Situated Dialogue and Spatial Organization: What, Where... and Why?

Geert-Jan M. Kruijff[1]; Hendrik Zender[1]; Patric Jensfelt[2] & Henrik I. Christensen[2]
[1]Language Technology Lab, German Research Center for Artificial Intelligence (DFKI GmbH),
Saarbrücken, Germany
[2]Centre for Autonomous Systems, Royal Institute of Technology (KTH),
Stockholm, Sweden
gj@dfki.de

Abstract: *The paper presents an HRI architecture for human-augmented mapping. Through interaction with a human, the robot can augment its autonomously learnt metric map with qualitative information about locations and objects in the environment. The system implements various interaction strategies observed in independently performed Wizard-of-Oz studies. The paper discusses an ontology-based approach to representing and inferring a 2.5-dimensional spatial organization, and to how knowledge of spatial organization can be acquired autonomously or through spoken dialogue interaction. Dialogue processing provides rich semantic representations of the meaning expressed by an utterance, making it possible to combine the conveyed description with inferences over ontologies modeling commonsense knowledge about indoor environments. The resulting conceptual descriptions are then used to represent qualitative knowledge about locations in the environment.*

Keywords: *Human-Robot Interaction, Conceptual Spatial Mapping, Situated Dialogue*

1. Introduction

More and more robots find their way into environments where their primary purpose is to interact with humans to help and solve a variety of service-oriented tasks. Particularly if such a service robot is mobile, it needs to have an understanding of the spatial and functional properties of the environment in which it operates. The problem we address is how a robot can acquire an understanding of the environment so that it can autonomously operate in it, and communicate about it with a human. We present an architecture that provides the robot with this ability through a combination of human-robot interaction and autonomous mapping techniques. The architecture captures various functions that independently performed Wizard-of-Oz studies have observed to be necessary for such a system.

The main issue is how to establish a correspondence between how a human perceives spatial and functional aspects of an environment, and what the robot autonomously learns as a map. Most existing approaches to robot map building, or Simultaneous Localization And Mapping (SLAM), use a metric representation of space. Humans, though, have a more qualitative, topological perspective on spatial organization (McNamara, 1986). We adopt an approach in which we build a multi-level representation of the environment, combining metrical maps and topological graphs (as an abstraction over metrical information), like (Kuipers, 2000). We extend these representations with conceptual descriptions that capture aspects of spatial and functional organization. The robot obtains these descriptions either through interaction with a human, or through inference combining its own observations (*I see a coffee machine*) with ontological knowledge (*Coffee machines are usually found in kitchens, so this is likely to be a kitchen!*). We store objects in the spatial representations, and so associate the functionality of a location with that of the functions of the objects present there.

Following (Topp & Christensen, 2005) and (Topp et al., 2006), we talk about *Human-Augmented Mapping* (HAM) to indicate the active role that human-robot interaction plays in the robot's acquisition of qualitative spatial knowledge. In §2 we discuss various observations that independently performed Wizard-of-Oz studies have made on typical interactions for HAM scenarios, and we indicate which we will be able to handle. In §3 we present our approach to a multi-layered conceptual-spatial represention and the mechanisms it uses to encode knowledge about spatial and functional aspects of the environment. In §4 we describe the natural language processing facilities that enable the robot to conduct a situated dialogue with its human user about their environment. We present the implementation of our approach in an HRI architecture in §5, followed by a discussion of our experiences with the system in §6. The paper closes with conclusions.

2. Observations on HAM

Various Wizard-of-Oz studies have investigated the nature of human-robot interaction in HAM. (Topp et al., 2006) discuss a study on how a human presents a familiar indoor environment to a robot, to teach the robot more about the spatial organization of that environment. (Shi & Tenbrink, 2005) study the different types of dialogues found when a subject interacts with a robot wheelchair (while being seated in it). Below we discuss several crucial insights these studies yield.

The experimental setup in (Topp et al., 2006) models a typical guided tour scenario. The human tutor guides the robot around and names places and objects. One result of the experiment is the observation that tutors employ many different strategies to introduce new locations. Besides naming whole rooms ("this is the kitchen" referring to the room itself) or specific locations in rooms ("this is the kitchen" referring to the cooking area), another frequently used strategy was to name specific locations by the objects found there ("this is the coffee machine"). Any combination of these individual strategies could be found during the experiments. Moreover, it has been found that subjects only name those objects and locations that they find interesting or relevant, thus *personalizing* the representation of the environment that the robot constructs.

In the study presented in (Shi and Tenbrink, 2005), the subjects are seated in a robot wheelchair and asked to guide it around using verbal commands. This setup has a major impact on the data collected. The tutors must use verbal commands containing deictic references in order to steer the robot. Since the perspective of the human tutor is identical to that of the robot, deictic references can be mapped one-to-one to the robot's frame of reference. One interesting finding is that people tend to name areas that are only passed by. This can either happen in a 'virtual tour' when giving route directions or in a 'real guided tour' ("here to the right of me is the door to the room with the mailboxes."). A robust conceptual mapping system must therefore be able to handle information about areas that have not yet been visited.

Next we discuss how we deal with the above findings, combining information from dialogue and commonsense knowledge about indoor environments.

3. Spatial Organization

In order for a robot to be able to understand and communicate about spatial organization, we must close the gap between the different ways humans and robots think of spatial organization. We discuss here our approach to representing the spatial and functional aspects of an environment at multiple levels of abstraction, thus closing this gap. *Spatial aspects* cover the organization of an environment in terms of connected areas and gateways that together constitute a conception of reachable space.

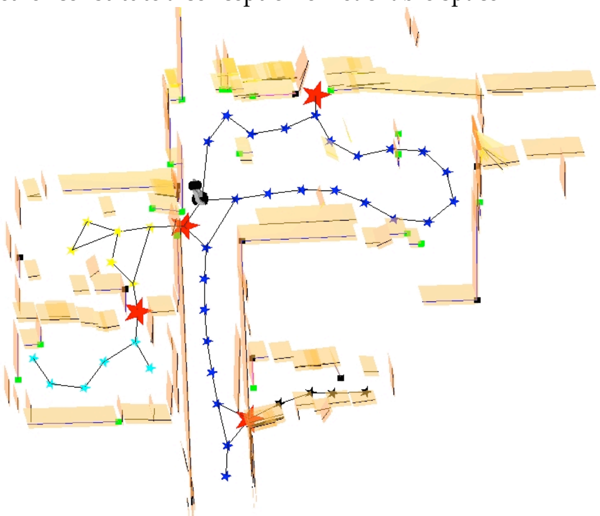


Fig. 1. Automatically acquired metric map

We associate *functional aspects* with an area on the basis of objects present in it. Through dialogue, we can build, query, and clarify these representations, and we point out how they are used in carrying out tasks. A core characteristic of our approach is that we analyze each utterance to obtain a representation of the meaning it expresses, and how it (syntactically) conveys that meaning – rather than just doing for example keyword spotting. This way, we can properly handle the variety of ways in which people may express assertions, questions, and commands. Furthermore, now that we have a representation of the meaning of the utterance we can combine it with further inferences over ontologies to obtain a complete conceptual description of the location or object being talked about.

