What are you talking about? Grounding dialogue in a perspective-aware robotic architecture

Séverin Lemaignan, Raquel Ros, Rachid Alami, Michael Beetz

What are you talking about? Grounding dialogue in a perspective-aware robotic architecture

Séverin Lemaignan*†, Raquel Ros*, Rachid Alami*, Michael Beetz†

*CNRS - LAAS, Université de Toulouse, UPS, INSA, INP, ISAE, LAAS, F-31077 Toulouse, France
{slemaign, rrosespi, rachid}@laas.fr

†Intelligent Autonomous Systems, Technische Universität München, Munich, Germany
{lemaigna, beetz}@in.tum.de

***Abstract*—While key for human-robot interaction, natural language interpretation is a notoriously difficult task, especially because the interaction context is at the same time essential for dialogue understanding, difficult to build for machines, and depends on each speaker point of view. However, robots as embodied artifacts, can perceive their environment and interactors, and hence compute symbolic models from various perspectives. This allows in turn to build symbolic contexts for dialogues. In this paper, we introduce DIALOGS, a component for natural language interpretation that relies on these structured symbolic models of the world to ground verbal interaction.**

I. GROUNDING HUMAN INTERACTION INTO THE ROBOT KNOWLEDGE

A. Situated speech acts

A messy table, covered with cardboard boxes, books, video tapes... Thomas is moving and packs everything with the help of Jido, its robot.

" – Jido, give me this", says Thomas, looking at a box that contains a video tape. The robot smoothly grasps the tape, and hands it to the human.

While this kind of interaction should hopefully sound quite familiar in a foreseeable future, our robots are not yet quite up to the task. Neither regarding natural language understanding nor plan-making and manipulation.

To be combined together, those abilities require an unambiguous and shared representation of concepts (objects, agents, actions...) underlying the interaction: what are the prerequisites for such a human sentence — "Jido, give me this" — to be understood by the robot, correctly interpreted in the spatial context of the interaction, and ultimately transformed into an action?

Austin [1] would have at first glance analyzed such kind of sentence as a *speech act*, comprising of *locutionary*, *illocutionary* and possibly *perlocutionary* acts. First, we want to understand the direct meaning of the sentence (*locutionary act*): we must acquire the sentence, convert it into a useful syntactic form (quite probably by mean of speech recognition), and understand the semantics of the sentence, *i.e.* , What is refered by "*Jido*"? What is "*give*"? What is "*me*"? And "*this*"?

Working in a situated context, we want furthermore to *resolve* these semantics atoms, *i.e.* ground them in the sensory-motor space of the robot. For instance, "*this*" is a demonstrative pronoun that refers in this context to the object the human is focusing on, whatever *focusing* means: here, Thomas is looking at something, which is a possible cue. But it could as well point at something or refer to some previously mentioned concept.

Fig. 1. Interacting with the robot in an everyday setup: the human asks for help in vague terms, the robot takes into account the human's spatial perspective to refine its understanding of the question.

Second, the *illocutionary force*, *i.e.* the *intent* of the utterance as thought by the agent must be extracted, and understood. In our example, Thomas obviously wants an action to be performed by the robot. The action parametrization is conveyed by the semantics attached to the words and the grammatical structures of the sentence. In our example, the type of action is given by the verb "*give*". Assuming the robot has some procedural knowledge attached to this symbol, the action type can be considered as grounded for the robot. We can as well understand that the recipient of the action is the human, the performer is the robot itself, and the object acted upon is the tape. These are the basic *thematic roles* [2] that can be extracted from the sentence that allow to fully ground the action.

