EU FP7 CogX
ICT-215181
May 1 2008 (52months)

DR 6.1:
Transparency in situated dialogue for interactive learning (in human-robot interaction)

Geert-Jan M. Kruijff, Miroslav Janiček, Ivana Kruijff-Korbayová, Pierre Lison, Raveesh Meena, Hendrik Zender

DFKI GmbH, Saarbrücken
⟨gj@dfki.de⟩

Due date of deliverable:	July 31 2009
Actual submission date:	July 24 2009
Lead partner:	DFKI
Revision:	final
Dissemination level:	PU

A robot can use dialogue to try to learn more about the world. For this to work, the robot and a human need to establish a mutually agreed-upon understanding of what is being talked about, and why. Thereby it is particularly important for the human to understand what the robot is after. The notion of *transparency* tries to capture this. It involves the relation between why a question is asked, how it relates to private and shared beliefs, and how it reveals what the robot does or does not know. For year 1, WP6 investigated means for establishing transparency in situated dialogue for interactive learning. This covered two aspects: how to phrase questions for knowledge gathering and -refinement, and how to verbalize knowledge. Results include methods for verbalizing what the robot does and does not know about referents and aspects of the environment, based on a mixture of prior and autonomously acquired knowledge and basic methods for self-understanding (Task 6.1); and, novel algorithms for determining content and context for question subdialogues to gather more information to help resolve misunderstandings or fill gaps (Task 6.2). WP6 also reports results on making spoken situated dialogue more robust, employing probabilistic models for using multi-modal information to reduce uncertainty in comprehension.

Executive Summary

One of the objectives of CogX is self-extension. This requires the robot to be able to actively gather information it can use to learn about the world. One of the sources of such information is dialogue. But for this to work, the robot needs to be able to establish with a human some form of mutually agreed-upon understanding – they need to reach a *common ground*. The overall goal of WP6 is to develop adaptive mechanisms for situated dialogue processing, to enable a robot to establish such common ground in situated dialogue.

WP6 primarily focuses on situated dialogue for continuous learning. In continuous learning, the robot is ultimately driven by its own curiosity, rather than by extrinsic motivations. The robot builds up its own understanding of the world – its own categorizations and structures, and the ways in which it sees these instantiated in the world. While learning, the robot can solicit help from the human, to clarify, explain, or perform something. This is where situated dialogue can help the robot to self-extend – and which is where transparency comes into play. The robot is acting on its own understanding, which need not be in any way similar to how a human sees the world. There is therefore a need for the robot to make clear what it is after: why the robot is requesting something from a human, what aspects of a common ground it appeals to, and how the request is related to what it does and does not know.

To achieve transparency in situated dialogue for continuous learning, WP6 investigated two important aspects in year 1: Verbalization of knowledge about classes and instances (Task 6.1), and phrasing questions as subdialogues (Task 6.2). WP6 developed novel methods for context- and content-determination in verbalizing knowledge about referents and aspects of the environment, with the possibility to combine a priori and autonomously acquired knowledge. As a result, the robot is capable to refer to instances in a contextually appropriate way, phrase their description relative to knowledge about other instances and classes, and talk about ontological knowledge it has. We base some of these methods in simple ways for the robot to introspect what it knows about an entity (self-understanding), and establish gaps in its understanding of that entity relative to ontological categories. Connected to these efforts, WP6 developed new algorithms for context- and content-determination for question subdialogues, setting such determination against the background of context models of multi-agent beliefs and intentions; and for realizing these dialogues with contextually appropriate intonation. The robot can now plan for how to request information from the user to clarify or extend its understanding. It does so in a manner that appropriately reflects how this request relates to private and shared beliefs, and intentions. In such a mixed-initiative dialogue, the robot can dynamically adapt its plan to achieve its knowledge-gathering intention.

In addition to the main focus on transparency, WP6 continued efforts in making spoken situated dialogue more robust. Further improvements in robustness were achieved by using context information in incremental, distributed processing covering speech recognition, parsing, and dialogue interpretation. This approach explicitly deals with the uncertainty in the speech recogniser in a Bayesian way, which is part of our broader approach in CogX to using probabilistic representations to capture uncertainty in initial interpretations of sensory signals. By maintaining all the possible hypotheses we can then use knowledge from other modalities to revise our interpretation or to bias inference. This is an example of a simple kind of self-understanding, since we are representing the possibilities and uncertainties in our interpretations.

Role of (transparent) situated dialogue in CogX

CogX investigates cognitive systems that self-understand and self-extend. In some of the scenarios explored within CogX such self-extension is done in a mixed-initiative, interactive fashion (e.g. the George and Dora scenarios). The robot interacts with a human, to learn more about the environment. WP6 contributes situated dialogue-based mechanisms to facilitate such interactive learning. Furthermore, WP6 explores several issues around the problems of self-understanding and self-extension in the context of dialogue processing. Dialogue comprehension and production is ultimately based in a belief model the robot builds up. This belief model captures beliefs and tasks, in a multi-agent fashion. We can attribute a belief/task to one agent (private), multiple agents (shared), or have an agent attribute a belief/task to another agent (private, attributed). Already at this level we thus see a range of possible forms of self-understanding and self-extension. The goal of transparency is to establish beliefs as shared, and thus, any belief that should be shared but currently is not represents a gap of sorts. The differentiation between private and shared status status is one aspect of context that helps determine how we produce references to entities in the world, and the way we produce questions about such entities. Furthermore, interpretations leading up to these beliefs and tasks may be uncertain. We use probabilistic models to help counter uncertainty in comprehension, fusing information from multiple modalities to guide comprehension. Should this fail, we can use clarification to overcome that uncertainty. Such clarification can also be used to resolve uncertainty about situated understanding, or in a more general way, to request information about entities in the world. WP6 presents a first attempt at an algorithm for identifying gaps in terms of unknown properties about an entity I relative to a category C. We use these gaps as a basis for verbalizing what a robot does and does not know about I, and to drive dialogue to gain more information about I.

Contribution to the CogX scenarios and prototypes

WP6 contributes directly to the George and Dora scenarios, in relation to work performed in WP 3 (Qualitative spatial cognition), WP 5 (Interactive continuous learning of cross-modal concepts), and WP 7 (Scenario-based integration). Robust dialogue processing, clarification, and verbalization are in principle used in both scenarios. In George we provide the possibility for the robot to ask about visual properties it is uncertain about, and to use verbalization and referencing to describe what it sees:

- **Human** places a red box on the table
- **Robot** Vision recognizes the object as a box, but is unsure about the color. A clarification request is triggered, handled by dialogue.
- **Robot** "Is that box red?" – dialogue provides indirect feedback it has recognized the object as a box, while at the same time asking for confirmation on the color.
- **Human** "Yes, this box is red."
- **Robot** Vision is provided with the information that the box is indeed red, and so can update its models.

In Dora we also explore the introspection on what the robot does and does not know about an area, to drive information requests to the user. (The method is in fact general enough to also drive active visual search in the environment.)

- **Human** guides the robot to a new area, and says "Here we are in the kitchen." This the second kitchen the human and the robot visit.
- **Robot** Place categorization can determine the area as a kitchen, with a particular size. Vision perceives a water cooker.
- **Robot** "Ok, this looks like a larger kitchen." – the robot can compare to other kitchen instances it has seen so far.
- **Robot** The robot can infer that kitchens typically have several objects, not only a water cooker but also a coffee machine. It understands that it does not know of a coffee machine here, though.
- **Robot** "I can see a water cooker. Is there also a coffee machine?" – the robot indicates what it does and does not know, and uses this as the background for extending its knowledge about the area.

1 Tasks, objectives, results

1.1 Planned work

Robots, like humans, do not always know or understand everything. Situated dialogue is a means for a robot to extend or refine its knowledge about the environment. For this to work, the robot needs to be able to establish with a human some form of mutually agreed-upon understanding – they need to reach a *common ground*. The overall goal of WP6 is to develop adaptive mechanisms for situated dialogue processing, to enable a robot to establish such common ground in situated dialogue.

WP6 primarily focuses on situated dialogue for continuous learning. In continuous learning, the robot is ultimately driven by its own curiosity, rather than by extrinsic motivations. The robot builds up its own understanding of the world – its own categorizations and structures, and the ways in which it sees these instantiated in the world. While learning, the robot can solicit help from the human, to clarify, explain, or perform something. This is where transparency comes into play. The robot is acting on its own understanding, which need not be in any way similar to how a human sees the world. There is therefore a need for the robot to make clear what it is after: why the robot is requesting something from a human, what aspects of a common ground it appeals to, and how the request is related to what it does and does not know. To achieve transparency in situated dialogue for continuous learning, WP6 investigated two important aspects in year 1: Verbalization of knowledge about classes and instances (Task 6.1), and phrasing questions as subdialogues (Task 6.2).

Task 6.1: Verbalising categorical knowledge *The goal is to enable the robot to verbalize its own categorical knowledge (or lack thereof) relative to a situation, and understand situated references. We will extend existing methods for comprehending and producing referring expressions to cover verbalization of relevant information from singular visual categories (WP5) and contextual reference.*

Task 6.2: Continual planning for clarification and explanation *We will extend strategies for planning clarification- and explanation dialogues using a continual planning approach. This offers the necessary flexibility to adjust a plan when interactively setting up an appropriate context, and provides a model of common ground in dialogue. These methods will be based in means for grounding the information expressed by clarifications and explanations in situated understanding.*

The intention behind Tasks 6.1 and 6.2 was to achieve that the robot would be able to enter into a dialogue with a human, to clarify something or to request more information. This could be either about dialogue itself,

or regard the situated context being talked about – thus spanning the entire range of Clark's grounding levels [11]. The robot uses belief models to represent private and shared beliefs, including private beliefs the robot attributes to other agents, and ontologies to capture its categorical knowledge about the world. Together, belief models and ontologies provide a rich epistemological background against which the robot can introspect what it does or does not know (e.g. whether another agent does understand something, or whether an observed instance is of a particular category). We use such self-understanding to guide verbalization and clarification, two interrelated functions to help the robot gather more information to self-extend. The role of verbalization in this process is to ensure that the why what and how of the question is clear to the human: why the robot asks, what it does and does not know, and how that gap should be addressed. The planning part is to take care of the planning and execution of the actual dialogue, to ensure human and robot eventually do achieve a common ground. In §1.2 we describe how we achieved these goals.

1.2 Actual work performed

Below we succinctly describe the achievements for the individual tasks. The descriptions refer to the relevant papers and reports in the annexes, for more technical detail. In §1.3 we place these achievements in the context of the state-of-the-art.

1.2.1 Verbalising categorical knowledge

The goal of Task 6.1 was to develop methods for the robot to verbalize its own categorical knowledge, or lack thereof. We have achieved the following:

Context-determination, bi-directionality in referencing A robot typically acts in an environment larger than the immediately perceivable situation. The challenge in referring to objects and places in such a large environment is to ensure that the agents participating in the dialogue can identify the appropriate context against which the resolve a reference. Zender et al (§2.1.1, §2.1.2) have developed novel methods for determining the appropriate context for comprehending and producing referring expressions.

A typical example Zender et al address is when the robot needs to refer to an object in a place other then where the robot currently is, talking to a human. Or when it needs to understand such a reference. For example, the robot has been sent to fetch a person to take a phone call in somebody else's office (e.g. GJ's). If this person is currently in her office, it would not do to say "there's a call for you on the phone." This could incorrectly identify the phone on that person's desk as the one to pick up, whereas the

point is to go to the GJ's office to take the call there. What Zender et al do is to use topological structure of the environment, to –literally– determine the appropriate context for identifying the object it needs to refer to. So, instead of just saying "the phone," the robot is able to say "there is a call for you on the phone on the desk in GJ's office." It uses the context to direct the human's attention to the appropriate location, where it can then identify the intended referent.

Verbalization of acquired properties Typically a robot is provided with an initial ontology, outlining the concepts and their relations considered relevant for understanding the environment in which the robot is to act. Over time, the robot can extend this ontology, for example with instances and properties that hold for these. Zender and Pronobis (§2.1.3) have developed a new method for verbalizing knowledge about autonomously acquired scalar properties for instances and their classes. The distributions of property values across instances, within a class, and across classes help define contextual standards [34, 15] against which the verbalization of scalar properties as comparatives can be determined in a contextually appropriate manner.

A scalar property is, simply, a property with values that are on a scale that makes them comparable. An example of a scalar property is size: A room can be smaller or larger than some other room, or of the same size. Scalars are typical material properties for the kinds of entities we want the robot to talk about. And, they are properties for which the robot can autonomously acquire quantitative models. The problem is, how to then talk about them. We cannot simply verbalize such a property at face value, e.g. as "the room is $14.67m^2$." Humans prefer more qualitative descriptions, like "large" or "smaller." Such qualitative descriptions are called *vague scalar predicates*. Their exact interpretation is left vague – that is to say, their exact interpretation is relative to a particular *contextual standard* which defines the scale along which comparisons are to be made. Zender and Pronobis propose a method to make it possible for the robot to introspect what variation it has perceived for a particular scalar property among instances of a class, or among classes as such. This form of self-understanding enables the robot to talk in a human-like, qualitative fashion about scalar properties, while at the same time (indirectly) indicating to the human what it considers as prototypical values (by comparison).

Verbalization of categorical knowledge Sometimes it is more important for the robot to make clear what it does *not* understand, than to say what it does know about. This helps the human to figure out what the robot might be after. Zender and Kruijff (§2.1.4) discuss a preliminary method for a robot to introspect the knowledge it has

about an entity in the world. The method establishes what the robot does and does not know about that entity relative to one or more categories in a known ontology. The resulting identified "gaps" are those properties for the entity that the robot does not know about, but which it would need to know to establish the entity as an instance of a particular category. Zender and Kruijff subsequently discuss how the robot can then verbalize this self-understanding, in terms of what the category, the instance and its known properties, and the missing properties identified as gaps.