3.1 Representing the environment

The spatial organisation of an (indoor) environment is represented at three levels (Fig. 2).

At the lowest level, we have a *metric map* (Fig. 1)., capturing observed spatial structures in the environment with a feature-based representation and establishing a notion of free and reachable space through a navigation graph.

The example of Fig. 2 shows line features, which typically correspond to walls. Each map primitive (line features and navigation nodes) is parameterized in world coordinates. A line is for example defined by a start- and an end-point. The metric map is automatically generated from sensor data as the robot moves around the environment. Using features has several advantages. Firstly, they give a compact representation, which, secondly, allows for efficient updates. Among the disadvantages, we find that the map does not explicitly model the free space of the environment as for example an occupancy-grid model would. Only structures that fit the model primitives (e.g. lines) will be captured. We therefore represent the free space and the connectivity of it by a *navigation graph*.

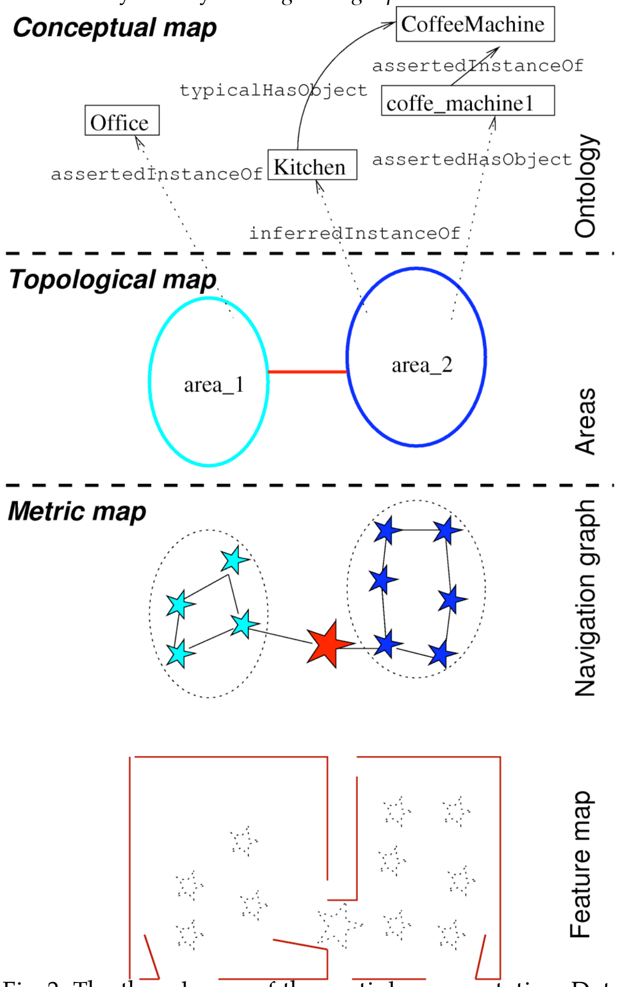


Fig. 2: The three layers of the spatial representation. Dotted lines denote connections between layers.

When the robot moves around, it adds nodes to the graph at the robot's current position if there is not already a node close by in the graph. This approach is inspired by the notion of 'free space markers', cf. (Newman et al., 2002). Each node is associated with a coordinate in the reference frame of the metric map and thus states that the area around that position was free from obstacles when it was added to the graph.

Assuming a mostly static environment, this location is likely to be free also when revisiting it. When the robot travels between nodes, edges are added to the graph to connect the corresponding nodes. We distinguish between two types of nodes: normal nodes and gateway nodes. The gateway nodes (large red stars in Fig. 1 and Fig. 2) encode passages between different areas (e.g. rooms) and typically correspond to doors. In the current implementation, the gateways are detected from the laser range data.

As an intermediate level of abstraction, we have *a topological map*, which divides the navigation graph into areas that are delimited by gateway nodes. This map is a first approximation of a humanlike perspective on space.

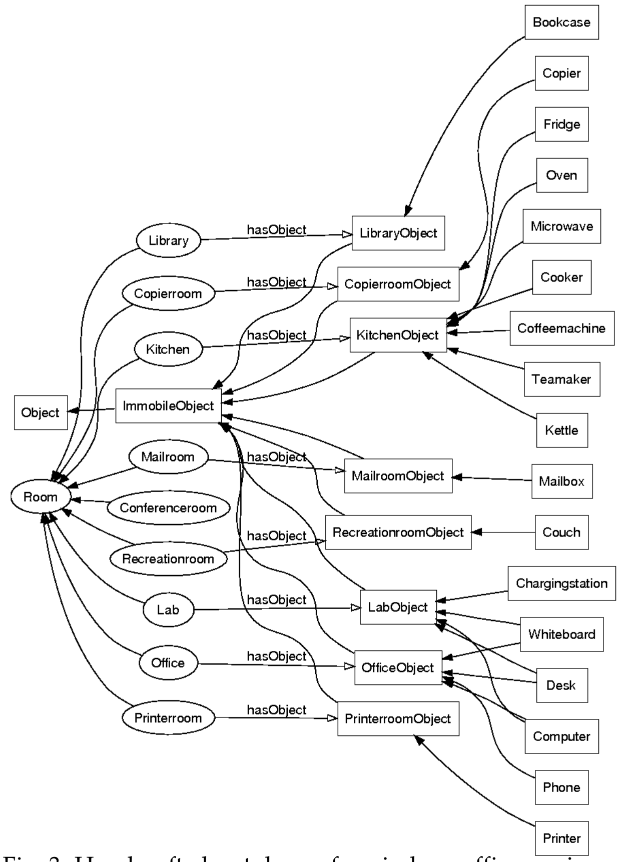


Fig. 3. Handcrafted ontology of an indoor office environment. Solid arrows denote the taxonomical is-a relation.

Finally, we have a *conceptual map* at the top level. In this layer, we store knowledge about names of areas and information about objects present therein in an ontology. Through fusion of acquired information and innate conceptual knowledge (given in a handcrafted commonsense ontology, cf. §3.2) a reasoner can infer new, additional knowledge. This includes inferences over how a room can be verbally referred to, what kinds of objects to expect in a room, and ultimately a functional understanding of what can be done where and why.

Fig. 1 shows a real example of a map that the robot has built. The metric map is represented by the lines, which have been extended to pseudo 3-D walls to indicate that they typically correspond to walls. The navigation graph is shown as the connected set of stars. The larger (red) stars are gateway nodes and, as can be seen, connect different rooms in this case. The grouping of the nodes in the conceptual map is illustrated by colouring the nodes. Each area has its own colour.