B. Building a symbolic model

Extracting these speech acts and turning them into a content processable by the robot is a difficult challenge in the general case. We base our approach on three distinct, inter-related cognitive functions:

1) Physical environment modeling and *spatial reasoning* (grouped under the term *situation assessment*) are in charge of building and maintaining a coherent model of the physical world. This model is realistic in the sense that it relies on accurate 3D models of both manipulated objects and humans. It also has dedicated mechanisms to manage disappearing or occluded objects. The geometric model is used to compute several spatial properties of the scene that actually convert the original sensory data into symbolic beliefs. This includes relative locations of objects, visibility state, gestures like pointing, etc. Assuming that other agents are as well represented in the model, the same computations are applied to analyze the scene from each agents' point of view (*i.e.* from their *perspectives*). This approach is presented in depth in [3].

2) Knowledge representation and management: the robot is endowed with an active knowledge base that provides a logically sound symbolic model of its beliefs on the world, as well as models for each cognitive agent the robot interacts with. Each of these models is independent and logically consistent. This enable reasoning on different perspectives of the world that would be considered otherwise inconsistent (for instance, an object can be visible for the robot but not for the human. This object can have at the same time the property `isVisible` **`true`** and `isVisible` **`false`**, in two different models). Our platform also features continuous storage, querying and event triggering over the pool of facts known by the robot. It relies on OWL ontologies (a decidable subset of the predicate logics). The knowledge base is presented in [4].

Used in combination with the situation assessment framework, the robot is thus able to maintain different models of the world, one per agent. This proves an essential feature ([5], [6]) to enable perspective-aware grounding of natural language, as we will see in next sections.

3) Dialogue input processing, including natural language parsing capabilities, disambiguation routines and interactive concept anchoring. We focused our efforts on three classes of utterance, commonly found in human-robot interaction: *statements* (*i.e.* new facts the human wants to inform the robot), *orders* (or more generically *desires*) and *questions on declarative knowledge* (whose answers do not require explicit planning). This would roughly cover the *representative* (sometimes referred as *assertives*) and *directives* type of illocutionary acts, in Searle [7] classification. This paper focuses on this last facet (dialogue processing).

C. Related work

Processing natural language in situated context is already an established research field. In [5], Roy summarizes what he sees as the main challenges to be tackled: cross-modal representation systems, association of words with perceptual and action categories, modeling of context, figuring out the right granularity of models, integrating temporal modeling and planning, the ability to match past (learned) experiences with the current interaction and the ability to take into account the human perspective.

Kruijff et al. provides in [6] an up-to-date survey of literature on situated human-robot dialogue, focusing on formal representation systems, bi-directionality of the interaction and context building. They point as well that, compared to the cognitive psychology community, the "situated AI" community started only recently to take into account agents focus, perspective and temporal projection abilities.

Dialogue processing on real robots have been explored by several teams. Scheutz [8] has contributions regarding natural language processing in an incremental way, and how this enables instant back-channel feedback (like nodding).

Hüwel et al. [9] propose the concept of *Situated Semantic Unit*: these meaning atoms are extracted from sentences and expose semantic links to other units. The parser tries to satisfy these links and rate accordingly the semantic interpretation of the sentence. Used in conjunction with ontologies, their approach offers good robustness to ungrammatical or partial utterances. They validated the approach with an extensive user-study.

While mostly implemented on virtual agents, the GLAIR cognitive architecture by Shapiro et al. [10] is an architecture explicitly built to tackle the grounding issue from the percept to the decision. The knowledge layer relies on a custom knowledge representation language, it has natural language processing capabilities similar to ours. It features explicit management of contexts of facts and memory models (long term/short term, episodic/semantic).

Also worth mentioning, Mavridis and Roy [11] propose the idea of a *grounded situation model* which is an amodal model of the world where different sensing modalities, including verbal ones (the robot is able to *imagine* objects), are merged. Their framework also allows management of the interaction history (the human can ask for a past event). They propose an implementation in an environment built on simple entities (a manipulator arm and color balls).

D. Contribution

Compared to previous contributions, our efforts have two foci: *(1)* integration between language processing and perception of the environment and the humans, from several perspectives; and *(2)* realistic human-robot interactions: realtime processing; open speech; complex, dynamic, partially unknown human environments; fully embodied autonomous robots with manipulation abilities.