Zender and Kruijff consider a simple, but often occurring form of "gap": namely, when a robot is lacking property information about an object or an area to fully determine whether it is an instance of a particular category. Consider again the example given earlier. A human and a robot enter a new room, which the human indicates is a kitchen. The robot can categorize the place as such, and even sees a water cooker. However, based on the knowledge it has about kitchens, it would also expect a coffee machine to be there. Zender and Kruijff show how the robot can determine such a property of "having a coffee machine" as a gap in its knowledge about this area (as being a kitchen). To convey this self-understanding, Zender and Kruijff discuss how the robot can then verbalize this gap, together with a description of what it does know about the area-as-a-kitchen. "Ok, this looks like a larger kitchen. [...] I can see a water cooker. Is there also a coffee machine?"

The novelty in all these methods is the role context plays in determining how a robot should understand or verbalize a reference, or what it knows about something (be that an instance or a class). Traditional methods focus primarily on *content*-determination, typically assuming a context to be given. Our methods combine content-determination with context-determination. Context-determination can thereby mean both situated context (e.g. references in large-scale space) and epistemological context (e.g. what beliefs a robot has, or attributes to other agents, or what it knows about how to compare across classes). With that we go beyond the original objectives of Task 6.1, which focused only on verbalizing knowledge about visual objects in a current scene.

1.2.2 Clarification

The goal of Task 6.2 was to develop methods so a robot could clarify or expand what it understands about the environment. These methods were to be continual, in the sense that it should be possible to monitor the execution of a plan, and where necessary adapt or expand it. We have achieved this goal in the following ways.

Determining epistemological context in questions Transparency and scaffolding in dialogue for learning depend on epistemological context: how questions appeal to common ground, what private beliefs they are based in – and what answers an interlocutor would like to have. Kruijff & Brenner (§2.2.1) explore methods for determining such appropriate epistemological contexts, considering transparency and scaffolding explicitly as referential qualities of a question. These contexts are then connected in a notion of question nucleus which reflects what is being asked after, in reference to beliefs (aboutness, transparency) and intentions (resolvedness, scaffolding). A question nucleus provides the basis for formulating a question as a dialogue.

A robot should not just go and blurt out a question – this may not lead to the human given the desired answer. A nice example of this is provided by the former CoSy Explorer system [37]. The robot classified every narrow passage it went through ($< 70cm$) as a door. Sometimes it would realize that some previous passage probably wasn't a door, just an artifact of driving around in a cluttered environment. At that point of realization, the robot would just ask "Is there a door here?" Out of the blue, without further indication of where there ought to be a door, a human would typically say "yes" – understanding the robot to mean, whether there would be a door to this room. Which, of course, was not what the robot meant. But what it failed to do was to properly take into account what the human would know (she didn't know that "here" was supposed to refer to that narrow passage), and how to formulate its question accordingly. Kruijff and Brenner look into how the robot could use its multi-agent belief models to determine how to best pose a question. They start by formulating ways for the robot to introspect its beliefs to determine what the human knows about something the robot wants to ask a question about. This determines how to refer the entity under discussion – making it transparent what the robot is talking about. A second step is to use what the robot holds as private knowledge and beliefs about the entity, to properly indicate what it would like to know more about.

Continual comprehension and production of clarification Brenner et al (§2.2.2) consider how a continual approach for planning and executing dialogues can be applied to human-robot interaction, in general. Kruijff & Janiček (§2.2.3) combine these insights with weighted abduction. The approach covers comprehending and producing dialogue and combines intention, attentional state, and multi-agent belief modeling. Kruijff & Janiček focus on clarification dialogues, covering Clark-style grounding from communicative levels to information requests concerning situated understanding.

Robots don't always understand everything. Sometimes they realize that, but sometimes they don't, and attribute some property to an object that is just plain wrong. Kruijff & Janiček try to capture such forms of collaboration between a human and a robot, dealing explicitly with the continual nature of such collaboration – things may go wrong and then need to be corrected. A typical example that they try to capture is the following:

(1) Human places an object on the table

(2) Robot: "That is a brown object."

(3) Human: "It is a red object."

(4) Robot: "Ok. What kind of object is it?"

(5) Human: "Yes."

(6) Robot: "Aha. But what KIND of object is it?"

(7) Human: "It is a box."

Kruijff & Janiček explicitly use the belief models of the robot, for the robot to figure out how it could use beliefs and observations to establish why a human may have said something, and how to best achieve what the robot itself is after (in terms of updating its beliefs). They make it possible for the robot to assert a belief ("this is a brown object") but then having to retract it when being corrected by a human ("it is a red object") and establishing the corrected belief as a shared belief about the scene ("ok"). At the same time, using what it understands to be shared, the robot can make safe assumptions about how it can refer to objects. Attributed beliefs also make it possible for the robot to assume that the human may know an answer to a question. In its reasoning the robot can then assert that the human will provide it with that information ("what kind of object is it?"). With that the robot first of all explicitly represents the gap in its knowledge (what it would like to know). But this also provides a level at which introspection can track the extent to which the gap has actually been resolved. The robot checks the updates it can make to its belief model in response to its question, and can use the "self"-understood failure to do so to persist in trying to get an appropriate answer from the human. Humans are not always fully cooperative, so when the human replies with "yes" (as in "coffee or tea? yes please") she does not provide an answer to the question. (Non-cooperative behavior is a problem usually "assumed away" in approaches to dialogue; Kruijff & Janiček don't, dealing with it in a continual way as argued for in Brenner et al, §2.2.2.) The robot can figure this out (using the approach of Kruijff & Brenner, §2.2.1), repeat the question, to then finally get the desired kind of answer ("it is a box.").

Contextually appropriate intonation for questions Kruijff-Korbayová et al (§2.2.4) develop new methods for determining information struc-

ture and intonation for a range of utterance types, including commands, assertions, and -most importantly- questions. Information structure and its reflection through e.g. intonation can make it clear to the hearer how the utterance relates to the preceding context, and what it focuses on. As with assertions, where intonation can change the dynamic potential of their interpretation, intonation in a question indicates its dynamics in terms of what it is after: what type of answers is expected. Kruijff-Korbayová et al outline experiments to verify these theoretical insights in an empirical way.

There is more to saying something than simply uttering a sequence of words. In English, the intonation of an utterance reflects what it is that someone is talking about, and what she would like to focus on. There is a marked difference between assertions like "this is a RED box" (capitalization indicating stress) versus "this is a red BOX," or questions like "is this a red BOX?" or "is this is a RED box?" Getting this right is crucial for the robot to convey what it is after. Kruijff-Korbayová et al (§2.2.4) describe how the robot can use private and shared beliefs, and what is currently attended to, to help drive how to formulate a contextually appropriate intonation for an utterance. In combination with the previous achievements, this rounds it all off: We can determine what beliefs and gaps play a role in formulating e.g. a question, we can manage a dialogue around that question, we can verbalize its content and references in a contextually appropriate way, and formulate all that with the right intonation.

The novelty in all these methods is thus how they achieve to flexibly combine intention, multi-agent beliefs and attentional state in continual processing of dialogue. Based on existing approaches, these methods explore how the robot can introspect the private and shared beliefs it entertains, situate beliefs and intentions, and then use that as a background against which it can handle and overcome pervasive aspects such as uncertainty, and the typically large-scale spatiotemporal nature of action and interaction.

1.2.3 Robust processing of spoken situated dialogue

The success of dialogue-based human-robot interaction ultimately stands or falls with how well a robot understands what a human says. Unfortunately, spoken dialogue is difficult to understand. Utterances are typically incomplete or contain disfluencies, they may be ungrammatical, or a speaker may correct herself and restart part of an utterance. This requires processing of spoken dialogue to be robust. At the same time, we cannot sacrifice deep understanding for robustness, as is often done. In the end a robot needs to understand what a human said, to be able to act on it. That is the whole point of situated dialogue as we consider it here.

In addition to Tasks 6.1 and 6.2, we have continued our efforts in robust

processing of spoken situated dialogue. These efforts started already in CoSy; the results reported here build up on these previous efforts but have been achieved entirely during year 1 in CogX.

Integration of context in speech recognition and incremental parsing Lison & Kruijff (§2.3.1) present a novel approach in which context information is used in a process combining speech recognition and incremental parsing. The approach considers the entire processing from incoming audio signal to establishing a contextually appropriate interpretation of an utterance. Lison & Kruijff show that substantial improvements on robustness in processing can be achieved (measured against a WoZ corpus) by including context information (e.g. salient objects, actions). This information is used to bias lexical activation probabilities in the language model for speech recognition, and to guide discriminative models for parse ranking applied at the end of an incremental parsing process.

Incremental contextual pruning in parsing Lison & Kruijff (§2.3.2) consider the application of discriminative models during incremental parsing. After each step, context-sensitive discriminative models are applied to rank analyses. Using a beam of width 30, Lison & Kruijff show how parsing time can be reduced by 50% without suffering any significant reduction in performance (measured on a WoZ corpus).

When a human processes visually situated dialogue, she uses what she sees in the scene and how she knows that scene to help her understand what someone else might be saying about that scene. Lison & Kruijff explore how this idea can be used to make spoken dialogue processing in human-robot interaction more robust. When a robot perceives objects in a scene, it uses that information to activate expressions it could associate with such objects. For example, if it sees a ball, it would activate expressions like "round," "pick up," etcetera. These expressions are phrases the robot expects to hear. They help the robot to anticipate what a human is likely to say, when talking about that scene. Lison & Kruijff show that the robot can use this information to deal with the uncertainty inherent to speech recognition. Doing it in a probabilistic way, it is part of the broader approach in CogX to using probabilistic representations to capture uncertainty in initial interpretations of sensory signals. By maintaining all the possible hypotheses we can then use knowledge from other modalities to bias how the audio signal is interpreted in terms of possible word sequences. This is an example of a very simple kind of self-understanding, since we are representing the possibilities and uncertainties in our interpretations. Lison & Kruijff take this even further, by using the same information about the context to then help parsing to discriminate between possible analyses, to end up with a parse that represents the most likely semantic interpretation of the

audio signal in the given context. Using this sort of discrimination-based-on-context during parsing actually helps to reduce the time needed to parse an utterance.

1.3 Relation to the state-of-the-art

Below we briefly discuss how the obtained results relate to the current state-of-the-art. We refer the reader to the annexes for more in-depth discussions.

1.3.1 Verbalisation

Task 6.1 considered the use of methods for comprehending and producing *referring expressions* to cover verbalization of knowledge, and contextual reference. For such expressions to appropriately refer to the intended referent, they need to meet a number of constraints, to help a hearer identify what is being talked about. First, an expression needs to make use of concepts that can be understood by the hearer. This becomes an important consideration when we are dealing with a robot which acquires its own models of the environment and is to talk about the contents of these. Second, the expression needs to contain enough information so that the hearer can distinguish the intended referent from other entities in the world or a belief state, the so-called *potential distractors*. For this it is necessary that the robot takes the differentiation between private and shared beliefs into account, as we already saw earlier. Finally, this needs to be balanced against the third constraint: Inclusion of unnecessary information should be avoided so as not to elicit false implications on the part of the hearer.

Zender & Pronobis (§2.1.3) particularly deal with the first aspect. Given that a robot autonomously acquires knowledge about the world, how can such properties be used to verbalize what the robot knows? Existing work on modeling *scalar properties* considers the use of *contextual standards*, to determine how to realize such properties as "vague" expressions involving gradable adjectives [15, 34]. This research primarily focuses on instances – in a given visual setting. Zender & Pronobis move beyond this, by considering how scalar properties can be modeled as probabilistic distributions over their values – and then use these distributions to construct contextual standards. This makes it possible to consider distributions solely across observed instances (like [15]), and also across instances within a class (considering values to be *prototypical* values within a class), and across classes. Within-class and across-class contextual standards are not considered (nor immediately possible) in [15]. They are, however, necessary to generate contextually appropriate verbalizations using comparatives. For example, consider the average office to have $8m^2$. Talking about two offices, with *office1* measuring $12m^2$ and *office2* $18m^2$, it would be more appropriate to talk about *office1* as "the smaller office," not as "the small office." The rea-

son being that it is still bigger than the average office. These ideas are based on insights in categorization and *prototypicality* originating with Brown [7] and Rosch [53]: some instances in a category are more prototypical of that category than others.

Zender et al (§2.1.1, §2.1.2) focus on the second and third aspect, namely the problem of including the right amount of information that allows the hearer to identify the intended referent. According to the seminal work on *generating referring expressions* (GRE) by Dale and Reiter [14], one needs to distinguish whether the intended referent is already in the hearer's *focus of attention* or not. This focus of attention can consist of a local visual scene (visual context) or a shared workspace (spatial context), but also contains recently mentioned entities (modeled as beliefs in the belief model associated with the dialogue context). If the intended referent is already part of the current context, the GRE task merely consists of singling out the referent among the other members of the context, which act as distractors. In this case the generated referring expression (RE) contains *discriminatory* information, e.g. "the red ball" if several kinds of objects with different colors are in the current context. If, on the other hand, the referent is not in the hearer's focus of attention, an RE needs to contain what Dale and Reiter call *navigational*, or *attention-directing* information. The example they give is "the black power supply in the equipment rack," where "the equipment rack" is supposed to direct the hearers attention to the rack and its contents.

While most existing GRE approaches assume that the intended referent is part of a given scene model, the *context set*, very little research has investigated the nature of references to entities that are not part of the current context. The domain of such systems is usually a small visual scene, e.g. a number of objects, such as cups and tables, located in the same room, other closed-context scenarios, including a human-robot collaborative table-top scenario [14, 31, 35, 33]. What these scenarios have in common is that they focus on a limited part of space, which is immediately and fully observable: *small-scale space*.