3.2 A commonsense ontology of an office environment

Since the robot may have observed only part of an area and the objects therein, and since, as we already pointed out in §2, humans do not necessarily convey complete information about a room, the robot needs to be able to infer knowledge on the basis of only partial information. For this, we use ontological knowledge of spatial and functional aspects.

We have handcrafted a commonsense ontology (Fig. 3) that models conceptual knowledge about an office environment. On the top level of the conceptual taxonomy, there are the two concepts *Object* and *Room*. The subconcepts of *Room* are defined by the instances of *Object* that are found there. This is encoded by the *hasObject* relation: if a given instance of *Room* is related with a specific instance of *Object* by an instance of the *hasObject* relation, this *Room* instance fulfills the conditions for being an instance of the respective specific subconcept of *Room*.

If for instance (Fig. 2) *area_1*, being an asserted instance of *Room* (not shown in the example), is asserted to contain an instance *coffee_machine1* of the concept *Coffeemachine*, it fulfills the conditions for being an (inferred) instance of *Kitchen*.

Based on the knowledge representation in the ontology, we use ontological reasoning to infer general names for rooms, places to look for specific objects, and to resolve linguistically given references to spatial entities (cf. §4.4).

Asserted knowledge about locations and objects is derived from structural descriptions of verbal input of a human tutor originating in the communication subsystem (cf. §4.1), or from automatically recognized objects provided by a visual object recognition subsystem (cf. §5.3).

4. Situated Dialogue

If robots are to enter the everyday lives of ordinary people, human-robot interaction should minimize the reluctance that people might have towards autonomous machines in their environment. Our natural language communiation system accommodates the fact that spoken interaction, *dialogue*, is the most intuitive way for humans to communicate.

(Lansdale & Ormerod, 1994) define dialogue as a "joint process of communication," which "involves sharing of information (data, symbols, context) between two or more parties." In the context of human-robot interaction (HRI), (Fong et al., 2003) claim that "dialogue, regardless of form, is meaningful only if it is grounded, i.e. when the symbols used by each party describe common concepts." In the previous section, we have presented our approach to establishing a common conceptual ground for a human-robot shared environment. In this section, we will present the linguistic methods used for natural language dialogue with a robot. We will also address the role of dialogue for supervised map acquisition and task execution.

4.1 Deriving the meaning of an utterance

On the basis of a string-based representation that is generated from spoken input through a speech recognition software, a Combinatory Categorial Grammar (CCG) (Steedman & Baldridge, 2003) parser analyzes the utterance syntactically and derives a semantic representation in the form of a Hybrid Logics Dependency Semantics (HLDS) logical form, (Kruijff, 2001) and (Baldridge & Kruijff, 2002). HLDS offers a dependency-based, compositional representation of different sorts of semantic meaning: *propositional content* and *intention*. Complex logical forms can be further differentiated by the ontological sort of their intention and their propositional content. We will present this ontology-based meaning mediation in the next paragraph.

The following examples show semantic representations of some utterances that would lead to the situation depicted in Fig. 2.

(1) "We are in the office."
$@_{\{B1:\mathrm{state}\}}$(**be**
& <*Mood*>**indicative**
& <*Restr*>(*W1: person* & **we**)
& <*Scope*>(*I1: region* & **in**
& <*Dir: Anchor*>(*L1: location* & **office**)))

(2) "Follow me!"
$@_{\{F3:\mathrm{action}\}}$(**follow**
& <*Mood*>**imperative**
& <*Actor*>(*R8: hearer* & **robot**)
& <*Patient*>(*I2: speaker* & **I**))

(3) "This is a coffeemachine."
$@_{\{B6:\mathrm{state}\}}$(**be**
& <*Mood*>**indicative**
& <*Restr*>(*T1: thing* & **this**
& <*VisualContext*>(*O1: visualobject*
& <*Proximity*>**proximal**))
& <*Scope*>(*C1: thing* & **coffeemachine**))

From these semantic representations, *structural descriptions* of the discourse entities they refer to are constructed. The conceptual map is then updated with the information encoded in those structural descriptions that can be resolved to spatial entities (Ex. 1) or objects in the environment (Ex. 3).

A structural description is an HLDS logical form of a nominal phrase – i.e. a syntactic constituent whose head is a noun or a pronoun – that ascribes properties to a discourse referent. The following examples show the structural descriptions that can be derived from the complex logical forms of Ex. 1 and Ex. 3.

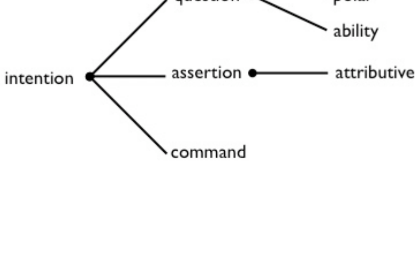
(4) $@_{\{L1:\mathrm{location}\}}$(**office** & <*Number*>**singular**)
(5) $@_{\{C1:\mathrm{thing}\}}$(**coffeemachine** & <*Number*>**singular**)

In dialogue analysis, the linguistic meaning of an utterance is related to the current dialogue context, in terms of how it *rhetorically* and *referentially* relates to preceding utterances. The *rhetorical relation* of an utterance indicates how the utterance extends the current discourse – for example, we try to relate an answer to a question that preceded it, to represent what the answer is an answer to. (This plays an important role in handling e.g. clarification questions.) The *referential* relations of an utterance indicate how *contextual references* like definite noun-phrases ("the box") and anaphora ("it") can be related to objects that have been mentioned in preceding utterances. For example, if a human shows an object to the robot and says "This is a box. It is red" we need to relate "it" with "box." Only this way we can incrementally update the qualitative description for the learnt visual model (Kruijff et al., 2006a). After the utterance is related to the preceding context in this way, an updated model of the dialogue context is obtained in the sense of e.g. (Asher & Lascarides, 2003) and (Bos et al., 2003).

4.2 Mediation of meaning

There are several reasons for why we may want to relate content across different modalities in an HRI architecture. One obvious reason is *symbol grounding*, i.e. the connection of symbolic representations with perceptual or motoric interpretations of a situation, to achieve a situated understanding of higher-level cognitive (symbolic) processes. Achieving such an understanding is an active process. We do not only use the fusion of different content to establish possible connections, but also want it to aid in disambiguating and completing information where and when needed. Finally, relating content may actively trigger processes in a modality (e.g. executing a motor action on the basis of a spoken command) or prime how information is processed (e.g. attentional priming). Altogether this means that we cannot just see content as a symbolic representation without further qualification. We conceive content as a tuple that provides a characterization in terms of *intention*, *propositional content*, and a *truth-value.*

An intention reflects *why* the content is provided to other modalities in the architecture. The intention influences what a connection with content in other modalities is expected to yield.