We do not claim any contribution to the field of computational linguists (see [6] for a survey of formal approaches to natural language processing in the robotics field): our main contribution here is the grounding of concepts involved in the human discourse through the robot's own knowledge.

Section II presents the overall grounding process, section III proposes an analysis of the processing of three prototypical sentences. Experimental results are presented in section IV. A discussion regarding the current limitations of our system concludes this article.

II. The Natural Language Grounding Process

Verbal interaction with human presents two categories of challenges: syntactic ones, and semantic ones. The robot must be able to process and analyze the structure of human utterances, *i.e.* natural language sentences, and then make sense of them. As stated in the introduction, we process three categories of sentences: *statements*, *desires* and *questions* that can be answered from the declarative knowledge present in the robot knowledge base (a choice similar to the *Behaviour Cycle* in the GLAIR architecture [10]). The grounding of the human discourse consists for us either in extracting the *informational* content of the sentence to produce statements or its *intentional* content (*i.e.* , performative value) to collect orders and questions.

We have developed a dedicated module called DIALOGS[1] that processes human input in natural language, grounds the concepts in the robot's knowledge and eventually translates the discourse in a set of declarative OWL/RDF statements.

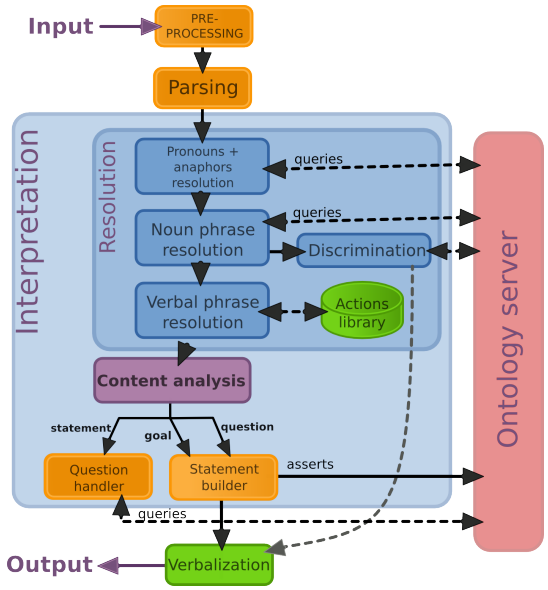

Fig. 2. The DIALOGS module has three main steps: the parsing, the interpretation and the verbalization. The interpretation module is responsible for both the *resolution* and the semantic content *analysis and translation*.

As shown in Figure 2, the DIALOGS module is composed of three main blocks. The user's input is first pre-processed. For instance, *I'm* constructs are expanded into *I am* and then parsed. The parser is a custom-made, rule-based (*i.e.* grammar-free) tool that extracts the grammatical structure from the user's sentence.

The result of the parsing is then sent to the *interpretation* module, the core of the approach. Interpretation consists in three distinct operations: the sentence *resolution* (concepts grounding), the *content analysis* (what is the intent of the utterance: information, question or desire) and the *statement building* (translation into RDF statements).

[1]DIALOGS is an open-source project. Source code is available from http://dialogs.openrobots.org.

The sentence resolution has three steps: *(1)* pronouns and anaphora are replaced by, respectively, the correct speaker ID and the ID of the last object spoken about (extracted from the dialogue history), *(2)* nominal groups are disambiguated and grounded (noun phrase resolution), and *(3)* verbal groups are resolved as well, and their associated thematic roles are retrieved (verbal phrase resolution). Algorithm II.1 describes the overall process. Next section describes specific examples to show how the noun and verbal phrase resolution takes place.