In contrast, mobile robots typically act in more complex environments. They operate in *large-scale space*, i.e. space "larger than what can be perceived at once" [39]. At the same time they do need the ability to understand and produce verbal references to things that are beyond the current visual and spatial context. When talking about remote places and things outside the current focus of attention, the task of *extending the context* becomes key.

Paraboni et al. [46] are among the few to address this problem. They present an algorithm for *context determination* in hierarchically ordered domains, e.g. a university campus or a document structure. Their approach is mainly targeted at producing textual references to entities in written documents (e.g. figures and tables in book chapters), and consequently they do

not touch upon the challenges that arise in a physically and perceptually situated dialogue setting. Nonetheless the approach presents a number of contributions towards GRE for situated dialogue in large-scale space. An appropriate context, as a subset of the full domain, is determined through *Ancestral Search*. This search for the intended referent is rooted in the "position of the speaker and the hearer in the domain" (represented as d), a crucial first step towards situatedness. Their approach suffers from the shortcoming that their GRE algorithm treats spatial relationships as one-place attributes. For example a spatial containment relation that holds between a room entity and a building entity ("the library in the Cockroft building") is given as a property of the room entity (BUILDING NAME = COCKROFT), rather than a two-place relation (`in(library,Cockroft)`). Thereby they avoid recursive calls to the GRE algorithm, which would be necessary if the intended referent is related to another entity that needs to be properly referred to. Zender et al argue that this imposes an unnecessary restriction onto the design of the knowledge base. Moreover, it makes it hard to use their context determination algorithm as a sub-routine of any of the many existing GRE algorithms. They show how these shortcomings can be overcome, in an approach that integrates context- and content-determination as separate routines. The approach is furthermore *bi-directional*, meaning it can be used for both producing and comprehending referring expressions.

Zender & Kruijff (§2.1.4) present preliminary research on a method that enables a robot to introspect what it knows and doesn't know about an instance, relative to a given category. The method is based on the idea of querying the robot's ontological knowledge to retrieve the properties that an entity would need to fulfill to be an instance of that given category. The robot can then compare these properties to those that it already knows for the instance. Working under an open world assumption, the robot can then consider any remaining properties as gaps, indicating ignorance. This basic idea is similar to slot-filling strategies in *information states-based dialogue management* [63]. An information state is a set of records of what we would like to know, and what we already know. Any open records identify "gaps" that we need to fill next – for example, if our state reflects booking a train ticket, records may indicate departure, arrival, destination, etc. A dialogue system for booking a ticket then will ask the user for all these bits of information, to ensure it can get the user the right ticket. Here we face something similar: obtain all the information for a set of properties so that we can establish the entity as an instance of a given category. Having said that, Zender & Kruijff indicate how the method has the potential to go beyond a slot-filling strategy, in several ways. They argue how the method can extended to deal with uncertainty in categorization, and use weighted abduction of the kind proposed by Kruijff & Janiček (§2.2.3) to provide a "lowest-cost" way of establishing the right category for the entity. This again follows up on the general CogX perspective, integrating different

sources of information to help overcome uncertainty in understanding (perceptual data, ontological knowledge) to drive inferences towards establishing an interpretation (weighted abduction). Zender & Kruijff extend a recent method for verbalizing ontological structure [55] to properly reflect what the robot knows about the category, the instance, and the gaps it has identified.

1.3.2 Clarification

Kruijff & Brenner (§2.2.1) propose the notion of *question nucleus*. This notion captures the information pertaining to a question. A description logic-like formalism is used to represent such information, as a conceptual structure in which propositions have ontological sorts and unique indices, and can be related through named relations. A question can then be represented as a structure in which we are querying one or more aspects of such a representation [23, 36]. The formalism allows everything to be queried: relations, propositions, sorts. The nucleus altogether comprises the situation (the "facts") and the beliefs that have led up to the question, the question itself, and the goal content which would resolve the question. The question nucleus thus integrates Ginzburg's notions of *aboutness* and *(potential) resolvedness*, and includes an explicit notion of what information is shared, and what is privately held information (cf. [42, 26]). Intuitively, it thus represents what the robot is asking about (aboutness), what it would like to know (resolvedness), and how it can appeal to shared beliefs or needs to make clear private beliefs when raising the question. The contributions the approach aims for are, briefly, as follows. Purver and Ginzburg develop an account for generating questions in a dialogue context [51, 50]. Their focus was, however, on clarification for the purpose of dialogue grounding. A similar observation can be made for recent work in HRI [41]. Kruijff & Brenner are more interested in formulating questions regarding issues in building up situation awareness, including the acquisition of new ways of understanding situations (cf. also [36]). In issue-based (or information state-based) dialogue systems [40], the problem of how to phrase a question is greatly simplified because the task domain is fixed. There is little need for paying attention to transparency or scaffolding, as it can be assumed the user understands the task domain. This is however an assumption that cannot be made for our setting.

Kruijff & Janiček (§2.2.3) provide a model for capturing the *continual* nature of *collaborative activity*. They base their approach on an algorithm in which a form of *weighted abduction* plays a core role. Weighted abduction is "inference to the best explanation" – meaning, in this context, the best explanation for why someone is saying something, and formulating that explanation in terms of an intention, an update to a belief model, and possible updates to an attentional state. Using weighted abduction for interpretation of natural language was introduced by Hobbs et al in [30]. Kruijff &

Janiček use an extended form, proposed by Stone & Thomason [60, 61, 62]. Stone & Thomason's approach integrates *attentional state*, intention, and beliefs. Their attentional state captures those entities that are currently "in focus" or highly salient in the context. (Kruijff & Janiček turn this into beliefs about such entities.) The approach is related to other collaborative models of dialogue [27, 42, 26], and provides a single model for both comprehension and production. Stone & Thomason's notion of "context" provides for a more flexible way of resolving contextual references than classical discourse theories, though. Beliefs, intentions, and attentional state can all co-determine the conditions on resolving a reference – rather than that resolution is solely determined by structural aspects of discourse (like in e.g. SDRT [2]). This provides a suitable bridge to the continuum between action and interaction, which Kruijff & Brenner have argued for, cf. Brenner et al §2.2.2. Kruijff & Janiček propose to extend Stone & Thomason's approach with a more explicit notion of situated multi-agent belief models, and they introduce assertions into proofs. An assertion is a statement whose "future necessary truth" needs to be assumed for a proof to conclude. This notion of assertion is taken from continual planning [6] where it is used to state the necessity of a future observation. Depending on the verification of the observation, an assertion triggers explicit expanding or revision of a plan. Within an abductive proof, an assertion turns the corresponding action plan into a continual plan, to achieve the inferred update to the agent's belief model and attentional state. Assertions thus make Stone & Thomason's intuitive idea of "checkpoints" more precise. Kruijff & Janiček explore the use of assertions in abductive proofs in the context of producing and comprehending clarification dialogues.

Kruijff-Korbayová et al (§2.2.4) explore intonation in situated dialogue, with a particular focus on intonation in questions like clarification requests. Intonation of clarification requests has so far received relatively little attention in the literature. Previous work on controling accent placement and type in dialogue system output based on information structure assigment w.r.t. the context all concentrated on the assignment of intonation in statements [49, 38, 4]. The seminal work of [51] which laid out a classification of the forms and functions of clarification requests based on extensive corpus analysis does not take intonation into account. Pioneering in this respect is the study of CRs in German task-oriented human-human dialogues in [52], who found that the use of intonation seemed to disambiguate clarification types, with rising boundary tones used more often to clarify acoustic problems than to clarify reference resolution. A series of production and perception experiments with one-word grounding utterances in Swedish has also shown differences in prosodic features depending on meaning (acknowledgment vs. clarification of understanding or perception), and that subjects differentiate between the meanings accordingly, and respond differently [17, 58]. The work by Kruijff-Korbayová et al extends the use of information

structure to control the intonation of dialogue system output beyond answers to information-seeking questions: they include acknowledgments as well as clarification requests, and ultimately other types of questions. They include both fragmentary grounding feedback and full utterances, and address varying placement of pitch accents depending on context and communicative intention.

1.3.3 Robust processing of spoken situated dialogue

Lison & Kruijff's work on robust processing (§2.3.1, §2.3.2) aims to address two central issues in spoken dialogue processing: (1) disfluencies in verbal interaction and (2) speech recognition errors.

We know from everyday experience that spoken language behaves quite differently from written language. We do not speak the way we write. The difference of communicative medium plays a major role in this discrepancy. A speech stream offers for instance no possibility for "backtracking" – once something has been uttered, it cannot be erased anymore. And, contrary to written language, the production of spoken language is strongly *time-pressured*. The pauses which are made during the production of an utterance do leave a trace in the speech stream. As a consequence, spoken dialogue is replete with *disfluencies* such as filled pauses, speech repairs, corrections or repetitions [56]. A speech stream is also more difficult to segment and delimitate than a written sentence with punctuation and clear empty spaces between words. In fact, the very concepts of "words" and "sentences", which are often taken as core linguistic objects, are much more difficult to define with regard to spoken language. When we analyse spoken language, we observe a continuous speech stream, not a sequence of discrete objects. Hence the presence of many *discourse markers* in spoken dialogue, which play an important role in determining discourse structure. A final characteristic of spoken dialogue which is worth pointing out is that few spoken utterances take the form of complete sentences. The most prototypical example is the "short answer" in response to queries, but many other types of fragments or *non-sentential utterances* can be found in real dialogues [19]. This is mainly due to the *interactive* nature of dialogue – dialogue participants heavily rely on what has been said previously, and seek to avoid redundancies. As a result of all these factors, spoken language contains much more disfluent, partial, elided or ungrammatical utterances than written language. The question of how to *accommodate* these types of ill-formed input is a major challenge for spoken dialogue systems.

A second, related problem is *automatic speech recognition* (ASR). Speech recognition is the first step in comprehending spoken dialogue, and a very important one. For robots operating in real-world, noisy environments, and dealing with utterances pertaining to complex, open-ended domains, this step is also highly error-prone. In spite of continuous technological advances,

the performance of ASR indeed remains for most tasks at least an order of magnitude worse than that of human listeners [44]. And contrary to human performance, ASR accuracy is usually unable to *degrade gracefully* when faced with new conditions in the environment (ambient noise, bad microphone, non-native or regional accent, variations in voice intensity, etc.) [12]. This less-than-perfect performance of ASR technology seriously hinders the robustness of dialogue comprehension systems, and new techniques are needed to alleviate this problem[1].

The papers included in this deliverable present an integrated approach to dealing with these problems. The approach has three defining characteristics:

1. It is a hybrid approach, combining symbolic and statistical methods to process spoken dialogue. The implemented mechanisms combine fine-grained linguistic resources (a CCG lexicon) with statistical information (the ASR language model and the discriminative model). The resulting system therefore draws from the best of both worlds and is able to deliver both *deep* and *robust* language processing.

2. It is also an integrated approach to spoken dialogue comprehension. It goes all the way from the signal processing of the speech input up to the logical forms and the pragmatic interpretation. The various components involved in dialogue processing interact with each other in complex ways to complement, coordinate and constrain their internal representations.

3. Finally, it is also a context-sensitive approach. Contextual information is used at each processing step, either as an *anticipatory* mechanism (to guide expectations about what is likely to be uttered next), or as a *discriminative* mechanism (to prune interpretations which are contextually unlikely). These mechanisms are implemented by the dynamic adaptation of the ASR language model and the use of contextual features in the discriminative model for robust parsing.

This approach compares to the state of the art in robust processing of spoken dialogue, as follows. Commercial spoken dialogue systems traditionally rely on shallow parsing techniques such as "concept spotting". In this approach, a small hand-crafted, task-specific grammar is used to extract specific constituents, such as locative phrases or temporal expressions, and turn these into basic semantic concepts [65, 32, 3, 16, 1]. These techniques are usually very efficient, but also present several important shortcomings,

[1]The speech recogniser included into our robotic platform – Nuance Recognizer v8.5 with statistical language models – yields for instance a word error rate (WER) of about 20 % when evaluated on real spoken utterances. Thus, more than *one word out of five* in each utterance is actually misrecognised by the system.

as they are often highly domain-specific, fragile, and require a lot of development and optimisation effort to implement. In more recent years, several new techniques emerged, mainly based on statistical approaches. In the CHORUS system [47], the utterances are modeled as Hidden Markov Models [HMMs], in which hidden states correspond to semantic concepts and the state outputs correspond to the individual words. HMMs are however a flat-concept model – the semantic representation is just a linear sequence of concepts with no internal structure. To overcome this problem, various stochastic parsing techniques have been proposed, based either on Probabilistic Context Free Grammars [43, 20], lexicalised models [13, 10], data-oriented parsing [5, 57], or constrained hierarchical models [29]. A few recent systems, such as the SOUP parser, also attempt to combine shallow parsing with statistical techniques, based on a hand-crafted grammar associated with probabilistic weights [22]. More rarely, we can also find in the literature some descriptions of spoken dialogue systems performing a real grammatical analysis, usually along with a "robustness" mechanism to deal with speech recognition errors, extra-grammaticality [64, 9] or ill-formed inputs [66].

Compared to the state of the art, our approach is unique in the sense that it is, to the best of our knowledge, the only one which attempts to combine deep grammatical analysis together with statistical discriminative models exploiting both linguistic and contextual information. This has arguably several advantages. Using a deep processing approach, we are able to extract full, detailed semantic representations, which can then be used to draw inferences and perform sophisticated dialogue planning. This is not possible with shallow or statistical methods. At the same time, due to the grammar relaxation mechanism and the discriminative model, we do not suffer from the inherent fragility of purely symbolic methods. Our parsing method is particularly robust, both to speech recognition errors and to ill-formed utterances. Finally, contrary to "concept spotting" techniques, our approach is much less domain-specific: the parser relies on a general-purpose lexicalised grammar which can be easily reused in other systems.