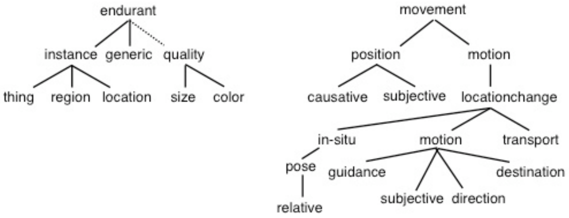


Fig. 4. Two parts of the complex semantic ontology: intention (top) and propositional content (bottom)

Fig. 4 illustrates an ontology of the types of intentions (top figure) we consider here. The ontology is inspired by theories of *speech acts* ((Searle, 1969) and (Core & Allen, 1997)). We discern commands, assertions, and questions. An assertion attributes a characteristic (*assertion.attributive*). A question can inquire after a fact (*question.factual*), whether a particular state is true (*question.polar*), or whether an agent is capable of performing a certain act (*question.ability*).

The intention types in Fig. 4 (top) provide a high-level characterization of the purpose of connecting the content to information in other modalities. When we combine them with the ontological characterization of the propositional content, we get a detailed qualification of what we need to try and relate the content to. Fig. 4 (bottom) gives part of the ontology used for sorting propositional content. It classifies objects (*endurant*) and different types of movement processes (*movement*). Endurants can be physical objects, regions, or locations, and may have qualities such as size or color.

The classification of movement processes is inspired by (Maierborn, 1990) and (Hamp & Feldweg, 1997). Movements are divided into *motion*, and the results of an action as a change of *position*. The latter may concern the agent itself (*subjective*), or regard an object (with the change performed by the agent, *causative*). Motions are movement actions that an agent performs on the spot (*in-situ*), changing pose (*pose.relative*), or that make the agent move to a new position. The latter types of movement can be relative to a person being followed (*guidance*), a subjective frame of reference (*subjective*), a *direction*, or an explicitly given *destination*. The examples below illustrate the types.

(6) several typical commands:
   command. …
(a) "go to the laboratory"
… *movement.motion.locationchange.motion.destination*
(b) "turn to the right"
… *movement.motion.locationchange.motion.direction*
(c) "follow me"
… *movement.motion.locationchange.motion.guidance*

(7) "we are in the office"
   assertion.attributive.endurant.perspective.spatial

We combine the types of the propositional content with the intentions of Fig.4. In the examples above, Ex. 6a shows a command to go to a particular destination: Ex. 6b also gives a command. If we change it into "Can you turn to the right?" we get a question after the ability of the agent to turn into a given direction: *question.ability. …*

Ex. 7 shows the semantic sort of the assertion in Ex. 1.

We create a characterization that includes intention and propositional content so we can determine which modality we need to try and connect this content to. How this connected-to modality should then deal with the provided content is given by the truth-value of the propositional content. The truth-value states how the content can be interpreted against the model of the sensorimotoric or cognitive modality in which the content originates. The interpretation is *dynamic* in that we try to update the model with the propositional content (Muskens et al., 1997). Instead of using a 2-valued truth system, we use a multi-valued system to indicate whether the content was already known in the model (*unknown, known*), and what the result of the update is: *true* if we can update the model, *false* if we cannot, and *ambiguous* if there are multiple ways in which the propositional content can be understood relative to content already present in the model.

To mediate between modalities we represent content using a shared representational formalism, following (Gurevych et al., 2003). We model content as an ontologically richly sorted, relational structure, as described above.

Once we have established the intention, propositional content, and truth value, we can establish *mediation*: We determine to which other modalities we need to establish relations, between the interpretation of the content in the originating modality, and interpretations in those other modalities. Because we determine mediation on the basis of ontological characterizations of content, rather than on its realized form in a modality-specific representation, we speak of *ontology-based content mediation*.

As we already pointed out, mediation can trigger new processes, and result in grounding through information fusion. We keep track of the results of mediation, i.e. the relations between interpretations of content across modalities, by creating *beliefs* that store the handles (identifiers) of the shared representations for the interpretations. We store beliefs at the mediation level. Beliefs thus provide a powerful means for cross-modal information fusion, without requiring individual modalities to commit to more than providing shared representations at the interface to other modalities that enable us to co-index references to interpreted content in individual modalities. For more detail we refer the reader to (Kruijff et al., 2006b).

4.3 Human-Augmented Mapping

In a typical HAM scenario, a human tutor takes the robot on a guided tour of the environment ("follow me!", cf. Ex. 2 and Ex. 6c). Our robotic system is able to follow its tutor, execute *near navigation* commands (e.g. "turn around!", "stop!", cf. (Severinson-Eklundh et al., 2003)), and explore its surroundings autonomously (e.g. "explore the corridor!", "look around the room!"). These individual behaviors can be freely combined and may be initiated or stopped at any point in time by the human tutor. This mixed control strategy – referred to as *sliding autonomy*, (Heger et al., 2005), or *adjustable autonomy*, (Goodrich et al., 2005) – combines the robot's autonomous capabilities where appropriate with different levels of telecontrol through the human user where needed. However, the human tutor preserves full control over the robot, as he can always stop it or give it new commands.

While thus guiding the robot around, he or she then presents and introduces *locations* ("this is the office.", cf. Ex. 1 and Ex. 7) and *objects* ("this is the coffee machine.", cf. Ex. 3). The issue here is how we can use this information to augment the spatial representation.

From language processing, we obtain a representation of the semantics of an utterance (§4.1). Depending on the kind of utterance (e.g. *question*, *command*, *assertion*), we decide in what modalities we need to process this content further (§4.2). A prototypical utterance in a HAM scenario makes an *assertion* about the kind of location the current area is. In this case, we create a structural description (§4.1, Ex. 4 & 5) from the semantics of the utterance, and try to update the conceptual map with the information it contains. The examples below (Ex. 8–10) illustrate a HAM guided tour that would lead to the spatial representation in Fig. 2.

(8) The robot is standing in the office (i.e. area_1). It has no initial map of its environment. The human tutor starts a guided tour by asking the robot to accompany him.
H.1 "Come with me!"
R *initialize people following;*
direct gaze to the tutor;
acknowledge understanding:
R.1 "Yes."
R *start following;*

(9) While the user shows the robot around the room, the robot constructs a metrical map with line features for SLAM and a navigation graph that covers the traveled route. The tutor informs the robot about their location.
H.2 "This is the office."
R *derive structural description:*
$@_{\{L1:\mathrm{location}\}}(\mathbf{office})$
add structural description to conceptual map;
create new instance in ontology:
instance(*area_1*, *Office***)**
acknowledge understanding:
R.2 "Yes."

If the human makes an assertion about an object, we take several steps. First, the vision system learns a model of the object, labeling the model with the structural description for the object (Kruijff et al., 2006a). Next, we anchor the occurrence of the object and its description at the different levels of the spatial representation: in the navigation graph (at the node nearest its position), in the conceptual map (an instance of the object's type is created and related to the individual that represents the current area). By using the same structural description for an object as label for its visual model and as pointer in the spatial representation, we can maintain associations across these representations. Continuing our previous examples, Ex. 10 illustrates this procedure.