Algorithm II.1: RESOLUTION($sentence, currentSpeaker$)

$\mathcal{G} \leftarrow$ PARSENOMINALGROUPS($sentence$)
for each $g \in \mathcal{G}$
do
- $\mathcal{D} \leftarrow$ GENERATEDESCRIPTION(g) (1)
- $candidates \leftarrow$ ONTOLOGY.FIND($\mathcal{D}$) (2)
- **if** $|candidates| = 0$
  - **then** **output** (Couldn't resolve the group!); **exit**
  - **else if** $|candidates| = 1$
  - **then** $id \leftarrow candidates[0]$ (3)
  - **else** **if** ONTOLOGY.CHECKEQUIVALENT($candidates$) **then** $id \leftarrow candidates[0]$; **else** $id \leftarrow$ DISCRIMINATION($candidates$)
- REPLACE($g, id, sentence$)

As represented on Figure 2, interpretation tightly relies on the communication with the knowledge base. All the concepts the robot manipulates are stored in the *ontology server* and retrieved through logical queries, except for the verbs that are currently stored in a dedicated library (the *action library* on the diagram).

III. Technical Analysis

In order to better understand the overall process we next describe the different steps based on three examples.

In this example, we assume some initial facts are present in the knowledge base, both in the robot's *own* model and in the *human's* model. Since the robot tries to ground a human utterance, all queries are sent to the *human* model, *i.e.* from the human perspective.

A. Informational content extraction

Figure 3 shows a first example of human discourse grounding and the extraction of informational content. We suppose that the robot knowledge base only contains two initial statements in the human model. The user asserts a new one: "The yellow banana is big!".

We need to resolve the nominal group *The yellow banana* to a known concept. A set of partial statements that describe the concept is generated based on the grammatical parsing of the sentence (algorithm II.1, *(1)*). In the example, a banana (`?obj` **`type`** `Banana`) that is yellow (`?obj` **`hasColor`** `yellow`)[2]. Based

[2]Predicates like `hasColor` or `hasSize` that bind `banana_01` to adjectives are extracted from a predefined database of [$Predicate \rightarrow AdjectiveCategory$], and falls back on the generic `hasFeature` predicate if the adjective is not known.

Initial knowledge model of human_01

```
banana_01 type Banana
banana_01 hasColor yellow
```

Human input

"The yellow banana is big!"

Generated partial statements

```
?obj type Banana
?obj hasColor yellow
  ⇒ ?obj = banana_01
```

Newly created statements

```
banana_01 hasSize big
```

Fig. 3. First example of natural language grounding: the nominal group "the yellow banana" is matched with the individual banana_01

Initial knowledge model of human_01

```
banana_01 type Banana
banana_01 hasColor yellow
```

Human input

"Give me the banana."

Generated partial statements

```
?obj type Banana
  ⇒ ?obj = banana_01
```

Newly created statements

```
human_01 desires situation_a3f74
situation_a3f74 type Give
situation_a3f74 performedBy myself
situation_a3f74 actsOnObject banana_02
situation_a3f74 receivedBy human_01
```

Fig. 4. Second example: processing an order.

on these partial statements a query is sent to the ontology server to retrieve possible instances that match the description (algorithm II.1, *(2)*).

In this first simple case, the concept banana_01 is unambiguously matched (since there is only one possible banana) and returned. We can then add the new information provided by the human, *i.e.* the new statement banana_01 **hasSize** big, to the human model in the ontology server.

B. Intentional content through verb resolution

The sentence in the first example is built with the state verb *be* at indicative. Let us examine a different example with an action verb at imperative mode (*i.e.* an order): "Give me the banana". The process is described in Figure 4.

In order to capture the intentional content of a sentence (for example, an order) we need to retain the semantics of the verb and its complements. *Thematic roles* allow to semantically link a verb to its complements. We use a small set of them that matches the relations the robot can actually achieve. In this second example, the verb *give* has three thematic roles: performedBy, actsOnObject and receivedBy.

Initial knowledge model of human_01

```
banana_01 type Banana
banana_01 hasColor yellow
banana_02 type Banana
banana_02 hasColor green
```

Human input

"The banana is good."