Our approach is also original in its tight integration of multiple knowledge sources – and particularly contextual knowledge sources – all through the utterance comprehension process. Many dialogue systems are designed in a classical modular fashion, where the output of a component serves as direct input for the next component, with few or no interactions other than this pipelined exchange of data[2]. Our strategy, however, is to put the tight, multi-level integration of linguistic and contextual information at the very center of processing.

As a final note, we would like to stress that our dialogue comprehension system also departs from previous work in the way we define "context".

[2]Some interesting exceptions to this design include integrated approaches such as [45, 21].

Many recent techniques have been developed to take context into account in language processing (see e.g. [28]). But the vast majority of these approaches take a rather narrow view of context, usually restricting it to the mere dialogue/discourse context. Our dialogue comprehension system is one of the only ones (with the possible exceptions of [54, 8, 25]) to define context in a multimodal fashion, with a special focus on situated context.

2 Annexes

2.1 Verbalization

2.1.1 Zender et al. "A Situated Context Model for Resolution and Generation of Referring Expressions" (ENLG'09)

Bibliography H. Zender, G.J.M. Kruijff, and I. Kruijff-Korbayová. "A Situated Context Model for Resolution and Generation of Referring Expressions." In: Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009). pp. 126–129. Athens, Greece. March 2009.

Abstract The background for this paper is the aim to build robotic assistants that can naturally interact with humans. One prerequisite for this is that the robot can correctly identify objects or places a user refers to, and produce comprehensible references itself. As robots typically act in environments that are larger than what is immediately perceivable, the problem arises how to identify the appropriate context, against which to resolve or produce a referring expression (RE). Existing algorithms for generating REs generally by-pass this problem by assuming a given context. In this paper, we explicitly address this problem, proposing a method for context determination in large-scale space. We show how it can be applied both for resolving and producing REs.

Relation to WP The paper makes it possible for the robot to discuss objects and places beyond the currently perceivable situation. That makes it unnecessary for a robot and a human to be in the very place where there is something a robot needs to be explained.

2.1.2 Zender et al. "Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants." (IJCAI'09)

Bibliography H. Zender, G.J.M. Kruijff, and I. Kruijff-Korbayová. "Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants." In: Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09). Pasadena, CA, USA. July 2009.

Abstract In this paper we present an approach to the task of generating and resolving referring expressions (REs) for conversational mobile robots. It is based on a spatial knowledge base encompassing both robot- and human-centric representations. Existing algorithms for the generation of referring expressions (GRE) try to find a description that uniquely identifies the referent with respect to other entities that are in the current context. Mobile robots, however, act in large-scale space, that is environments that are larger than what can be perceived at a glance, e.g. an office building with different floors, each containing several rooms and objects. One challenge when referring to elsewhere is thus to include enough information so that the interlocutors can extend their context appropriately. We address this challenge with a method for context construction that can be used for both generating and resolving REs two previously disjoint aspects. Our approach is embedded in a bi-directional framework for natural language processing for robots.

Relation to WP The paper further explores how a robot can discuss objects and places outside the current situation (cf. also §2.1.1). The paper shows how determining the appropriate context for a reference can be integrated in a bi-directional approach, to enable the robot to both produce and comprehend contextually appropriate references.

2.1.3 Zender and Pronobis. "Verbalizing vague scalar predicates for autonomously acquired ontological knowledge" (report)

Bibliography H. Zender and A. Pronobis. "Verbalizing vague scalar predicates for autonomously acquired ontological knowledge" (report)

Abstract The paper reports on ongoing research in generating and understanding verbal references to entities in the robot's environment. The paper focuses on features of spatial entities that are commonly expressed as vague scalar predicates in natural language, such as, e.g., size. The paper proposes an approach for characterizing such features in terms of properties and distributions over their values. This leads to a basic notion of prototypicality of property-values. Using this notion, the paper shows how different types of contextual standards can be defined, which determine the contextually appropriate use of a vague scalar predicate in linguistically describing a feature of a spatial entity. The approach goes beyond existing work in that it allows for a variety of contextual standards (in class, across classes, across instances) in describing features as vague scalar predicates, and by ultimately basing these standards in models of the robot's perceptual experience.

Relation to WP Typically a robot is provided with an initial ontology, outlining the concepts and their relations considered relevant for understanding the environment in which the robot is to act. Over time, the robot can extend this ontology, for example with instances and properties that hold for these. The report develops a new method for verbalizing knowledge about autonomously acquired scalar properties for instances and their classes.

2.1.4 Zender and Kruijff. "Verbalizing classes and instances in ontological knowledge" (report)

Bibliography H. Zender and G.J.M. Kruijff. "Verbalizing classes and instances in ontological knowledge" (report)

Abstract The paper reports preliminary research on verbalizing a robot's knowledge about an instance I of a particular category C. This covers both what a robot knows, and what it does not (yet) know about the instance. The paper considers a "gap" to be that information the robot misses to establish a given property P for I, knowing that that property typically applies to instances of C. The paper proposes a method for determining which properties are classifiable as gaps for an instance relative to a category. This method operates on the T- and A-box of an ontology. It provides a general method for determining gaps, and is not specific to situated dialogue. The paper shows how the resulting characterization of available and missing knowledge about I relative to C can then be verbalized, following up an approach recently presented in [55]. The paper illustrates the method on an example involving spatial entities, and discusses further research on extending the method.

Relation to WP The report provides a first attempt at verbalizing ontological knowledge about classes and instances, with a particular focus on verbalizing what a robot does not yet know about a particular instance (i.e. a "gap").

2.2 Clarification

2.2.1 Kruijff and Brenner. "Phrasing Questions" (AAAI SS'09)

Bibliography G.J.M. Kruijff and M. Brenner. "Phrasing Questions." In: Proceedings of the AAAI 2009 Spring Symposium on Agents that Learn from Human Teachers. Stanford, CA. March 2009.

Abstract In a constructive learning setting, a robot builds up beliefs about the world by interacting – interacting with the world, and with other agents. Asking questions is key in such a setting. It provides a mechanism for interactively exploring possibilities, to extend and explain the robot's beliefs. The paper focuses on how to linguistically phrase questions in dialogue. How well the point of a question gets across depends on how it is put. It needs to be effective in making transparent the agent's intentions and beliefs behind raising the question, and in helping to scaffold the dialogue such that the desired answers can be obtained. The paper proposes an algorithm for deciding what to include in formulating a question. Its formulation is based on the idea of considering transparency and scaffolding as referential aspects of a question.

Relation to WP The paper considers what beliefs to use as context for a question (considered as a subdialogue). The paper defines the notion of a question nucleus. This structure identifies beliefs that provide a background for the question, the expected answers to the question, and a plan for formulating the question. The identified beliefs provide the basis for determining how to achieve transparency in phrasing the question, by relating aspects of the question nucleus to private and shared beliefs.

2.2.2 Brenner et al. "Continual Collaborative Planning for Situated Interaction" (report)

Bibliography M. Brenner, G.J.M. Kruijff, I. Kruijff-Korbayová, and N.A. Hawes. "Continual Collaborative Planning for Situated Interaction."

Abstract When several agents are situated in a common environment they usually interact both verbally and physically. Human-Robot Interaction (HRI) is a prototypical case of such situated interaction. It requires agents to closely integrate dialogue with behavior planning, physical action execution, and perception. The paper describes a framework called Continual Collaborative Planning (CCP) and its application to HRI. CCP enables agents to autonomously plan and realise situated interaction that intelligently interleaves planning, acting, and communicating. The paper analyses the behavior and efficiency of CCP agents in simulation, and on two robot implementations.

Relation to WP The paper argues for the continual nature of dialogue processing, reacting to the dynamics of the collaborative activity encompassing the actions of the different agents, and their interaction.

2.2.3 Kruijff and Janiček. "Abduction for clarification in situated dialogue" (report)

Bibliography G.J.M. Kruijff and M. Janiček. "Abductive inference for clarification in situated dialogue" (report)

Abstract A robot can use situated dialogue with a human, in an effort to learn more about the world it finds itself in. When asking the human for more information, it needs to be clear to the human, what the robot is talking about. The robot needs to make transparent what it would like to know more about, what it does know (or doesn't), and what it is after. Otherwise, the human is less likely to provide a useful answer to the robot. They need to establish a common ground in. The paper presents ongoing research on developing an approach for comprehending and producing (sub-)dialogues for clarifying or requesting information about the world in which establishing common ground in beliefs, intentions, and attention plays an explicit role. The approach is based on Stone & Thomason's abductive framework [60, 61, 62]. This framework integrates intention, attentional state, and dynamic interpretation to abductively derive an explanation on what assumptions and intentions communicated content can be interpreted as updating a belief context. The approach extends the framework of Stone & Thomason with assertions, to provide an explicit notion of checkpoint, and a more explicit form of multi-agent beliefs [6]. The approach uses these notions to formulate clarification as continual process of comprehension and production set in dialogue as a collaborative activity.

Relation to WP The report details a continual approach for managing clarification dialogues, based on an extended form of weighted abductive inference. The inference process covers both comprehension and production, in an interleaved fashion. The approach integrates intention, attentional state, and multi-agent belief models in a continual way of dealing with dialogue as a collaborative activity.

2.2.4 Kruijff-Korbayová et al. "Contextually appropriate intonation of clarification in situated dialogue" (report)

Bibliography I. Kruijff-Korbayová, R. Meena, and G.J.M. Kruijff. "Contextually appropriate intonation of clarification in situated dialogue." Report.

Abstract When in doubt, ask. This paradigm very much applies to autonomous robots which self-understand and self-extend in the environment they find themselves. For this, it is essentially for these systems to learn continuously, driven mainly by their own curiosity about the surroundings. Spoken dialogue is a means through which a robot can clarify or extend the acquired knowledge about the situated environment. This ability to self-initiate a dialogue to actively seek information or clarifications besides adding autonomy to a robot's behavior also allows the robot to connect its belief system to that of its listener. This access to respective belief systems in a dialogue helps the participating agents in dialogue *grounding*. However, for conversational robots raising clarification requests to seeking information is not only limited to contextually appropriate lexical selection and utterance content planning, but extends further to the generation of contextually appropriate intonation. In the absence of contextually appropriate intonation, dialogue participants might be lead to maintain incongruous belief state in wake of situational ambiguities that may arise in situated dialogue. Use of contextually appropriate intonation in clarification statements will enable the robot to rightly express its intentions to the human interlocutor. In this work we develop an approach for determining contextually appropriate intonation in clarification statements, for resolving situated ambiguities. Following the approaches [24, 50, 51] to clarification in human dialogue, we develop clarification strategies in human-robot dialogue for continuous and cross-modal learning. Working in the lines of Steedman's theory of *information structure* [59, 48] and [18], we propose and develop the notion of information packaging in our clarification statements. We evaluate our approach to generation of contextually appropriate intonations using psycholinguistically plausible experimental setup.

Relation to WP When a robot raises a question, or more in general says something in a given context, it is important for it to be clear how the utterance relates to the preceding context – and what it focuses on. Intonation is one such means to indicate this relation to context.

2.3 Robust processing of spoken situated dialogue

Increased robustness, ultimately reflected as an improvement in understanding what the user has said, contributes to efficient and effective dialogue: the better the understanding, the less need for corrective measures (e.g. clarification).

2.3.1 Lison and Kruijff. "An integrated approach to robust processing of situated spoken dialogue." (SRSL'09)

Bibliography P. Lison and G.J.M. Kruijff. "An integrated approach to robust processing of situated spoken dialogue." In: Proceedings of the Second International Workshop on the Semantic Representation of Spoken Language (SRSL'09). Athens, Greece. April 2009

Abstract Spoken dialogue is notoriously hard to process with standard NLP technologies. Natural spoken dialogue is replete with disfluent, partial, elided or ungrammatical utterances, all of which are difficult to accommodate in a dialogue system. Furthermore, speech recognition is known to be a highly error-prone task, especially for complex, open-ended domains. The combination of these two problems - ill-formed and/or misrecognised speech inputs - raises a major challenge to the development of robust dialogue systems. We present an integrated approach for addressing these two issues, based on an incremental parser for Combinatory Categorial Grammar. The parser takes word lattices as input and is able to handle ill-formed and misrecognised utterances by selectively relaxing its set of grammatical rules. The choice of the most relevant interpretation is then realised via a discriminative model augmented with contextual information. The approach is fully implemented in a dialogue system for autonomous robots. Evaluation results on a Wizard of Oz test suite demonstrate very significant improvements in accuracy and robustness compared to the baseline.

Relation to WP The paper describes an approach in which context information (salient entities, properties, and actions) is used to anticipate likely word sequences (biasing the lexical activations of words in a language model), and to discriminate (complete) parses. This yields improvements in robustness, resulting in a lower word error rate (WER) and an improvement in partial- and exact-matches of semantic representations against a WoZ corpus.

2.3.2 Lison and Kruijff, "Efficient parsing of spoken inputs for human-robot interaction" (RO-MAN'09)

Bibliography P. Lison and G.J.M. Kruijff. "Efficient parsing of spoken inputs for human-robot interaction." In: Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'09). Toyama, Japan. September 2009.

Abstract The use of deep parsers in spoken dialogue systems is usually subject to strong performance requirements. This is particularly the case in human-robot interaction, where the computing resources are limited and must be shared by many components in parallel. A real-time dialogue system must be capable of responding quickly to any given utterance, even in the presence of noisy, ambiguous or distorted input. The parser must therefore ensure that the number of analyses remains bounded at every processing step. The paper presents a practical approach to address this issue in the context of deep parsers designed for spoken dialogue. The approach is based on a word lattice parser combined with a statistical model for parse selection. Each word lattice is parsed incrementally, word by word, and a discriminative model is applied at each incremental step to prune the set of resulting partial analyses. The model incorporates a wide range of linguistic and contextual features and can be trained with a simple perceptron. The approach is fully implemented as part of a spoken dialogue system for human-robot interaction. Evaluation results on a Wizard-of-Oz test suite demonstrate significant improvements in parsing time.