(10) The tutor then takes the robot to the next room – a kitchen. The robot detects a doorway, creates a gateway node in the navigation graph, and thus creates a new area in the topological map. The tutor shows the robot the coffee machine. Since the robot has no further knowledge about the area (i.e. area_2), it assumes it to be of the general type *Room*.
H.3 "This is a coffee machine."
R *derive structural description:*
$@_{\{C1:\mathrm{thing}\}}(\mathbf{coffeemachine})$
initiate vision subsystem:
acquire a visual model of the perceived object and store it together with the structural description;
add structural description to conceptual map;
create new instance in ontology:
instance(*coffee_machine1*, *Coffeemachine***)**
instance(*area_2*, *Room***)**
hasObject(*area_2*, *coffee_machine1***)**
acknowledge understanding:
R.3 "Yes."

4.4 Answering questions about locations and objects

Given the robot's conceptual map, we can at any given time ask the robot about where it thinks it is. If a structural description of the current room has been given before, the robot retrieves this information from the conceptual map (Ex. 11). If the robot has not explicitly been given a general name (such as 'kitchen', 'office', or 'lab') for the current area, the system can try to generate a linguistic expression to refer to the given room. This mechnism makes use of the ontological representation of acquired and innate conceptual knowledge to generate a description (Ex. 12). The description of the area is then returned to the dialogue system, which generates a contextually appropriate utterance to convey the given information (Kruijff, 2006).

(11) The robot has the spatial representation acquired in the previous examples. It is standing in the office.
H.1 "Where are you?"
R *retrieve structural description for the current area (area_1) from conceptual map:*
$@_{\{L1:\mathrm{location}\}}(\mathbf{office})$
generate answer with truth value known_true:
R.1 "I am in the office."

(12) Now, the robot is standing in the kitchen, whose general name has not been explicitly given. On the basis of its knowledge about the presence of a coffee machine in the room, the robot can infer that it can linguistically refer to that particular area as "Kitchen".
H.1 "Where are you?"

R *no structural description of the current area in conceptual map;*
query the ontological reasoner:
most-specific-instantiators(*area_2***)**
returns: *Kitchen*
generate structural description:
$@_{\{X0:\mathrm{location}\}}$(**kitchen**)
generate answer with truth value known_true:
R.1 "I am in the kitchen."

If asked about the location of an object, we retrieve occurrences of the desired object from the conceptual map or the ontology. We then generate a structural description of the room where that object can be found, and provide this description to the dialogue system to convey it.

(13)H.2 "Where is the coffee machine?"
R *retrieve object from conceptual map and generate a referring expression for the respective area:*
$@_{\{X0:\mathrm{location}\}}$(**kitchen**)
generate answer with truth value known_true:
R.2 "It is in the kitchen."

Both when asked for rooms or objects, if the system fails to produce a structural description to answer a question, i.e. the information can neither be retrieved from the conceptual map nor inferred through ontological reasoning, the robot generates a negative answer.

(14)H.3 "Where is the laboratory?"
R *information unavailable;*
generate answer with
truth value unknown_false:
R.3 "I am sorry. I do not know."

4.5 Clarification

Existing dialogue-based approaches to HRI usually implement a *master/slave* model of dialogue: the human speaks, the robot listens, e.g. (Bos et al., 2003). However, situations naturally arise in which the robot needs to take the initiative, e.g. to clarify an issue with the human. This is one form of *mixed-initiative* interaction, enabling a robot to recognize when help is needed from a human, and learn from this interaction (Bruemmer & Walton, 2003). A situation that may require is for example when uncertainty arises in automatic area classification: Doors provide important knowledge about spatial organization, but are difficult to recognize robustly and reliably. Clarification dialogues can help to improve the quality of the spatial representation the robot constructs, and to increase the robot's robustness in dealing with uncertain information.

We have extended an approach to processing clarification questions in multi-modal dialogue systems. For space reasons, we refer the reader to (Kruijff et al., 2006c) for technical details. The basic idea is to allow for any modality to raise an *issue*. An issue is essentially a query for information, which is sent into the architecture. Different modalities, e.g. vision or dialogue, can then respond with a statement that they can handle the query. Once an answer to the query is found, it is then returned to the modality that raised the issue.

For example, when mapping is unsure about the presence of a door in a given location, an issue is raised, which is then addressed through interaction with the human. The robot can take the initiative in the dialogue, and phrase a *(clarification) question* about objects (``What is this thing near me?'') or about the truth of a proposition (``Is there a door here?''). Once the dialogue system obtains an answer to the clarification question, both answer and question are provided to the mapping subsystem to resolve the outstanding issue.

4.6 Carrying out tasks

Guiding the robot around an environment is only one step in working with a service robot. The main purpose of a service robot, and of most domestic robots, is to carry out tasks. The multi-level represention of the environment we build up provides an important basis for that. Eventually, we can combine knowledge about what objects are needed to perform particular actions, with the knowledge of where they are.

The simplest action to be performed by a mobile robot is the *go-to task*. The next step in terms of complexity is the *fetch-and-carry task* of locating a specific object or place, going there, possibly fetching the desired object or doing some manipulation with the object in question, and returning.

The current system can be instructed to go to a particular place or object. If the robot knows the location it is sent to, it will just go there. If it has never been shown a place with the respective general name before, it will employ reasoning mechanisms to determine possible locations. If it is sent to an object it has neither been shown nor visually recognized by itself, it will make use of its innate (ontological) knowledge to determine areas that are likely to contain such an object. There, the robot can make use of its autonomous exploration facilities to visually search the area for the desired object.

The current model, however, does not contain functional knowledge about how manipulating and combining objects result in new objects (such as preparing a coffee by placing a cup under a coffee machine and then pressing the start button). As our robot is not equipped with any manipulators (e.g. a gripper or a robotic arm), the physical actions involved in fetching a simple object can only be simulated or replaced by verbally asking for help (Wilske and Kruijff, 2006). The following examples should thus be regarded as proof-of-concept illustrations. (Wilske and Kruijff, 2006) also describe how the system can deal with commands phrased as *indirect speech acts*.

(15)H.1 "I would like to have a tea…"
R *recognize indirect speech act;*
generate a plan to fetch tea
(NB: no planner in the current implementation);
planner queries the conceptual spatial representation for location of a tea kettle;
ontological reasoning yields area_2, the kitchen, as a location likely to contain a tea kettle;
planner succeeds in generating a preliminary plan;
acknowledge understanding:
R.1 "I will get you a tea."
execute generated plan;

5. Implementation

We have implemented the approach of §3 & §4 in a distributed architecture that integrates different sensorimotor and cognitive modalities. The architecture enables a mobile robot to move about in an indoor environment, and have a situated dialogue with a human about various aspects of the environment.