Generated partial statements

```
?obj type Banana
```

Robot output speech

"The yellow one or the green one?"

Human answer

"The green one."

Newly created statements

```
banana_02 hasFeature good
```

Fig. 5. Ambiguity resolution: in this example, "banana" can refer to the yellow banana (banana_01) or the green one (banana_02). Discrimination routines handle the disambiguation process.

The list of actions the robot can plan for (currently *take*, *place*, *give*, *show*, *hide* and *move*) along with possible synonyms and their associated thematic roles are stored in a predefined library of actions (Figure 2). For each action we identify and store the role of the subject of the sentence — always performedBy; the role of the direct object (for instance, actsOnObject); and the role of each of the indirect objects with their optional prepositions (for instance, receivedBy)[3]. Moreover, we check with the help of the ontology that each holder of a role has a consistent semantic. For instance, action *Give* must have a manipulable physical item (*Artifact*) as direct object. Thus, if the concept the robot finds for the thematic role *actsOnObject* can not be inferred to be an artifact, it goes back to the human saying it does not understand.

Once the sentence is completely resolved and translated into a formal representation (a human desire in this case[4]), we store it in the ontology server. The robot's decisional/executive layers should then decide whether to execute the order or not.

C. Informational content extraction requiring clarification

This last example (Figure 5) shows the resolution of ambiguous concepts. In this case the user refers to "the banana" while two instances of the Banana class exist in the ontology. The robot needs to find out to which instance the user is actually referring to. To this end, disambiguation routines [12] find differences between the instances (in the example, one banana is yellow while the other one is green) and build a sentence through the *verbalization* module to ask the user a closed question that will help clarify the ambiguity: "Is it yellow or green?" The user's answer is parsed and added

[3]Note that in example 2, "give me the banana", the pronoun "me" appears before "banana", while it is an indirect complement — "give it **to me**". The parser correctly handles these cases.

[4]Orders are here represented as human desires: the human desires a specific new situation.

to the previous sentence. The resulting, augmented, sentence (*i.e.* "Give me the green banana") goes again through all the interpretation steps. This process is repeated until no ambiguities arise. In the example, the `banana_02` is finally returned.

Several other strategies are used in parallel to disambiguate concepts without having to ask for more information to the human:

- Which objects are currently visible to the human? If only one of them, then it is probably the one the user is talking about.
- Did a previous interaction involved a specific object that would still be the subject of the current sentence?
- Is the user looking or pointing to a specific object?

While no examples involving questions have been detailled, *W-* questions and *yes/no* questions can be processed in a similar way by DIALOGS. For instance, a question like: *What is on the table?* is grounded (to extract the relation *isOn* and to find what *table* refers to) and transformed into the following kind of query: `find ?var [?var isOn table1]`. Answers are converted back to a full sentence and uttered to the human.

IV. EXPERIMENTAL RESULTS

In order to illustrate the approach presented in this paper, we have designed the following daily life situation. Tom and Jerry are moving to London, so they are packing things in boxes. The scenario takes places in the living-room, where Jido (our robot) is observing while they move things here and there. To assess the reasoning abilities of the robot they ask Jido for information (entered through keyboard). Ideally, the robot should also perform actions when required (e.g. hand an object when asking "give me..."). However, since it is out of the scope of this work, we do not include any motion from the robot's side.

Perception of objects is done through a tag-based system and humans are detected through motion capture. The robot knowledge base is pre-loaded with the *ORO Commonsense Ontology*[5]. We next describe in detail two situations where we can follow the internal robot's reasoning and the interaction with the users.

1) Implicit disambiguation through visual perspective taking: Tom enters the room while carrying a big box (Figure 1, page 1). He approaches the table and asks Jido to handle him the video tape: "Jido, can you give me the video tape". The DIALOGS module queries the ontology to identify the object the human is referring to: `?obj` **`type`** `VideoTape`.