Relation to WP Whereas the (SRSL'09) paper only considers the uses of discriminative models at the end of the parsing process, the current paper employs discriminative models after each incremental step during parsing. A discriminative models ranks all partial analyses, after which the top-30 ranked analyses are selected for further processing. The paper shows a 50% improvement in parsing time, without any significant loss in performance (partial/exact match). Improvements in processing time make it possible for the system to have a faster response-time.

References

[1] J.F. Allen, B.W. Miller, E.K. Ringger, and T. Sikorski. A robust system for natural spoken dialogue. In *ACL'96: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, USA, 1996. Association for Computational Linguistics.

[2] N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, Cambridge, UK, 2003.

[3] Harald Aust, Martin Oerder, Frank Seide, and Volker Steinbiss. The philips automatic train timetable information system. *Speech Communications*, 17(3-4):249–262, 1995.

[4] Rachel Baker, Robert A. J. Clark, and Michael White. Synthetizing contextually appropriate intonation in limited domains. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, 2004.

[5] Rens Bod. Context-sensitive spoken dialogue processing with the dop model. *Natural Language Engineering*, 5(4):309–323, 1999.

[6] M. Brenner and B. Nebel. Continual planning and acting in dynamic multiagent environments. *Journal of Autonomous Agents and Multiagent Systems*, 2008.

[7] Roger Brown. How shall a thing be called? *Psychological Review*, 65(1):14–21, 1958.

[8] J. Y. Chai and Sh. Qu. A salience driven approach to robust input interpretation in multimodal conversational systems. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing 2005*, pages 217–224, Vancouver, Canada, October 2005. Association for Computational Linguistics.

[9] J.-P. Chanod. Robust parsing and beyond. In Gertjan van Noord and J. Juncqua, editors, *Robustness in Language Technology*. Kluwer, 2000.

[10] Eugene Charniak. Immediate-head parsing for language models. In *ACL '01: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 124–131, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[11] H. Clark. *Using Language*. Cambridge University Press, Cambridge, UK, 1996.

[12] Ronald A. Cole and Victor Zue. Spoken language input. In Ronald A. Cole, Joseph Mariana, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge, 1997.

[13] M. Collins. Three generative, lexicalised models for statistical parsing. In *ACL-35: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[14] Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995.

[15] D. DeVault and M. Stone. Interpreting vague utterances in context. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, 2004.

[16] John Dowding, Robert Moore, Francois Andry, and Douglas Moran. Interleaving syntax and semantics in an efficient bottom-up parser. In *ACL-94: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 110–116. Association for Computational Linguistics, 1994.

[17] Jens Edlund, Davod House, and Gabriel Skantze. The effects of prosodic features on the interpretation of clarification ellipses. In *Proceedings of Interspeech. Lisbon, Portugal*, pages 2389–2392, 2005.

[18] E. Engdahl. Infromation packaging in questions. *Empirical Issues in Syntax and Semantics*, 6(1):93–111, 2006.

[19] R. Fernández and J. Ginzburg. A corpus study of non-sentential utterances in dialogue. *Traitement Automatique des Langues*, 43(2):12–43, 2002.

[20] Shai Fine. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

[21] Malte Gabsdil and Johan Bos. Combining acoustic confidence scores with deep semantic analysis for clarification dialogues. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5*, pages 137–150, 2003.

[22] Marsal Gavaldà. Soup: a parser for real-world spontaneous speech. In *New developments in parsing technology*, pages 339–350. Kluwer Academic Publishers, Norwell, MA, USA, 2004.

[23] J. Ginzburg. The semantics of interrogatives. In S. Lappin, editor, *Handbook of Contemporary Semantic Theory*. Blackwell, 1995.

[24] J. Ginzburg. Interrogatives: Questions, facts and dialogue. In *The Handbook of Contemporary Semantic Theory*, pages 385–422. Blackwell, 1996.

[25] P. Gorniak and D. Roy. Situated language understanding as filtering perceived affordances. *Cognitive Science*, 31(2):197–231, 2007.

[26] B.J. Grosz and S. Kraus. The evolution of shared plans. In A. Rao and M. Wooldridge, editors, *Foundations and Theories of Rational Agency*, pages 227–262. Springer, 1999.

[27] B.J. Grosz and C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[28] A. Gruenstein, C. Wang, and S. Seneff. Context-sensitive statistical language modeling. In *Proceedings of INTERSPEECH 2005*, pages 17–20, 2005.

[29] Yulan He and S. Young. Semantic processing using the hidden vector state model. computer speech and language. *Computer Speech and Language*, 19:85–106, 2005.

[30] J.R. Hobbs, M. Stickel, D. Appelt, and P. Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1-2):69–142, 1993.

[31] Helmut Horacek. An algorithm for generating referential descriptions with flexible interfaces. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conf. of the EACL*, Madrid, Spain, 1997.

[32] Eric Jackson, Douglas Appelt, John Bear, Robert Moore, and Ann Podlozny. A template matcher for robust nl interpretation. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 190–194, Morristown, NJ, USA, 1991. Association for Computational Linguistics.

[33] J. Kelleher and G.J.M. Kruijff. Incremental generation of spatial referring expressions in situated dialogue. In *Proc. Coling-ACL-2006*, Sydney, Australia, 2006.

[34] C. Kennedy. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45, February 2007.

[35] E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R.Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*. CSLI Publications, Stanford, CA, USA, 2002.

[36] G.J.M. Kruijff, M. Brenner, and N.A. Hawes. Continual planning for cross-modal situated clarification in human-robot interaction. In *Proceedings of the 17th International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, Munich, Germany, 2008.

[37] G.J.M. Kruijff, H. Zender, P. Jensfelt, and H.I. Christensen. Clarification dialogues in human-augmented mapping. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI'06)*, 2006.

[38] Ivana Kruijff-Korbayová, Stina Ericsson, Kepa Joseba Rodríguez, and Elena Karagjosova. Producing contextually appropriate intonation is an information-state based dialogue system. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 227–234. ACL, 2003.

[39] B. Kuipers. *Representing Knowledge of Large-scale Space*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1977.

[40] Staffan Larsson. *Issue-Based Dialogue Management*. Phd thesis, Department of Linguistics, Göteborg University, Göteborg, Sweden, 2002.

[41] S. Li, B. Wrede, and G. Sagerer. A computational model of multi-modal grounding. In *Proceedings of the ACL SIGdial workshop on discourse and dialog*, pages 153–160, 2006.

[42] K. Lochbaum, B.J. Grosz, and C.L. Sidner. Discourse structure and intention recognition. In R. Dale, H. Moisl, , and H. Somers, editors, *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. Marcel Dekker, New York, 1999.

[43] Scott Miller, Richard Schwartz, Robert Bobrow, and Robert Ingria. Statistical language processing using hidden understanding models. In *HLT '94: Proceedings of the workshop on Human Language Technology*, pages 278–282, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[44] R. K. Moore. Spoken language processing: piecing together the puzzle. *Speech Communication: Special Issue on Bridging the Gap Between Human and Automatic Speech Processing*, 49:418–435, 2007.

[45] Robert Moore, John Dowding, J. M. Gawron, and Douglas Moran. Combining linguistic and statistical knowledge sources in natural-language processing for atis. In *ARPA Spoken Language Technology Workshop*, 1995.

[46] I. Paraboni, K. van Deemter, and J. Masthoff. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, June 2007.

[47] Roberto Pieraccini, Evelyne Tzoukermann, Zakhar Gorelov, Esther Levin, Chin-Hui Lee, and Jean-Luc Gauvain. Progress report on the Chronus system: ATIS benchmark results. In *HLT '91: Proceedings*

of the workshop on Speech and Natural Language, pages 67–71, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[48] Scott A. Prevost. An information structural approach to spoken language generation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 294–301, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[49] Scott A. Prevost. *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. Phd thesis, University of Pennsylvania, Institute for Research in Cognitive Science Technical Report, Pennsylvania, USA, 1996.

[50] M. Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, King's College, University of London, 2004.

[51] M. Purver, J. Ginzburg, and P. Healey. On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*, pages 235–255. Kluwer Academic Publishers, 2003.

[52] Kepa J. Rodríguez and David Schlangen. Form, intonation and function of clarification requests in german task oriented spoken dialogues. In *Proceedings of Catalog '04 (The 8th Workshop on the Semantics and Pragmatics of Dialogue, SemDial04), Barcelona, Spain, July* , 2004.

[53] Eleanor Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1978.

[54] Deb Roy. Situation-aware spoken language processing. In *Royal Institute of Acoustics Workshop on Innovation in Speech Processing*, Stratford-upon-Avon, England, 2001.

[55] Niels Schütte. Generating natural language descriptions of ontology concepts. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 106–109, Athens, Greece, March 2009.

[56] E. Shriberg. Disfluencies in switchboard. In *Proceedings of ICSLP '96*, volume supplement, Philadelphia, PA, 1996.

[57] Khalil Sima'an. Robust data oriented spoken language understanding. In *New developments in parsing technology*, pages 323–338. Kluwer Academic Publishers, Norwell, MA, USA, 2004.

[58] Gabriel Skanze, David House, and Jens Edlund. User responses to prosodic variation in fragmentary grounding utterances in dialogue. In *Proceedings of Interspeech ICSLP. Pittsburgh PA, USA*, pages 2002–2005, 2006.

[59] M. Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31:649–689, 2000.

[60] M. Stone and R.H. Thomason. Context in abductive interpretation. In *Proceedings of EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue*, 2002.

[61] M. Stone and R.H. Thomason. Coordinating understanding and generation in an abductive approach to interpretation. In *Proceedings of DIABRUCK 2003: 7th workshop on the semantics and pragmatics of dialogue*, 2003.

[62] R.H. Thomason, M. Stone, and D. DeVault. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. In D. Byron, C. Roberts, and S. Schwenter, editors, *Presupposition Accommodation*. (to appear).

[63] David Traum and Staffan Larsson. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie Smith, editors, *Current and New Directions in Discourse and Dialogue*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

[64] G. van Noord, G. Bouma, R. Koeling, and M.-J. Nederhof. Robust grammatical analysis for spoken dialogue systems. *Journal of Natural Language Engineering*, 1999.

[65] Wayne Ward. Understanding spontaneous speech. In *HLT '89: Proceedings of the workshop on Speech and Natural Language*, pages 137–141, Morristown, NJ, USA, 1989. Association for Computational Linguistics.

[66] L. S. Zettlemoyer and M. Collins. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687, 2007.

A Situated Context Model for Resolution and Generation of Referring Expressions

Hendrik Zender and **Geert-Jan M. Kruijff** and **Ivana Kruijff-Korbayová**
Language Technology Lab, German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{zender, gj, ivana.kruijff}@dfki.de

Abstract

The background for this paper is the aim to build robotic assistants that can "naturally" interact with humans. One prerequisite for this is that the robot can correctly identify objects or places a user refers to, and produce comprehensible references itself. As robots typically act in environments that are larger than what is immediately perceivable, the problem arises how to identify the appropriate context, against which to resolve or produce a referring expression (RE). Existing algorithms for generating REs generally bypass this problem by assuming a given context. In this paper, we explicitly address this problem, proposing a method for context determination in large-scale space. We show how it can be applied both for resolving and producing REs.

1 Introduction

The past years have seen an extraordinary increase in research on robotic assistants that help users perform daily chores. Autonomous vacuum cleaners have already found their way into people's homes, but it will still take a while before fully conversational robot "gophers" will assist people in more demanding everyday tasks. Imagine a robot that can deliver objects, and give directions to visitors on a university campus. This robot must be able to verbalize its knowledge in a way that is understandable by humans.

A conversational robot will inevitably face situations in which it needs to refer to an entity (an object, a locality, or even an event) that is located somewhere outside the current scene, as Figure 1 illustrates. There are conceivably many ways in which a robot might refer to things in the world, but many such expressions are unsuitable in most

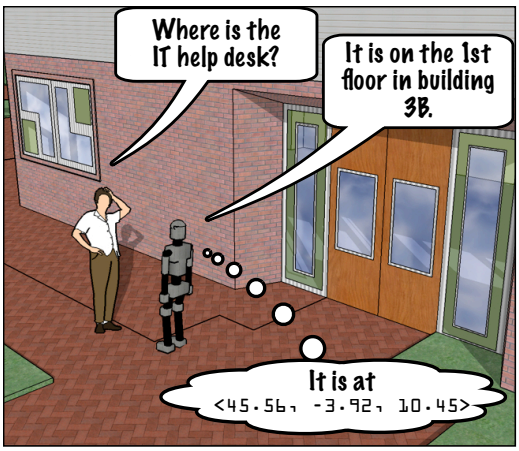


Figure 1: Situated dialogue with a service robot

human-robot dialogues. Consider the following set of examples:

1. "position $P = \langle 45.56, -3.92, 10.45 \rangle$"
2. "Peter's office no. 200 at the end of the corridor on the third floor of the Acme Corp. building 3 in the Acme Corp. complex, 47 Evergreen Terrace, Calisota, Earth, (...)"
3. "the area"

These REs are valid descriptions of their respective referents. Still they fail to achieve their *communicative goal*, which is to specify the right amount of information that the hearer needs to uniquely identify the referent. The next REs *might* serve as more appropriate variants of the previous examples (*in certain contexts!*):

1. "the IT help desk"
2. "Peter's office"
3. "the large hall on the first floor"

The first example highlights a requirement on the knowledge representation to which an algorithm for generating referring expressions (GRE) has access. Although the robot needs a robot-centric representation of its surrounding space that allows it to safely perform actions and navigate its world, it should use human-centric qualitative descriptions when talking about things in the world. We

do not address this issue here, but refer the interested reader to our recent work on multi-layered spatial maps for robots, bridging the gap between robot-centric and human-centric spatial representations (Zender et al., 2008).