Fig. 5 shows the ActivMedia Pioneer 3 PeopleBot used in the experiments. It is equipped with a SICK LMS291 laser scanner, which is the main navigation sensor and is used for building the metric map, performing obstacle detection, tracking people, etc. On top of the robot, there is a Directed Perception pan-tilt unit with a stereo-vision system from Videre Design on it. Bumpers in the front and back are used to detect contact with the environment. The robot hardware is interfaced using the Player/Stage software. Speech recognition, natural language processing, conceptual spatial reasoning, and people tracking are performed off-board and communicate via wireless Ethernet with the on-board computer.

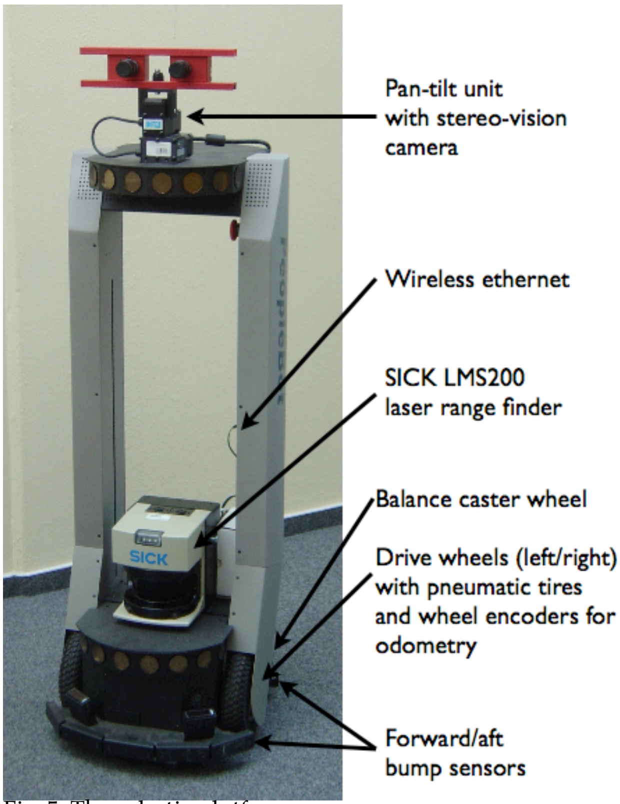


Fig. 5. The robotic platform

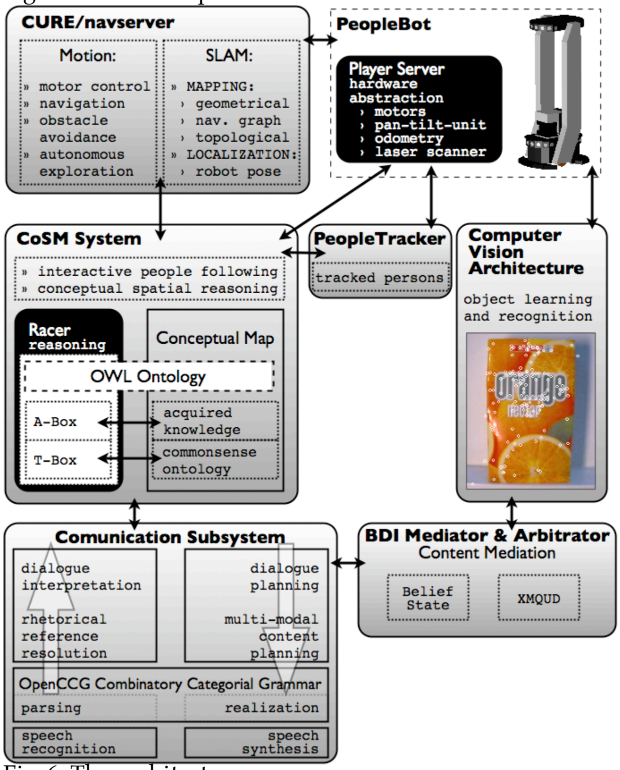


Fig. 6. The architecture

Fig. 6 shows the relevant aspects of the architecture, with subsystems for situated dialogue, spatial localization and mapping, and visual processing. A BDI-mediator (Belief, Desire, Intention) is used to mediate between subsystems. By this we mean that beliefs provide a common ground between different modalities, rather than being a layer on top of these. Beliefs provide a means for cross-modal information fusion, in its minimal form by co-indexing references to information in individual modalities.

The BDI mediator decides what modalities should further process linguistically conveyed information, and how to handle requests for clarifying issues that have arisen. We describe each of these components in more detail below.

5.1 *The communication subsystem*

The communication subsystem consists of several components for the analysis and production of natural utterances in situated dialogue. The purpose of this system is twofold. Firstly, to take an audio-signal as input, recognize what is being said, and then produce a representation of the contextually appropriate meaning of the utterance. As mentioned before, this then enables combining the conveyed meaning with further inferencing over ontological knowledge. Secondly, to take a representation of meaning *to be conveyed* as input, produce a plan of how the robot can communicate that meaning, and carry out that plan.

The communication subsystem has been implemented as a distributed architecture using the Open Agent Architecture (Cheyer & Martin, 2001).

On the analysis side, we use the Nuance speech recognition engine (http://www.nuance.com) with a domain-specific speech grammar. The string-based output of Nuance is then parsed with an OpenCCG parser. OpenCCG (http://openccg.sf.net) uses a combinatory categorial grammar (Baldridge & Kruijff, 2003) to yield a representation of the linguistic meaning for the recognized string/utterance (Baldridge & Kruijff, 2002). These representations are in the same framework used to mediate content between modalities. This enables us to combine linguistically conveyed meaning with further inferences over ontologies

To produce flexible, contextually appropriate interaction, we use several levels of dialogue planning. Based on a need to communicate, arising from the current dialogue flow or from another modality, the dialogue planner establishes a communicative goal. We then plan the content to express this goal, possibly in a multi-modal way using non-verbal (pose, head moves) and verbal means. During planning, we can inquire the models of the situated context (e.g. dialogue context, visually scene) to ensure the plan is contextually appropriate (Kruijff, 2006). The system realizes verbal content using the OpenCCG realizer. OpenCCG takes logical forms representing meaning as input, and then generates a string for the utterance using the same grammar as we use for parsing utterances (White, 2004). Finally, we synthesize the resulting string using the Mary (http://mary.dfki.de) text-to-speech system.

5.2 *Conceptual Spatial Localization & Mapping*

The subsystem for SLAM (Simultaneous Localization And Mapping) creates the metric map uses a lines as map primitives. The underlying feature representation is flexible and other types of features can be used (Folkesson et al., 2005). The basis for integrating the feature observations is the extended Kalman filter (EKF).

