prefer separate kinds of messages that use different communication properties, such as the publishing rate, which is higher for the pose and lower for the holding.

IV. PROOF OF CONCEPT AND EVALUATION

To provide a proof of concept, we present scenarios two indoor assistance scenarios, with additional technical detail in a supplementary video[2]. We also present a qualitative comparison of the state of the art in NLU for HRI.

A. Proof of Concept Scenarios

To demonstrate the NLU-system, we present two scenarios from the Assisted Living domain in an environment as depicted in Figure 5. Herein, we focus on natural language problems, and assume that reasonably good speech recognition, indoor navigation and object recognition methods by respective experts are available. We currently have integrated some basic ad-hoc solutions for these problems within ECG's problem solver and the CQI.

Scenario 1: *Anaphora resolution and conditionals with PR2.* The first scenario is triggered by the sentence "PR2, if a soda can is on the kitchen counter, please bring it to the dining table, otherwise get a new one from the fridge". Here we have a conditional if-then-else command with reference resolution. Figure 5 shows that our simulated PR2 has found the can of soda and is driving to the dining table.

Scenario 2: *Spatial clarification dialog for reference grounding with DARwin-OP.* Here we show how our system deals with the sentence "Darwin, pick up the marker under the table" (see Fig. 5). Darwin knows that there are two markers which fell down from the table, a blue and a red one. Hence, it tries to resolve this ambiguity by asking for clarification: "Which one?". In this case we answer "The blue one" and Darwin walks to the table to pick the blue marker up. Note that such clarification dialogs could also involve multiple steps. For example, if there were two blue markers of different sizes, Darwin would continue to resolve the ambiguity by asking for the size.

[2] https://www.youtube.com/watch?v=BAqoZEi1IVA

Feature	Example	Supported by										
		[X]	[31]	[11]	[5]	[3]	[30]	[43]	[8]	[25]	[12]	[45]
Conceptual and constructional compositionality	Abstract transitive and ditransitive constructions affect the meaning of verbs.	✓	(✓)	(✓)	-	-	-	(✓)	(✓)	(✓)	(✓)	-
Conditionals	“If there is a can of coke on the table, please bring it!”	✓	-	-	-	-	-	(✓)	(✓)	-	-	-
Clarification and Dialog	“Please pick up the marker under the table!” – “Which one, the red one or the blue one?”	✓	-	✓	-	-	-	✓	-	✓	✓	✓
Indirect assertions via relative sentences or appositions	“Please bring the plate, which fell from the table, to the dishwasher.”	✓	-	-	-	✓	-	-	(✓)	✓	-	-
Modalities	“Are you able to order pizza?”	✓	-	-	-	-	-	-	(✓)	-	-	-
Metaphor	“Robot, fly over here!” (move fast)	-	-	-	-	-	-	-	-	-	-	-
Indirect speech acts	“Do you know where I have left my keys?”	-	-	✓	-	-	-	-	-	-	-	-
Sloppy or erroneous input	Erroneous input: “Please bring me the grass” correct input: “Please bring me the glass”	-	-	✓	-	-	-	✓	-	-	-	-
Disfluency analysis and repair	“There is a blue kit, erm, I mean a blue box, at the end of the table”	-	-	✓	(✓)	-	-	✓	-	✓	-	-
Verbal interlocutor feedback	“So turn right” – “okay” – “and walk a little bit and turn right again”	-	-	✓	-	-	-	-	-	(✓)	-	-
Robot implementation		✓	-	✓	-	✓	(✓)	✓	✓	-	✓	✓

TABLE I: Features of different language understanding frameworks for robots ([X] represents this work).

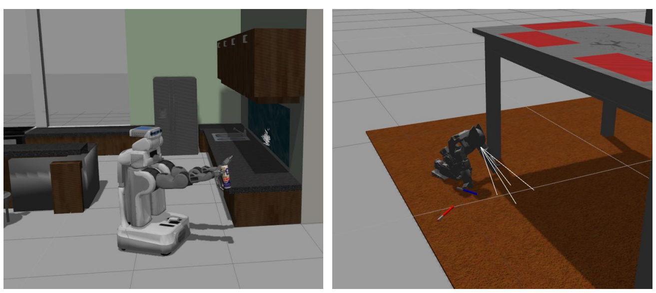

Fig. 5: PR2 carrying a can of soda in the kitchen environment (left) and DARwin-OP picking up the blue marker (right).

B. Comparison of Features of NLU-Systems for HRI

To evaluate our work wrt. the state of the art, we look at the features of related systems found in literature and compare it to our system. Results are depicted in Table I. We not only highlight the capabilities of our system, but also looked at the literature and compiled the features that were highlighted together. Cases marked as (✓) are borderline, e.g. when constructional but no conceptual compositionality is supported. These cases are explained in Section II-C.