The other examples point out another important consideration: how much information does the human need to single out the intended referent among the possible entities that the robot could be referring to? According to the seminal work on GRE by Dale and Reiter (1995), one needs to distinguish whether the intended referent is already in the hearer's *focus of attention* or not. This focus of attention can consist of a local visual scene (visual context) or a shared workspace (spatial context), but also contains recently mentioned entities (dialogue context). If the referent is already part of the current context, the GRE task merely consists of singling it out among the other members of the context, which act as distractors. In this case the generated RE contains *discriminatory* information, e.g. "the red ball" if several kinds of objects with different colors are in the context. If, on the other hand, the referent is not in the hearer's focus of attention, an RE needs to contain what Dale and Reiter call *navigational*, or *attention-directing* information. The example they give is "the black power supply in the equipment rack," where "the equipment rack" is supposed to direct the hearers attention to the rack and its contents.

In the following we propose an approach for context determination and extension that allows a mobile robot to produce and interpret REs to entities outside the current visual context.

2 Background

Most GRE approaches are applied to very limited, visual scenes – so-called *small-scale space*. The domain of such systems is usually a small visual scene, e.g. a number of objects, such as cups and tables, located in the same room), or other closed-context scenarios (Dale and Reiter, 1995; Horacek, 1997; Krahmer and Theune, 2002). Recently, Kelleher and Kruijff (2006) have presented an incremental GRE algorithm for situated dialogue with a robot about a table-top setting, i.e. also about small-scale space. In all these cases, the context set is assumed to be identical to the visual scene that is shared between the interlocutors. The intended referent is thus already in the hearer's *focus of attention*.

In contrast, robots typically act in *large-scale space*, i.e. space "larger than what can be perceived at once" (Kuipers, 1977). They need the ability to understand and produce references to things that are beyond the current visual and spatial context. In any situated dialogue that involves entities beyond the current focus of attention, the task of *extending the context* becomes key.

Paraboni et al. (2007) present an algorithm for *context determination* in hierarchically ordered domains, e.g. a university campus or a document structure. Their approach is mainly targeted at producing textual references to entities in written documents (e.g. figures, tables in book chapters). Consequently they do not address the challenges that arise in physically and perceptually situated dialogues. Still, the approach presents a number of good contributions towards GRE for situated dialogue in large-scale space. An appropriate context, as a subset of the full domain, is determined through Ancestral Search. This search for the intended referent is rooted in the "position of the speaker and the hearer in the domain" (represented as d), a crucial first step towards situatedness. Their approach suffers from the shortcoming that spatial relationships are treated as one-place attributes by their GRE algorithm. For example they transform the spatial containment relation that holds between a room entity and a building entity ("the library in the Cockroft building") into a property of the room entity (BUILDING NAME = COCKROFT) and not a two-place relation (`in(library,Cockroft)`). Thus they avoid recursive calls to the algorithm, which would be needed if the intended referent is related to another entity that needs to be properly referred to.

However, according to Dale and Reiter (1995), these related entities do not necessarily serve as discriminatory information. At least in large-scale space, in contrast to a document structure that is conceivably transparent to a reader, they function as *attention-directing elements* that are introduced to build up *common ground* by incrementally extending the hearer's focus of attention. Moreover, representing some spatial relations as two-place predicates between two entities and some as one-place predicates is an arbitrary decision.

We present an approach for context determination (or *extension*), that imposes less restrictions on its knowledge base, and which can be used as a sub-routine in existing GRE algorithms.

3 Situated Dialogue in Large-Scale Space

Imagine the situation in Figure 1 did not take place somewhere on campus, but rather inside building 3B. Certainly the robot would not have said "the IT help desk is on the 1st floor in building 3B." To avoid confusing the human, an utterance like "the IT help desk is on the 1st floor" would have been appropriate. Likewise, if the IT help desk happened to be located on another site of the university, the robot would have had to identify its location as being "on the 1st floor in building 3B on the new campus." The hierarchical representation of space that people are known to assume (Cohn and Hazarika, 2001), reflects upon the choice of an appropriate context when producing REs.

In the above example the physical and spatial situatedness of the dialogue participants play an important role in determining which related parts of space come into consideration as potential distractors. Another important observation concerns the verbal behavior of humans when talking about remote objects and places during a complex dialogue (i.e. more than just a question and a reply). Consider the following example dialogue:

> Person A: "Where is the exit?"
>
> Person B: "You first go down this corridor. Then you turn right. After a few steps you will see the big glass doors."
>
> Person A: "And the bus station? Is it to the left?"

The dialogue illustrates how utterances become grounded in previously introduced discourse referents, both temporally and spatially. Initially, the physical surroundings of the dialogue partners form the context for anchoring references. As a dialogue unfolds, this point can conceptually move to other locations that have been explicitly introduced. Discourse markers denoting spatial or temporal cohesion (e.g. "then" or "there") can make this move to a new anchor explicit, leading to a "mental tour" through large-scale space.

We propose a general principle of *Topological Abstraction* (TA) for context extension which is rooted in what we will call the *Referential Anchor* a.[1] TA is designed for a multiple abstraction hierarchy (e.g. represented as a lattice structure rather than a simple tree). The Referential Anchor a, corresponding to the current focus of attention, forms the nucleus of the context. In the simple case, a corresponds to the hearer's physical location. As illustrated above, a can also move along the "spatial progression" of the most salient discourse entity during a dialogue. If the intended referent is outside the current context, TA extends the context by incrementally ascending the spatial abstraction hierarchy until the intended referent is an element of the resulting sub-hierarchy, as illustrated in Figure 2. Below we describe two instantiations of the TA principle, a TA algorithm for reference generation (TAA1) and TAA2 for reference resolution.

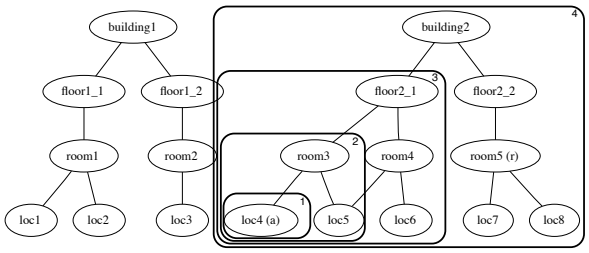


Figure 2: Incremental TA in large-scale space

Context Determination for GRE TAA1 constructs a set of entities dominated by the Referential Anchor a (and a itself). If this set contains the intended referent r, it is taken as the current utterance context set. Else TAA1 moves up one level of abstraction and adds the set of all child nodes to the context set. This loop continues until r is in the context set. At that point TAA1 stops and returns the constructed context set (cf. Algorithm 1).

TAA1 is formulated to be neutral to the kind of GRE algorithm that it is used for. It can be used with the original Incremental Algorithm (Dale and Reiter, 1995), augmented by a recursive call if a relation to another entity is selected as a discriminatory feature. It could in principle also be used with the standard approach to GRE involving relations (Dale and Haddock, 1991), but we agree with Paraboni et al. (2007) that the mutually qualified references that it can produce[2] are not easily resolvable if they pertain to circumstances where a confirmatory search is costly (such as in large-scale space). More recent approaches to avoiding infinite loops when using relations in GRE make use of a graph-based knowledge representation (Krahmer et al., 2003; Croitoru and van Deemter, 2007). TAA1 is compatible with these approaches, as well as with the salience based approach of (Krahmer and Theune, 2002).

[1] similar to Ancestral Search (Paraboni et al., 2007)

[2] An example for such a phenomenon is the expression "the ball on the table" in a context with several tables and several balls, but of which only one is on a table. Humans find such REs natural and easy to resolve in visual scenes.

Algorithm 1 TAA1 (for reference generation)

Require: a = referential anchor; r = intended referent
  Initialize context: $C = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  if $r \in C$ **then**
    return C
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ **do**
      for each $p \in topologicalParents(n)$ **do**
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      if $r \in C$ **then**
        return C
      end if
    end for
    return failure
  end if

Algorithm 2 TAA2 (for reference resolution)

Require: a = ref. anchor; $desc(x)$ = description of referent
  Initialize context: $C = \{\}$
  Initialize possible referents: $R = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  $R = desc(x) \cap C$
  if $R \neq \{\}$ **then**
    return R
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ **do**
      for each $p \in topologicalParents(n)$ **do**
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      $R = desc(x) \cap C$
      if $R \neq \{\}$ **then**
        return R
      end if
    end for
    return failure
  end if

Resolving References to Elsewhere Analogous to the GRE task, a conversational robot must be able to understand verbal descriptions by its users. In order to avoid overgenerating possible referents, we propose TAA2 (cf. Algorithm 2) which tries to select an appropriate referent from a relevant subset of the full knowledge base. It is initialized with a given semantic representation of the referential expression, $desc(x)$, in a format compatible with the knowledge base. Then, an appropriate entity satisfying this description is searched for in the knowledge base. Similarly to TAA1, the description is first matched against the current *context set* C consisting of a and its child nodes. If this set does not contain any instances that match $desc(x)$, TAA2 increases the context set along the spatial abstraction axis until at least one possible referent can be identified within the context.

4 Conclusions and Future Work

We have presented two algorithms for context determination that can be used both for resolving and generating REs in large-scale space.

We are currently planning a user study to evaluate the performance of the TA algorithms. Another important item for future work is the exact nature of the spatial progression, modeled by "moving" the referential anchor, in a situated dialogue.

Acknowledgments

This work was supported by the EU FP7 ICT Project "CogX" (FP7-ICT-215181).

References

A. G. Cohn and S. M. Hazarika. 2001. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46:1–29.

M. Croitoru and K. van Deemter. 2007. A conceptual graph approach to the generation of referring expressions. In *Proc. IJCAI-2007*, Hyderabad, India.

R. Dale and N. Haddock. 1991. Generating referring expressions involving relations. In *Proc. of the 5th Meeting of the EACL*, Berlin, Germany, April.

R. Dale and E. Reiter. 1995. Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

H. Horacek. 1997. An algorithm for generating referential descriptions with flexible interfaces. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conf. of the EACL*, Madrid, Spain.

J. Kelleher and G.-J. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialogue. In *In Proc. Coling-ACL 06*, Sydney, Australia.

E. Krahmer and M. Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R.Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*. CSLI Publications, Stanford, CA, USA.

E. Krahmer, S. van Erk, and A. Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1).

B. Kuipers. 1977. *Representing Knowledge of Large-scale Space*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

I. Paraboni, K. van Deemter, and J. Masthoff. 2007. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, June.

H. Zender, O. Martínez Mozos, P. Jensfelt, G.-J. Kruijff, and W. Burgard. 2008. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502, June.

Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants*

Hendrik Zender and **Geert-Jan M. Kruijff** and **Ivana Kruijff-Korbayová**
Language Technology Lab, German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{zender, gj, ivana.kruijff}@dfki.de

Abstract

In this paper we present an approach to the task of generating and resolving referring expressions (REs) for conversational mobile robots. It is based on a spatial knowledge base encompassing both robot- and human-centric representations. Existing algorithms for the generation of referring expressions (GRE) try to find a description that uniquely identifies the referent with respect to other entities that are in the current context. Mobile robots, however, act in large-scale space, that is, environments that are larger than what can be perceived at a glance, e.g., an office building with different floors, each containing several rooms and objects. One challenge when referring to elsewhere is thus to include enough information so that the interlocutors can extend their context appropriately. We address this challenge with a method for context construction that can be used for both generating and resolving REs – two previously disjoint aspects. Our approach is embedded in a bi-directional framework for natural language processing for robots.

1 Introduction

The past years have seen an extraordinary increase in research on robotic assistants that help the users perform their daily chores. Although the autonomous vacuum cleaner "Roomba" has already found its way into people's homes and lives, there is still a long way until fully conversational robot "gophers" will be able to assist people in more demanding everyday tasks. For example, imagine a robot that can deliver objects and give directions to visitors on a university campus. Such a robot must be able to verbalize its knowledge in a way that is understandable by humans, as illustrated in Figure 1.

A conversational robot will inevitably face situations in which it needs to refer to an entity (e.g., an object, a locality, or even an event) that is located somewhere outside the current scene. There are conceivably many ways in which a robot might refer to things in the world, but many such expressions are unsuitable in most human-robot dialogues. Consider the following set of examples:

*Supported by the EU FP7 Project "CogX" (FP7-ICT-215181).

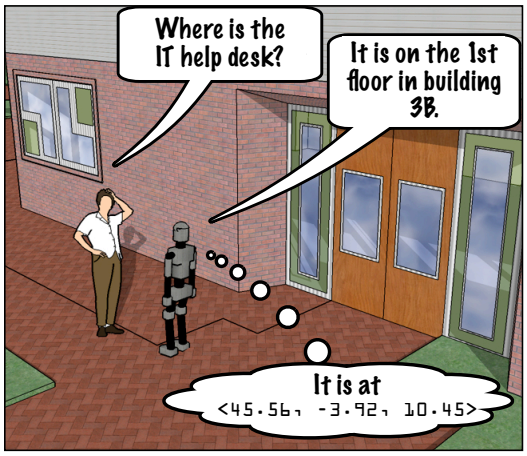


Figure 1: Situated dialogue with a campus service robot

1. "position $P = \langle 45.56, -3.92, 10.45 \rangle$"
2. "the area"
3. "Peter's office at the end of the corridor on the third floor of the Acme Corp. building 7 in the Acme Corp. complex, 47 Evergreen Terrace, Calisota, Earth, (...)"