In the current implementation, the robot adds a new node to the navigation graph when it has moved 1m assuming that there is no old node close by. It builds the conceptual map automatically from the navigation graph by labeling the nodes into different areas and thus partitioning it. Our strategy rests on the simple observation that the robot passes a door to move between rooms. Whenever the robot passes a door a node marked as a door is added to the navigation graph and consecutive nodes are given a new area label. Currently, door detection is simply based on detecting when the robot passes through a narrow opening. The fact that the robot has to *pass* through an opening removes many false doors that would result from simply looking for narrow openings everywhere. However, this alone will still lead to some false doors in cluttered rooms. A *loop closing* algorithm is used to spot inconsistencies (Kruijff et al., 2006c) arising from falsely recognized doors, and then trigger a clarification dialogue (§4.5).

5.3 Vision

The vision subsystem provides visual scene understanding based on three cues: identity, color, and size of objects in the scene. We use an implementation of SIFT (Scale Invariant Feature Transform) features (Lowe, 2004) and visual codebooks (Fritz et al., 2005) to recognize object identity, and bounding boxes to establish size and color. The subsystem maintains a qualitative interpretation of the spatial organization of objects in the scene, based on topological and projective spatial relations (Kelleher & Kruijff, 2005).

5.4 Ontological Reasoning

The ontological representation is part of the conceptual map. Ontological reasoning is used to fuse knowledge about types and instances of types in the world. We have built a common-sense ontology of an indoor (office) environment (Fig. 3) as an *OWL ontology*, having *concepts*, *instances* (individuals belonging to concepts) and *relations* (binary relations between individuals). The ontology covers types of locations and typical objects. A priori, as the robot has not yet learnt anything, the ontology does not contain any instances. The robot creates instances as the it discovers its environment (§4.3). For each new area, a new instance of concept *Room* is created. When the robot is in a room, and is shown or visually detects an object, we create a new instance of the corresponding *Object* subcconcept, and relate the object's instance and the room's instance using the *hasObject* relation.

We use RACER (http://www.racer-systems.com) to reason over *TBoxes* (terminological knowledge / concepts in our ontology) and *ABoxes* (assertional knowledge / instances). We use *assertions* about instances and relations to represent knowledge that the robot learns as it discovers the world. This includes explicit introductions by the tutor or autonomously acquired information. We do not change the TBox at runtime.

If the conceptual map does not contain a structural description that is relevant for the current task (cf. §4.4 and §4.6), we try to infer the missing information. We use ABox retrieval functions as a first reasoning attempt. The reasoner checks if it can infer that an instance is consistent with the given description. If so, this instance is taken. Else, we use TBox reasoning as a second attempt to resolve uncertainties, e.g. when the robot has not been shown explicitly the occurence of a relevant object. The robot can thus make use of its a priori knowledge about typical occurences of objects and use this as a basis for autonomous planning.

Fig. 2 has already briefly sketched how partial information can be fused.

5.5 Interactive people following

In order to follow the tutor, we use a laser range based people tracking software (Schulz et al., 2003) that uses a Bayesian filtering algorithm. The people tracker derives robust tracking information of dynamic objects within the robot's perceptual range. Given the tracking data, the people following module calculates appropriate motion commands that are sent to the robot control system to follow the tutor's trajectory, while preserving a socially appropriate distance to the tutor when standing still. The system is *interactive* in that it actively gives the tutor feedback about its state. A pan-tilt-unit with a stereo vision device is moved to always point to the tutor, thus giving a gaze-feedback such that he or she is aware that the robot is actually following the tutor and nobody else. Also, should the people tracker lose track of the tutor, the robot provides simple verbal grounding feedback (i.e. ``Oops!'') to quickly inform the tutor. This gives the tutor the possibility to immediately react and wait for the robot to recover. Once the person is found again, which typically takes about a second, another grounding feedback (i.e. ``Ah!'') is given to the tutor who can then proceed. The visual grounding feedback provided by the gaze helps detecting false recovery attempts quickly.

6. Experience

Our main experience with the implemented system is that there are a couple of principal behaviors needed for HAM. If we want a human to guide a robot around an environment, then the robot must be able to (a) follow the human, (b) use information it gets from the human to augment its map, (c) take the initiative to ask the human for clarification; and (d) we need to be able to verify, and correct, what the robot has (not) understood. Where is the system successful, and where is it not? Videos illustrating sample runs with our system are available at http://www.dfki.de/cosy/www/media.

a) Although people tracking/following works fairly smoothly, the robot tends to loose track when the human e.g. passes around a corner. We are now studying how to predict the *path* where a tracked human is going, to overcome this problem and to reduce misclassifications of static objects as dynamic (due to laser data noise). We have also found that having a notion of what human behavior to expect is important: when a human moves to open a door, the robot should not follow the human behind the door, but go through it. The robot needs to reason over functionality of regions/objects in the environment to raise such expectations. We are currently investigating how we can make use of the knowledge that the robot has about its environment to allow for a *smarter* behavior in situations like mentioned before.

b) The question here is not just whether the robot can use information from the human – there is also the issue of how easy or difficult it is for the human to convey that information to the robot in the first place. In our grammar, we have *lexical families* that specify different types of syntactic structures and the meaning they convey, and *lexical entries* specifying how words belong to specific lexical families. This way we can specify many ways in which one can convey the same information (*synonymy*). Dialogue can thus be more flexible, as there is less need for the human to know and give the precise formulation (controlled language).

c) Clarification often concerns aspects of the environment which need to be explicitly referred to, e.g. "Is there a door *here*?" The difficulty lies in generating deictic references with a robot with a limited morphology. Although we can generate spatial referring expressions, non-verbal means would be preferable. However, body- and head-pose may not be distinctive enough. We may thus have to drive to a place (the "HERE") to make the deictic reference explicit, while avoiding disturbing the interaction.

d) Because we have reliable speech recognition (recognition rate is >90%), misunderstanding is primarily a semantic issue. This raises two main questions. First, how does the human understand that the robot understood what was said, without asking the robot? Various systems have the robot repeat what it has just heard. We have not done this; the robot only indicates whether it has understood ("yes"/"okay"/"no"). We have not experienced problems with this, but we are investigating now more explicit nonverbal cues for grounding feedback (e.g. gaze). Second, we need to study what types of misunderstanding may occur in HRI for HAM, and to what extent they may have a *relevant* effect on the robot's behavior. This is an issue we now investigate.

7. Conclusions

We presented an HRI architecture for human-augmented mapping and situated dialogue with a human partner about the environment they share. We discussed the multi-level representations we build of the environment, including spatial organization and functional aspects (based on salient objects present in areas). The system uses autonomous mapping, visual processing, human-robot interaction, and ontological reasoning to construct structural descriptions with which the multi-level representations are annotated. The approach has been fully implemented, and helps bridging the gap between robot and human conceptions of space. We showed its functionality, inspired by independently performed Wizard-of-Oz studies, on several running examples. For future research we want to study more detailed spatial organizations of regions and objects within rooms, to create 3-dimensional representations.