There are two video tapes in the scene: one on the table, and another one inside the cardboard box. Thus, the knowledge base returns both: ⇒ `?obj = [videoTape1, videoTape2]`.

However, only one is visible for Tom (the one on the table). Thus, although there is an ambiguity from the robot's perspective (since it can see both video tapes), based on the perspective of its human partner it infers that Tom is referring to the video tape on the table, and not the one inside the box which is not visible from his view. Therefore, non-visible objects are removed obtaining: `?obj =[videoTape1]`.

[5]This ontology can be downloaded from http://oro.openrobots.org/.

Robot's beliefs about itself (*robot's model*):
`videoTape1` **`type`** `VideoTape`
`videoTape1` **`isOn`** `table`
`videoTape1` **`isVisible`** *`true`*
`videoTape2` **`type`** `VideoTape`
`videoTape2` **`isIn`** `cardBoardBox`
`videoTape2` **`isVisible`** *`true`*
Robot's beliefs about Tom (*Tom's model*):
`videoTape1` **`type`** `VideoTape`
`videoTape1` **`isOn`** `table`
`videoTape1` **`isVisible`** *`true`*
`videoTape2` **`type`** `VideoTape`
`videoTape2` **`isIn`** `cardBoardBox`
`videoTape2` **`isVisible`** *`false`*

TABLE I
ROBOT'S BELIEFS ABOUT ITSELF AND ITS HUMAN PARTNER.

Since only one object is available, the robot infers that the human refers to it and would eventually execute the command, *i.e.* give it to the human. Alternatively, the robot could first verify with the human if that was the object being referred to or not before proceeding to execute the action. Table I lists the robot's beliefs about itself and its human partner involved in this situation.

Fig. 6. Jerry asks Jido for the content of the box by pointing at it.

2) Explicit disambiguation through verbal interaction and gestures: In this situation, Jerry enters the living room without knowing where Tom had placed the video tapes. So he first asks Jido: "What's in the box?". Before the robot can answer the question it has to figure out which box Jerry is talking about. Similar to the previous situation, there are two available boxes:

```
?obj type box
    ⇒ ?obj = [cardBoardBox, toolbox]
```

However both are visible and the cognitive ambiguity resolution cannot be applied. The only option is to ask Jerry which box he is referring to: "Which box, the toolbox or the cardboard box?" Jerry could now simply answer the question. Instead, he decides to point at it while indicating: "This box" (Figure 6). The robot's perception identifies the `cardBoardBox` as being pointed at and looked at by the human and updates the ontology with this new

information using a rule available in the commonsense ontology (**pointsAt**(?ag, ?obj) ∧ **looksAt**(?ag, ?obj) → **focusesOn**(?ag, ?obj)) The DIALOGS module is then able to merge both sources of information, verbal ("this") and gestural to distinguish the box Jerry refers to.

```
Jerry pointsAt carboardBox
Jerry looksAt carboardBox
→ Jerry focusesAt carboardBox
    ⇒ ?obj = [cardBoardBox]
```

Finally, the DIALOGS queries the ontology about the content of the box and the question can be answered: "Jido-E". Note that the object's label is used instead of its ID. This way we enhance interaction using familiar names given by the users.

```
?obj isIn cardBoardBox
    ⇒ ?obj = videoTape2
```

At this point Jerry wants to know where the other tape is, and that is exactly what he asks Jido: "And where is the other tape?". In this occasion, the DIALOGS module is able to interpret that Jerry is not referring to the video which they were just talking about, but to the other one:

```
?obj type VideoTape
?obj differentFrom videoTape2
    ⇒ ?obj = [videoTape1]
```

Since there is only one possible "other" video (there are only two videos in the scene), it can directly answer Jerry: "The other tape is on the table and next to the toolbox."

```
videoTape1 isOn table
videoTape1 isNextTo toolbox
```

V. FUTURE DEVELOPMENTS

While perspective-awareness proves to play an important role in dialogue grounding, several improvements need to be considered.