Clearly, these REs are valid descriptions of the respective entities in the robot's world representation. Still they fail to achieve their *communicative goal*, which is to specify the right amount of information so that the hearer can easily uniquely identify what is meant. The following expressions *might* serve as more appropriate variants of the previous examples (*in certain situations!*):

1. "the IT help desk"
2. "the large hall on the first floor"
3. "Peter's office"

However, the question remains how a natural language processing (NLP) system can generate such expressions which are suitable in a given situation. In this paper we identify some of the challenges that an NLP system for situated dialogue about large-scale space needs to address. We present a situated model for generating and resolving REs that addresses these issues, with a special focus on how a conversational mobile robot can produce and interpret such expressions against an appropriate part of its acquired knowledge base (KB). One benefit of our approach is that most components, including the situated model and the linguistic resources, are bi-directional, i.e., they use the same representa-

tions for comprehension and production of utterances. This means that the proposed system is able to understand and correctly resolve all the REs that it is able to generate.

The rest of the paper is organized as follows. We first briefly discuss relevant existing approaches to comprehending and producing REs (Section 2). We then motivate our approach to context determination for situated interaction in large-scale space (Section 3), and describe its implementation in a dialogue system for an autonomous robot (Section 4). We conclude in Section 5.

2 Background

The main purpose of an RE is to enable a hearer to correctly and uniquely identify the target entity to which the speaker is referring, the so-called *intended referent*. The GRE task is thus to produce a natural language expression for a KB entity that fulfills this purpose.

As can be seen from the examples in the previous section, an RE needs to meet a number of constraints in order to be successful. First, it needs to make use of concepts that can be understood by the hearer. This becomes an important consideration when we are dealing with a robot which acquires its own models of the environment and is to talk about the contents of these. Second, it needs to contain enough information so that the hearer can distinguish the intended referent from other entities in the world, the so-called *potential distractors*. Finally, this needs to be balanced against the third constraint: Inclusion of unnecessary information should be avoided so as not to elicit false implications on the part of the hearer.

We will only briefly mention how to address the first challenge, and refer the reader to our recent work on multi-layered conceptual spatial maps for robots that bridge the gap between robot-centric representations of space and human-centric conceptualizations [Zender *et al.*, 2008].

The focus in this paper lies on the second and third aspect, namely the problem of including the right amount of information that allows the hearer to identify the intended referent. According to the seminal work on GRE by Dale and Reiter [1995], one needs to distinguish whether the intended referent is already in the hearer's *current context* or not. This context can consist of a local visual scene (visual context) or a shared workspace (spatial context), but also contains recently mentioned entities (dialogue context). If the intended referent is already part of the current context, the GRE task merely consists of singling out the referent among the other members of the context, which act as distractors. In this case the generated RE contains *discriminatory* information, e.g., "the red ball" if several kinds of objects with different colors are in the current context. If, on the other hand, the referent is not in the hearer's focus of attention, an RE needs to contain what Dale and Reiter call *navigational*, or *attention-directing* information. The example they give is "the black power supply in the equipment rack," where "the equipment rack" is supposed to direct the hearers attention to the rack and its contents.

While most existing GRE approaches assume that the intended referent is part of a given scene model, the *context set*, very little research has investigated the nature of references to entities that are not part of the current context.

The domain of such systems is usually a small visual scene, e.g., a number of objects, such as cups and tables, located in the same room, other closed-context scenarios, including a human-robot collaborative table-top scenario [Dale and Reiter, 1995; Horacek, 1997; Krahmer and Theune, 2002; Kelleher and Kruijff, 2006]. What these scenarios have in common is that they focus on a limited part of space, which is immediately and fully observable: *small-scale space*.

In contrast, mobile robots typically act in more complex environments. They operate in *large-scale space*, i.e., space "larger than what can be perceived at once" [Kuipers, 1977]. At the same time they do need the ability to understand and produce verbal references to things that are beyond the current visual and spatial context. When talking about remote places and things outside the current focus of attention, the task of *extending the context* becomes crucial.

Paraboni et al. [2007] are among the few to address this problem. They present an algorithm for *context determination* in hierarchically ordered domains, e.g., a university campus or a document structure. Their approach is mainly targeted at producing textual references to entities in written documents (e.g., figures and tables in book chapters), and consequently they do not touch upon the challenges that arise in a physically and perceptually situated dialogue setting. Nonetheless their approach presents a number of contributions towards GRE for situated dialogue in large-scale space. An appropriate context, as a subset of the full domain, is determined through Ancestral Search. This search for the intended referent is rooted in the "position of the speaker and the hearer in the domain" (represented as d), a crucial first step towards situatedness. Their approach suffers from the shortcoming that their GRE algorithm treats spatial relationships as one-place attributes. E.g., a spatial containment relation that holds between a room entity and a building entity ("the library in the Cockroft building") is given as a property of the room entity (BUILDING NAME = COCKROFT), rather than a two-place relation (`in(library,Cockroft)`). Thereby they avoid recursive calls to the GRE algorithm, which are necessary for intended referents related to another entity that needs to be properly referred to. We claim that this imposes an unnecessary restriction onto the KB design. Moreover, it makes it hard to use their context determination algorithm as a sub-routine of any of the many existing GRE algorithms.

3 Situated Dialogue in Large-Scale Space

Imagine the situation in Figure 1 did not take place somewhere on campus, but rather inside building 3B. It would have made little or no sense for the robot to say that "the IT help desk is on the 1st floor in building 3B." To avoid confusion, an utterance like "the IT help desk is on the 1st floor" would be appropriate. Likewise, if the IT help desk happened to be located on another site of the university, the robot would have had to identify its location as being, e.g., "on the 1st floor in building 3B on the new campus". This illustrates that the hierarchical representation of space that humans adopt [Cohn and Hazarika, 2001] reflects upon the choice of an appropriate context when producing referential descriptions that involve attention-directing information.

Thus, the physical and spatial situatedness of the dialogue participants plays an important role when determining which related parts of space come into consideration as potential distractors. Another important observation concerns the verbal behavior of humans when talking about remote objects and places in a complex dialogue (i.e., more than just a question and a reply). E.g., consider the following dialogue:

Person A: "Where is the exit?"

Person B: "First go down this corridor. Then turn right. After a few steps you'll see the big glass doors."

Person A: "And the bus station? Is it to the left?"

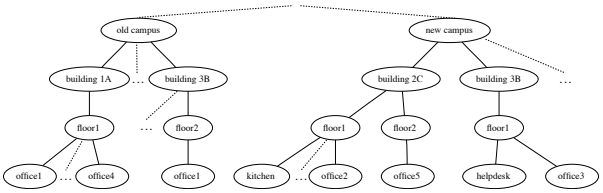
As can be seen, an utterance in such a collaborative dialogue is usually grounded in previously introduced discourse referents, both temporally and spatially. Initially, the physical surroundings of the dialogue partners form the context to which references are related. Then, as the dialogue unfolds, this point can conceptually move to other locations that have been explicitly introduced. Usually, a discourse marker denoting spatial or temporal cohesion (e.g., "then" or "there") establishes the last mentioned referent as the new anchor, creating a "mental tour" through large-scale space.

3.1 Context Determination Through Topological Abstraction

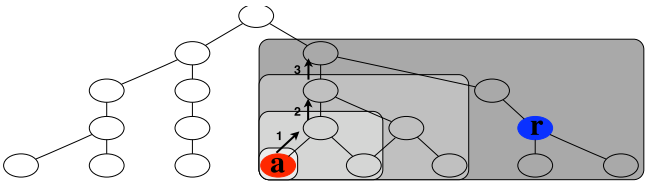
To keep track of the correct referential context in such a dialogue, we propose a general principle of *Topological Abstraction*[1] (TA) for context extension. TA is applied whenever a reference cannot be generated or resolved with respect to the current context. In such a case TA incrementally extends the context until the reference can be established. TA is designed to operate on a spatial abstraction hierarchy; i.e., a decomposition of space into parts that are related through a tree or lattice structure in which edges denote a containment relation (cf. Figure 2a). Originating in the *Referential Anchor* a, TA extends the context by incrementally ascending the spatial abstraction hierarchy until the intended referent is in the resulting sub-hierarchy (cf. Figure 2b). When no other information, e.g., from a preceding dialogue, is present, a is assumed to correspond to the spatio-visual context that is shared by the hearer and the speaker – usually their physical location and immediate surroundings. During a dialogue, however, a corresponds to the most salient discourse entity, reflecting how the *focus of attention* moves to different, even remote, places, as illustrated in the example dialogue above.

Below we describe two instantiations of the TA principle, a TA algorithm for reference generation (TAA1) and one for reference resolution (TAA2). They differ only minimally, namely in their use of an intended referent r or an RE $desc(x)$ to determine the conditions for entering and exiting the loop for topological abstraction. The way they determine a context through topological abstraction is identical.

Context Determination for GRE TAA1 (cf. Algorithm 1) constructs a set of entities dominated by the Referential Anchor a (including a itself). If this set contains the intended referent r, it is taken as the current utterance context set. Else TAA1 moves up one level of abstraction and adds the set of all child nodes to the context set. This loop continues until r is in the thus constructed set. At that point TAA1 stops and returns the constructed context set.



(a) Example for a hierarchical representation of space



(b) Illustration of the TA principle: starting from the Referential Anchor (a), the smallest sub-hierarchy containing both a and the intended referent (r) is formed incrementally

Figure 2: Topological Abstraction in a spatial hierarchy

Algorithm 1 TAA1 (for reference generation)

Require: a = referential anchor; r = intended referent
  Initialize context: $C = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  if $r \in C$ **then**
    return C
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ **do**
      for each $p \in topologicalParents(n)$ **do**
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      if $r \in C$ **then**
        return C
      end if
    end for
    return failure
  end if

TAA1 is formulated to be neutral to the kind of GRE algorithm that it is used for. It can be used with the original Incremental Algorithm [Dale and Reiter, 1995], augmented by a recursive call if a relation to another entity is selected as a discriminatory feature. It could in principle also be used with the standard approach to GRE involving relations [Dale and Haddock, 1991], but we agree with Paraboni et al. [2007] that the mutually qualified references that it can produce[2] are not easily resolvable if they pertain to circumstances where a confirmatory search is costly (such as in large-scale space). More recent approaches to avoiding infinite loops when using relations in GRE make use of a graph-based knowledge representation [Krahmer *et al.*, 2003; Croitoru and van Deemter, 2007]. TAA1 is compatible with these approaches, as well as with the salience based approach of Krahmer and Theune [2002].

[1] similar to Ancestral Search [Paraboni *et al.*, 2007]

[2] Stone and Webber [1998] present an approach that produces sentences like "take the rabbit from the hat" in a context with several hats and rabbits, but of which only one is in a hat. Humans find such REs natural and easy to resolve in visual scenes.

Algorithm 2 TAA2 (for reference resolution)

Require: a = ref. anchor; $desc(x)$ = description of referent
  Initialize context: $C = \{\}$
  Initialize possible referents: $R = \{\}$
  $C = C \cup topologicalChildren(a) \cup \{a\}$
  $R = desc(x) \cap C$
  if $R \neq \{\}$ **then**
    return R
  else
    Initialize: $SUPERNODES = \{a\}$
    for each $n \in SUPERNODES$ **do**
      for each $p \in topologicalParents(n)$ **do**
        $SUPERNODES = SUPERNODES \cup \{p\}$
        $C = C \cup topologicalChildren(p)$
      end for
      $R = desc(x) \cap C$
      if $R \neq \{\}$ **then**
        return R
      end if
    end for
    return failure
  end if

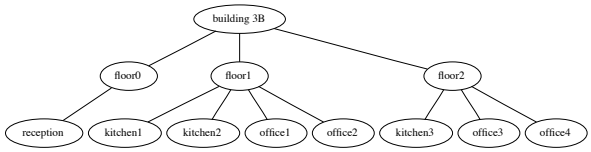
Context Determination for Reference Resolution A conversational robot must also be able to understand verbal descriptions by its users. In order to avoid overgenerating possible referents, we propose TAA2 (cf. Algorithm 2) which tries to select an appropriate referent from a relevant subset of the full KB. It is initialized with a given semantic representation of the referential expression, $desc(x)$, in a format compatible with the KB. We will show how this is accomplished in our framework in Section 4.1. Then, an appropriate entity satisfying this description is searched for in the KB. Similarly to TAA1, the description is first matched against the current *context set* C consisting of a and its child nodes. If this set does not contain any instances that match $desc(x)$, TAA2 enlarges the context set along the spatial abstraction axis until at least one possible referent can be identified within C.

4 Implementation

Our approach for resolving and generating spatial referring expressions has been fully integrated with the dialogue functionality in a cognitive system for a mobile robot [Zender *et al.*, 2008; Kruijff *et al.*, 2009]. The robot is endowed with a *conceptual spatial map* [Zender and Kruijff, 2007], which represents knowledge about places, objects and their relations in an OWL-DL[3] ontology. We use the Jena reasoning framework[4] with its built-in OWL reasoning and rule inference facilities. Internally, Jena stores the facts of the *conceptual map* as RDF[5] triples, which can be queried through SPARQL[6] queries. Figure 3 shows a subset of such a KB.

Below, we use this example scenario to illustrate our approach to generating and resolving spatial referring expressions in the robot's dialogue system. We assume that the interaction takes place at the reception on the ground floor ("floor0"), so that for TAA1 and TAA2 a =reception.

[3] http://www.w3.org/TR/owl-guide/
[4] http://jena.sourceforge.net
[5] http://www.w3.org/RDF
[6] http://www.w3.org/TR/rdf-sparql-query



(a) Topological abstraction hierarchy

```
(kitchen1 rdf:type Kitchen), (...)
(office1 rdf:type Office), (...)
(kitchen2 size big), (...)
(bob rdf:type Person), (bob name Bob),
(bob owns office1), (...)
(floor1 contains kitchen1), (...)
(floor2 contains office3), (...)
(floor1 ordNum 1), (floor2 ordNum 2), (...)
```
(b) RDF triples in the conceptual map (namespace URIs omitted)

Figure 3: Part of a representation of an office environment

4.1 The Comprehension Side

In situated dialogue processing, the robot needs to build up an interpretation for an utterance which is linked both to the dialogue context and to the (referenced) situated context. Here, we focus on the meaning representations.