8. Acknowledgements

The research reported of in this article was supported by the EU FP6 IST Cognitive Systems Integrated Project *Cognitive Systems for Cognitive Assistants* "CoSy" FP6-004250-IP.

9. References

Asher, N. & Lascarides, A. (2003). *Logics of Conversation*, Cambridge University Press, New York, NY, USA

Baldridge, J. & Kruijff, G.-J. M. (2002). Coupling CCG and Hybrid Logics Dependency Semantics, Proceedings of the 40th Annual Meeting of theAssociation for Computational Linguistics (ACL 2002), pp. 319—326, Philadelphia, PA, USA, 2002

Baldridge, J. & Kruijff, G.-J. M. (2003). Multi-modal combinatory categorical grammar, Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), Budapest, Hungary, 2003

Bos, J., Klein, E. & Oka, T. (2003). Meaningful conversation with a mobile robot. Proceedings of the Research Note Sessions for the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), pp. 71—74, Budapest, Hungary, 2003

Bruemmer, D. J. & Walton, M. (2003). Collaborative tools for mixed teams of humans and robots, Proceedings of the Workshop on Multi-Robot Systems, Washington, D.C., 2003

Cheyer, A. & Martin, D. (2001). The Open Agent Architecture, *Journal of Autonomous Agents and Multi-Agent Systems*, Vol. 4, No. 1, pp. 143—148

Core, M. & Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme, Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines, Boston, MA, USA, 1997

Folkesson, J., Jensfelt, P. & Christensen, H. I. (2005). Vision SLAM in the measurement subspace, Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005), Barcelona, Spain

Fong, T., Nourbakhsh, I. & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, Vol. 42, pp. 143—166

Fritz, M., Leibe, B., Caputo, B. & Schiele, B. (2005). Integrating representative and discriminant models for object category detection, Proceedings of the International Conference on Computer Vision (ICCV05), Beijing, China, 2005

Goodrich, M. A., Olsen, D. R., Crandall, J. W. & Palmer, T. J. (2001). Experiments in adjustable autonomy, Proceedings of the IJCAI-01 Workshop on Autonomy, Delegation and Control: Interacting with Intelligent Agents, Seattle, WA, USA, 2001

Gurevych, I., Porzel, R., Slinko, E., Pfleger, N., Alexandersson, A. & Merten, S. (2003). Less is more: Using a single knowledge representation in dialogue systems, Proceedings of the HLT-NAACL Workshop on Text Meaning, Edmonton, Alberta, Canada, 2003

Hamp, B. & Feldweg, H. (1997). GermaNet – A lexical-semantic net for German, Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, Spain, 1997

Heger, F. W., Hiatt, L. M., Sellner, B., Simmons, R. & Singh, S. (2005). Results in sliding autonomy for multi-robot spatial assembly, Proceedings of the 8th International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS 2005), Munich, Germany, 2005

Kelleher, J. D. & Kruijff, G.-J. M. (2005). A context-dependent model of proximity in physically situated environments, Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions, Colchester, Essex, UK, 2005

Kruijff, G.-J. M. (2001). *A Categorial-Modal Logical Architecture of Informativiy: Dependency Grammar Logic & Information Structure*. PhD thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Kruijff, G.-J. M. (2006). Dependency grammar. *The Encyclopedia of Language and Linguistics*, 2nd edition, Elsevier Publishers

Kruijff, G.-J. M., Kelleher, J. D., Berginc, G. & Leonardis, A. (2006a). Structural descriptions in human-assisted robot visual learning, Proceedings of the 1st ACM Conference on Human-Robot Interaction (HRI 2006), Salt Lake City, UT, USA, March 2006

Kruijff, G.-J. M., Kelleher, J. D. & Hawes, N. (2006b). Information fusion for visual reference resolution in dynamic situated dialogue, In: *Perception and Interactive Technologies (PIT 2006)*, Andre, E., Dybkjaer, L., Minker, W., Neumann, H. & Weber, M. (Eds.), Springer Verlag, 2006

Kruijff, G.-J. M., Zender, H., Jensfelt, P. & Christensen, H. I. (2006c). Clarification dialogues in human-augmented mapping, Proceedings of the 1st ACM Conference on Human-Robot Interaction (HRI 2006), Salt Lake City, UT, USA, March 2006

Kuipers, B. (2000). The Spatial Semantic Hierarchy. *Artificial Intelligence*, Vol. 119, pp. 191—233

Lansdale, M. & Ormerod, T. (1994). *Understanding Interfaces*. Academic Press, London, UK

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, pp. 91—110

Maienborn, C. (1990). Position und Bewegung: Zur Semantik lokaler Verben, Technical Report, IWBS-Report No. 138, IBM Stuttgart, Stuttgart, Germany

McNamara, T. (1968). Mental representations of spatial relations. *Cognitive Psychology*, Vol. 18, pp. 87—121

Muskens, R., van Benthem, J. & Visser, A. (1997). Dynamics, In: *Handbook of Logic and Language*, van Benthem, J. & ter Meulen, A. (Eds.), Elsevier, 1997

Newman, P., Leonard, J., Tardós, J. & Neira, J. (2002). Explore and return: Experimental validation of real-time concurrent mapping and localization, Proceedings of the 2002 IEEE International Conference on Robotics and Automation (ICRA 2002), pp. 1802—1809, Washington, D.C., USA, 2002

Schulz, D., Burgard, W., Fox, D. & Cremers, A. B. (2003). People tracking with a mobile robot using sample-based joint probabilistic data association filters, *International Journal of Robotics Research*, Vol. 22, No. 2, pp. 99—116,

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, Cambridge, UK

Severinson-Eklundh, K., Green, A. & Hüttenrauch, H. (2003). Social and collaborative aspects of interaction with a service robot, *Robotics and Autonomous Systems*, Vol. 42, pp. 223—234

Shi, H. & Tenbrink, T. (2005). Telling Rolland where to go: HRI dialogues on route navigation, Proceedings of the Workshop on Spatial Language and Dialogue (5th Workshop on Language and Space), Delmenhorst, Germany, 2005

Steedman, M. & Baldridge, J. (2003). Combinatory Categorial Grammar (draft 4.0)

Topp, E. A. & Christensen, H. I. (2005). Tracking for following and passing persons, Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), pp. 70—76, Edmonton, Alberta, Canada, August 2005

Topp, E. A., Hüttenrauch, H., Christensen, H. I. & Severinson-Eklundh, K. (2006). Acquiring a shared environment representation, Proceedings of the 1st ACM Conference on Human-Robot Interaction (HRI 1006), pp. 361—362, Salt Lake City, UT, USA, March 2006

White, M. (2004). Efficient realizations of coordinate structures in combinatory categorial grammar, *Research on Language and Computation*

Wilske, S. & Kruijff, G.-J. M. (2006). Service robots dealing with indirect speech acts, Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006), Beijing, China, October 2006