While Table I provides a good overview on the state of the art, it is not intended to reflect details about the depth of the individual author’s focus. For example, [45] focus heavily on clarification and dialog, and this aspect of their work is much more sophisticated than our clarification methods. However, we think that a successful HRI system should be able to capture as many of the listed features as possible to be successful. We also do not claim that the table is exhaustive, but we believe that it includes the vast majority of problems that occur in NLU for robotics.

V. CONCLUSION

We demonstrate the importance of capturing deep semantics in language for robotic applications using examples from the indoor assistance domain. Herein, we connect the ECG analyzer tool to ROS and extend the ECG core grammar towards more robot-specific tasks. As a proof-of-concept, we implement two scenarios which illustrate a subset of the capabilities of the ECG analyzer. To realize our toolchain, we provide a modular abstraction layer (the CQI), which maps high-level actions like grasp and move to robot-specific low-level motor commands. So far, we have implement basic CQI modules for the PR2 and the DARwin-OP robot.

We also present a survey on NLU for robotics, which illustrates shortcomings of our and other approaches. The survey supports our hypothesis: By building on cognitive linguistic theories, in the form of core ECG schemas and constructions, we are able to analyze a combination of several kinds of sentences that no other NLU system for robotics can currently interpret correctly and reliably. We demonstrate this in our scenarios, with conditional sentences in combination with anaphora resolution and grounding.

Our approach is straight-forward to extend to other not yet supported problems that we depict in Table I. For example, indirect speech acts could be resolved by adding an intention detection mechanism to the problem solver, as realized by [48], and similarly for erroneous input. There is also ongoing work to make the ECG core grammar capable of understanding metaphor [15].

In addition to extending the ECG core functionality, we want to investigate how learning-based approaches for language understanding could leverage our knowledge-based approach, and vice versa. Specifically, we think that compositionality makes learning easier because it allows one to maintain a universal core grammar set for learning new constructions and words. Hence, by building high-level, domain-independent semantics, we can expand to new domains in a semi-automated way. The data that is needed for the learning can be extracted from Wizard-of-Oz experiments, but also from existing resources like FrameNet [2] to build domain-specific low-level constructions and tokens that naturally compose with the core grammar.

ACKNOWLEDGMENT

This work is supported by the Office of Naval Research grant number N000141110416 and a research grant from Google. Manfred Eppe received support from the German Academic Exchange Service (DAAD) via the FITweltweit program.

REFERENCES

[1] Luigia Carlucci Aiello, Emanuele Bastianelli, Luca Iocchi, Daniele Nardi, Vittorio Perera, and Gabriele Randelli. Knowledgeable talking robots. *Lecture Notes in Computer Science - AGI*, 2013.

[2] Collin Baker, Charles Filmore, and John Lowe. The Berkeley FrameNet project. In *International Conference on Computational Linguistics (COLING)*, 1998.

[3] Daniel Paul Barrett, Scott Alan Bronikowski, Haonan Yu, and Jeffrey Mark Siskind. Robot Language Learning, Generation, and Comprehension. 2015.

[4] Emanuele Bastianelli, Giuseppe Castellucci, Danielo Croce, and Roberto Basili. Textual Inference and Meaning Representation in Human Robot Interaction. In *Joint Symposium on Semantic Processing*, 2013.

[5] Emanuele Bastianelli, Giuseppe Castellucci, Danielo Croce, Roberto Basili, Daniele Nardi, and Luca Iocchi. HuRIC : a Human Robot Interaction Corpus. In *LREC*, 2014.

[6] Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Roberto Basili, and Daniele Nardi. Effective and Robust Natural Language Understanding for Human Robot Interaction. In *European Conference on Artificial Intelligence*, 2014.

[7] Katrien Beuls, Remi Van Trijp, and Pieter Wellens. Diagnostics and repairs in fluid construction grammar. In *Language Grounding in Robots*. Springer, 2012.

[8] Johan Bos and Tetsushi Oka. A spoken language interface with a mobile robot. *Artificial Life and Robotics*, 11(1):42–47, 2007.

[9] Cynthia Breazeal. Jibo, the world's first family robot.

[10] John Edward Bryant. *Best-Fit Constructional Analysis*. PhD thesis, University of California at Berkeley, 2008.

[11] Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and Xuan Wu. Robust spoken instruction understanding for HRI. In *International Conference on Human-Robot Interaction (HRI)*, 2010.

[12] Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. Collaborative effort towards common ground in situated human-robot dialogue. In *International Confernce on Human-Robot-Interaction (HRI)*, 2014.

[13] David L. Chen and Raymond J. Mooney. Learning to Interpret Natural Language Navigation Instructions from Observations. *AAAI Conference on Artificial Intelligence*, (August):859–865, 2011.

[14] Robin Deits, Stefanie Tellex, Pratiksha Thaker, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Clarifying Commands with information-Theoretic Human-Robot Dialog. *Journal of Human-Robot Interaction*, 2012.

[15] Ellen Dodge, Jisup Hong, and Elise Stickles. MetaNet : Deep semantic automatic metaphor analysis. In *Workshop on Metaphor in NLP, at NAACL*, 2015.