For instance, we would like to add temporal reasoning abilities: currently, all the interaction takes place in a model that only stores the current state of the world, with basic extensions like management of dialogue history.

The current framework also lacks a proper management of uncertainty which is essential for real world environments. A probabilistic layer could be added by attaching truth probabilities to statements, similar to [13].

Non-verbal communication (so-called "back-channel communication": nodding, social gaze, cues based on small movements, ...) should be also largely extended, both as new input percepts and as new communication behaviours [8].

We plan eventually to conduct a user study with non-expert humans to validate our hypotheses regarding the importance of perspective awareness for the natural language grounding.

VI. CONCLUSION

In this paper we have presented the DIALOGS module that converts natural language utterances into either symbolic facts (OWL statements) or natural language answers, depending on the intent conveyed by the original sentence.

Grounding of referent is done by relying on a symbolic knowledge base that is able to store several *perspectives* on the world state, one for each agent. Our system takes also into account non-verbal communication cues, like gaze or pointing gestures.

Using so-called thematic roles and the symbolic reasoning capabilities of the knowledge base, semantic correctness of utterances can be checked by the robot who can react accordingly.

We have demonstrated this module in an experiment involving a service robot, in an everyday environment. This experiment showed how ambiguous referents are successfully resolved by the robot, using multi-modal communication cues.

VII. ACKNOWLEDGMENTS

Part of this work has been conducted within the EU CHRIS project (`http://www.chrisfp7.eu/`) funded by the E.C. Division FP7-IST under Contract 215805. Part of this work has been supported by a Marie Curie Intra European Fellowship. We would like to thank as well Patrick Tsemengue and Madhi Chouayakh for their great work on DIALOGS.

REFERENCES

[1] J. Austin, J. Urmson, and M. Sbisà, *How to do things with words*. Harvard University Press, 1962.

[2] J. Gruber, *Studies in lexical relations*. PhD thesis, Massachusetts Institute of Technology, 1965.

[3] E. Sisbot, R. Ros, and R. Alami, "Situation assessment for human-robot interaction," in *20th IEEE International Symposium in Robot and Human Interactive Communication*, 2011.

[4] S. Lemaignan, R. Ros, L. Mösenlechner, R. Alami, and M. Beetz, "ORO, a knowledge management platform for cognitive architectures in robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[5] D. Roy and E. Reiter, "Connecting language to the world," *Artificial Intelligence*, 2005.

[6] G. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, I. Kruijff-Korbayová, and N. Hawes, "Situated dialogue processing for human-robot interaction," *Cognitive Systems*, pp. 311–364, 2010.

[7] J. R. Searle, "A classification of illocutionary acts," *Language in society*, vol. 5, no. 01, pp. 1–23, 1976.

[8] T. Brick and M. Scheutz, "Incremental natural language processing for HRI," in *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, 2007.

[9] S. Huwel, B. Wrede, and G. Sagerer, "Robust speech understanding for multi-modal human-robot communication," in *The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 2006.

[10] S. Shapiro and J. Bona, "The GLAIR cognitive architecture," in *Biologically Inspired Cognitive Architectures-II: Papers from the AAAI Fall Symposium, FS-09-01 (AAAI Press, Menlo Park, CA)*, pp. 141–152, 2009.

[11] N. Mavridis and D. Roy, "Grounded situation models for robots: Bridging language, perception, and action," in *AAAI-05 Workshop on Modular Construction of Human-Like Intelligence*, 2005.

[12] R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Which one? grounding the referent based on efficient human-robot interaction," in *19th IEEE International Symposium in Robot and Human Interactive Communication*, 2010.

[13] D. Jain, L. Mösenlechner, and M. Beetz, "Equipping robot control programs with first-order probabilistic reasoning capabilities," in *International Conference on Robotics and Automation (ICRA)*, 2009.