[16] Manfred Eppe and Mehul Bhatt. A History Based Approximate Epistemic Action Theory for Efficient Postdictive Reasoning. *Journal of Applied Logic*, to appear, 2015.

[17] Manfred Eppe, Mehul Bhatt, and Frank Dylla. Approximate Epistemic Planning with Postdiction as Answer-Set Programming. In *International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR)*, 2013.

[18] Jerome Feldman, John Edward Bryant, and E Dodge. Embodied Construction Grammar. In *The Oxfrod Handbook of Computational Linguistics*, pages 38 – 111. Oxford University Press, 2009.

[19] Jerome Feldman, Ellen Dodge, and John Bryant. A neural theory of language and embodied construction grammar. *The Oxford Handbook of Linguistic Analysis.*, pages 111 —- 138, 2009.

[20] Charles Fillmore. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254, 1985.

[21] Open Source Robotics Foundation. Ros.org.

[22] Adele Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, 1995.

[23] Hans Kamp, Josef van Genabith, and Uwe Reyle. Discourse Representation Theory. In *Handbook of Philosophical Logic*, pages 125 – 394. 2011.

[24] Huda Khayrallah, Sean Trott, and Jerome Feldman. Natural Language For Human Robot Interaction. In *International Conference on Human-Robot Interaction (HRI)*, 2015.

[25] Geert-Jan Kruijff, Pierre Lison, Trevor Benjamin, Henrik Jacobsen, Hendrik Zender, and Ivana Kruijff-Korbayová. Situated dialogue processing for human-robot interaction. *Cognitive Systems*, 2010.

[26] Geert Jan Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik Christensen. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems*, 2007.

[27] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

[28] George Lakoff and Mark Johnson. *Philosophy in the Flesh*. Basic Books, 1999.

[29] Peter Lindes and John E Laird. Toward Integrating Cognitive Linguistics and Cognitive Language Processing. In *International Conference on Cognitive Modeling (ICCM)*, 2016.

[30] M MacMahon, Brian Stankiewicz, and Benjamin Kuipers. Walk the talk: Connecting language, knowledge, and action in route instructions. In *AAAI*, 2006.

[31] Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. *IntâĂŹl Symposium on Experimental Robotics (ISER)*, 2012.

[32] Miriam Petruck. Frame Semantics. In *Handbook of Pragmatics*. John Benjamin Publishing Company, 1996.

[33] Geoffrey Pullum and Gerald Gazadar. Natural languages and context-free languages. *Linguistics and Philosophy*, 1982.

[34] Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Tom Williams, and Richard Veale. Novel mechanisms for natural human-robot interactions in the diarc architecture. In *AAAI*, 2013.

[35] Matthias Scheutz, Rehj Cantrell, and Paul Schermerhorn. Toward Humanlike Task-Based Dialogue Processing for Human Robot Interaction. *AI Magazine*, 32(4):77–84, 2011.

[36] Matthias Scheutz and Kathleen Eberhard. Towards a framework for integrated natural language processing architectures for social robots. In *International Workshop on Natural Language Processing and Cognitive Science*, 2008.

[37] Steve Sinha. *Answering Questions about Complex Events*. PhD thesis, University of California at Berkeley, 2008.

[38] Michael Spranger and Luc Steels. Co-Acquisition of Syntax and Semantics âĂŤ An Investigation in Spatial Language. In *International Joint Conference on Artificial Intelligence*, 2015.

[39] Michael Spranger, Jakob Suchan, and Mehul Bhatt. Robust Natural Language Processing âĂŤ Combining Reasoning , Cognitive Semantics , and Construction Grammar for Spatial Language. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[40] Mark Steedman. *The syntactic process*. MIT Press, 2000.

[41] Luc Steels. The origins of ontologies and communication conventions in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 1(2):169–194, 1998.

[42] Luc Steels. *The Talking Heads Experiment*. 2015.

[43] Luc Steels, Joachim De Beule, and Pieter Wellens. Fluid Construction Grammar on Real Robots. In *Language Grounding in Robotics*. Springer, 2012.

[44] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI Conference on Artificial Intelligence*, 2011.

[45] Stefanie Tellex, Pratiksha Thaker, Robin L H Deits, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Toward Information Theoretic Human-Robot Dialog. *Robotics: Science and Systems Conference*, 2012.

[46] Sean Trott, Aurélien Appriou, Jerome Feldman, and Adam Janin. Natural Language Understanding and Communication for Multi-Agent Systems. In *AAAI Fall Symposium*, pages 137–141, 2015.

[47] Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. Learning Semantic Maps from Natural Language Descriptions. *Robotics Science and Systems*, pages 1–8, 2013.

[48] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. Going Beyond Literal Command-Based Instructions : Extending Robotic Natural Language Interaction Capabilities. In *AAAI*, 2015.

[49] Thomas Wisspeintner, Tijn van der Zant, Luca Iocchi, and Stefan Schiffer. RoboCup@Home: Scientific Competition and Benchmarking for Domestic Service Robots. *Interaction Studies*, 10(3):392–426, 2009.