We represent meaning as a logical form (LF) in a description logic [Blackburn, 2000]. An LF is a directed acyclic graph (DAG), with labeled edges, and nodes representing propositions. Each proposition has an ontological sort, and a unique index. We write the resulting ontologically sorted, relational structure as a conjunction of elementary predications (EPs): $@_{idx:sort}(\mathbf{prop})$ to represent a proposition **prop** with ontological sort $sort$ and index idx, $@_{idx1:sort1}\langle Rel\rangle(idx2 : srt2)$ to represent a relation Rel from index $idx1$ to index $idx2$, and $@_{idx:sort}\langle Feat\rangle(\mathbf{val})$ to represent a feature $Feat$ with value **val** at index idx. Representations are built compositionally, parsing the word lattices provided by speech recognition with a Combinatory Categorial Grammar [Lison and Kruijff, 2008]. Reversely, we use the same grammar to realize strings (cf. Section 4.2) from these meaning representations [White and Baldridge, 2003].

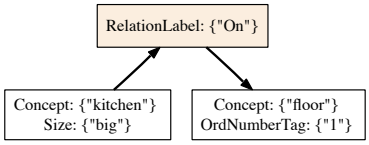
An example is the meaning we obtain for "the big kitchen on the first floor," (folding EPs under a single scope of @). It illustrates how each propositional meaning gets an index, similar to situation theory. "kitchen" gets one, and also modifiers like "big," "on" and "one." This enables us to single out every aspect for possible contextual reference (Figure 4a).

Next, we resolve contextual references, and determine the possible dialogue move(s) the utterance may express. Contextual reference resolution determines how we can relate the content in the utterance meaning, to the preceding dialogue context. If part of the meaning refers to previously mentioned content, we associate the identifiers of these content representations; else, we generate a new identifier. Consequently, each identifier is considered a dialogue referent.

Once we have a representation of utterance meaning in dialogue context, we build a further level of representation to facilitate connecting dialogue content with models of the robot's situation awareness. This next level of representation is essentially an a-modal abstraction over the linguistic aspects of meaning, to provide an a-modal conceptual structure

$@_{l1:e-place}(\mathbf{kitchen} \wedge$
$\langle Delimitation \rangle \mathbf{unique} \wedge$
$\langle Num \rangle \mathbf{sg} \wedge \langle Quantification \rangle \mathbf{specific} \wedge$
$\langle Modifier \rangle (b1 : q - size \wedge \mathbf{big}) \wedge$
$\langle Modifier \rangle (o1 : m - location \wedge \mathbf{on} \wedge$
$\langle Anchor \rangle (f1 : thing \wedge \mathbf{floor} \wedge$
$\langle Delimitation \rangle \mathbf{unique} \wedge$
$\langle Num \rangle \mathbf{sg} \wedge \langle Quantification \rangle \mathbf{specific} \wedge$
$\langle Modifier \rangle (n1 : number - ordinal \wedge \mathbf{1}))))$

(a) Logical form



(b) A-modal directed acyclic graph

```
SELECT ?x0 ?x1 WHERE {
  ?x0 rdf:type Kitchen.
  ?x0 size big.
  ?x1 rdf:type Floor.
  ?x1 ordNum 1.
  ?x0 containedIn ?x1.
```

(c) SPARQL query
In the previous example this would resolve `?x0` to `kitchen2`

Figure 4: Logical form, a-modal DAG and corresponding SPARQL query for "the big kitchen on the first floor"

[Jacobsson *et al.*, 2008]. Abstraction is a recursive translation of DAGs into DAGs, whereby the latter (conceptual) DAGs are typically flatter than the linguistic DAGs (Figure 4b).

The final step in resolving an RE is to construct a query to the robot's KB. In our implementation we construct a SPARQL query from the a-modal DAG representations (Figure 4c). This query corresponds to the logical description of the referent $desc(r)$ in TAA2. TAA2 then incrementally extends the context until at least one element of the result set of $desc(r)$ is contained within the context.

4.2 The Production Side

Production covers the entire path from handling dialogue goals to speech synthesis. The dialogue system can itself produce goals (e.g., to handle communicative phenomena like greetings), and it accepts goals from a higher level planner. Once there is a goal, an utterance content planner produces a content representation for achieving that goal, which the realizer then turns into one or more surface forms to be synthesized. Below we focus on utterance content planning.

A dialogue goal specifies a goal to be achieved, and any content that is associated with it. A typical example is to convey an answer to a user: the goal is to tell, the content is the answer. Content is given as a conceptual structure, *proto LF*, abstracting away from linguistic specifics, similar to the a-modal structures we produce for comprehension.

Content planning turns this proto LF into an LF which matches the specific linguistic structures defined in the grammar we use to realize it. "Turning into" means extending the proto LF with further semantic structure. This may be non-monotonic in that parts of the proto LF may be rewritten, expanding into locally connected graph structures.

Planning is agenda-based, and uses a planning domain defined as a (systemic) grammar network alike [Bateman, 1997; Kruijff, 2005]. A grammar network is a collection of systems that define possible sequences of operations to be performed on a node with characteristics matching the applicability conditions for the system. A system's decision tree determines which operations are to be applied. Decisions are typically context-sensitive, based on information about the shape of the (entire) LF, or on information in context models (dialogue or otherwise). While constructing an LF, the planner cycles over its nodes, and proposes new agenda items for nodes which have not yet been visited. An agenda item consists of the node, and a system which can be applied to that node.

A system can explicitly trigger the generation of an RE for the node on which it operates. It then provides the dialogue system with a request for an RE, with a pointer to the node in the (provided) LF. The dialogue system resolves this request by submitting it to GRE modules which have been registered with the system. (Registration allows us to plug-and-play with content-specific GRE algorithms.) Assuming a GRE module produces an LF with the content for the RE, the planner gets this LF and integrates it into the overall LF.

For example, say the robot in our previous example is to answer the question "Where is Bob?". We receive a communicative goal (see below) to inform the user, specifying the goal as an assertion related to the previous dialogue context as an answer. The content is specified as an ascription e of a property to a target entity. The target entity is t which is specified as a person called "Bob" already available in the dialogue context, and thus familiar to the hearer. The property is specified as topological inclusion (TopIn) within the entity k, the reference to which is to be produced by the GRE algorithm (hence the type "rfx" and the "RefIndex" which is the address of the entity).

$@_{d:dvp}(c - goal \wedge$
$\langle SpeechAct \rangle \mathbf{assertion} \wedge$
$\langle Relation \rangle \mathbf{answer} \wedge$
$\langle Content \rangle (e : ascription \wedge$
$\langle Target \rangle (t : person \wedge Bob \wedge$
$\langle InfoStatus \rangle \mathbf{familiar}) \wedge$
$\langle TopIn \rangle (p : rfx \wedge RefIndex)))$

The content planner makes a series of decisions about the type and structure of the utterance to be produced. As it is an assertion of a property ascription, it decides to plan a sentence in indicative mood and present tense with "be" as the main verb. The reference to the target entity makes up the copula restriction, and a reference to the ascribed property is in the copula scope. This yields an expansion of the goal content:

$@_{e:ascription}(\mathbf{be} \wedge$
$\langle Tense \rangle \mathbf{pres} \wedge$
$\langle Mood \rangle \mathbf{ind} \wedge$
$\langle Cop - Restr \rangle (t : entity \wedge$
$Bob \wedge \langle InfoStatus \rangle \mathbf{familiar}) \wedge$
$\langle Subject \rangle (t : entity) \wedge$
$\langle Cop - Scope \rangle (prop : m - location \wedge$
$\mathbf{in} \wedge \langle Anchor \rangle (p : rfx \wedge RefIndex)))$

The next step consists in calling the GRE algorithm to produce an RE for the entity p. In our NLP system we use a slightly modified implementation of the Incremental Algorithm [Dale and Reiter, 1995]. The context set C is determined using TAA1. Let's assume that Bob is currently in

kitchen3. In our example (a =reception) the GRE algorithm hence produces the following result, which is then returned to the planner and inserted into the proto LF created so far:

$$@_{p:entity}(\mathbf{kitchen} \wedge \langle TopOn\rangle(f : entity \wedge floor \wedge \langle Unique\rangle\mathbf{true} \wedge \langle Number\rangle(n : quality \wedge 2)))$$

The planner then makes further decisions about the realization, expanding this part of the LF to the following result:

$$@_{p:entity}(\mathbf{kitchen} \wedge \langle Delimitation\rangle\mathbf{unique} \wedge \langle Num\rangle\mathbf{sg} \wedge Quantification\rangle\mathbf{specific} \wedge \langle Modifier\rangle(o1 : m - location \wedge on \wedge \langle Anchor\rangle(f : thing \wedge floor \wedge \langle Delimitation\rangle\mathbf{unique} \wedge \langle Num\rangle\mathbf{sg} \wedge \langle Quantification\rangle\mathbf{specific} \wedge \langle Modifier\rangle(t1 : number - ordinal \wedge 2))))$$

Once the planner is finished, the resulting overall LF is provided to a CCG realizer [White and Baldridge, 2003], turning it into a surface form ("Bob is in the kitchen on the second floor"). This string is synthesized to speech using the MARY TTS software [Schröder and Trouvain, 2003].

5 Conclusions and Future Work

We have presented an algorithm for context determination that can be used both for resolving and generating referring expressions in a large-scale space domain. We have presented an implementation of this approach in a dialogue system for an autonomous mobile robot.

Since there exists no suitable evaluation benchmark for situated human-robot dialogue to compare our results against, we are currently planning a user study to evaluate the performance of the TA algorithm. Another important item for future work is the exact nature of the spatial progression in situated dialogue, modeled by "moving" the referential anchor.

References

[Bateman, 1997] J. A. Bateman. Enabling technology for multilingual natural language generation: the KPML development environment. *Journal of Natural Language Engineering*, 3(1):15–55, 1997.

[Blackburn, 2000] P. Blackburn. Representation, reasoning, and relational structures: a hybrid logic manifesto. *Journal of the Interest Group in Pure Logic*, 8(3):339–365, 2000.

[Cohn and Hazarika, 2001] A. G. Cohn and S. M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46:1–29, 2001.

[Croitoru and van Deemter, 2007] M. Croitoru and K. van Deemter. A conceptual graph approach to the generation of referring expressions. In *Proc. IJCAI-2007*, Hyderabad, India, 2007.

[Dale and Haddock, 1991] R. Dale and N. Haddock. Generating referring expressions involving relations. In *Proc. EACL-1991*, Berlin, Germany, April 1991.

[Dale and Reiter, 1995] R. Dale and E. Reiter. Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, 1995.

[Horacek, 1997] H. Horacek. An algorithm for generating referential descriptions with flexible interfaces. In *Proc. ACL/EACL-1997*, Madrid, Spain, 1997.

[Jacobsson *et al.*, 2008] H. Jacobsson, N. Hawes, G. J. Kruijff, and J. Wyatt. Crossmodal content binding in information-processing architectures. In *Proc. HRI-2008*, Amsterdam, The Netherlands, 2008.

[Kelleher and Kruijff, 2006] J. Kelleher and G. J. Kruijff. Incremental generation of spatial referring expressions in situated dialogue. In *In Proc. Coling-ACL-2006*, Sydney, Australia, 2006.

[Krahmer and Theune, 2002] E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R.Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*. CSLI Publications, Stanford, CA, USA, 2002.

[Krahmer *et al.*, 2003] E. Krahmer, S. van Erk, and A. Verleg. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1), 2003.

[Kruijff *et al.*, 2009] G. J. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, I. Kruijff-Korbayová, and N. Hawes. Situated dialogue processing for human-robot interaction. In H. I. Christensen, G. J. Kruijff, and J. Wyatt, editors, *Cognitive Systems*. Springer, 2009. to appear.

[Kruijff, 2005] G. J. Kruijff. Context-sensitive utterance planning for CCG. In *Proc. ENLG-2005*, Aberdeen, Scotland, 2005.

[Kuipers, 1977] B. Kuipers. *Representing Knowledge of Large-scale Space*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1977.

[Lison and Kruijff, 2008] P. Lison and G. J. Kruijff. Salience-driven contextual priming of speech recognition for human-robot interaction. In *ECAI 2008*, 2008.

[Paraboni *et al.*, 2007] I. Paraboni, K. van Deemter, and J. Masthoff. Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, 33(2):229–254, June 2007.

[Schröder and Trouvain, 2003] M. Schröder and J. Trouvain. The german text-to-speech synthesis system MARY: A tool for research, development and teaching. *Int. Journal of Speech Technology*, 6:365–377, 2003.

[Stone and Webber, 1998] M. Stone and B. Webber. Textual economy through close coupling of syntax and semantics. In *Proc. INLG-1998*, pages 178–187, Niagara-on-the-Lake, ON, Canada, 1998.

[White and Baldridge, 2003] M. White and J. Baldridge. Adapting chart realization to CCG. In *Proc. ENLG-2003*, Budapest, Hungary, 2003.

[Zender and Kruijff, 2007] H. Zender and G. J. Kruijff. Multi-layered conceptual spatial mapping for autonomous mobile robots. In *Control Mechanisms for Spatial Knowledge Processing in Cognitive / Intelligent Systems*, AAAI Spring Symposium 2007, March 2007.

[Zender *et al.*, 2008] H. Zender, O. Martínez Mozos, P. Jensfelt, G. J. Kruijff, and W. Burgard. Conceptual spatial representations for indoor mobile robots. *Robotics and Autonomous Systems*, 56(6):493–502, June 2008.

Verbalization of Vague Scalar Predicates Based on Autonomously Acquired Models (preliminary report)

Hendrik Zender and Andrzej Pronobis

Abstract— **The paper reports on ongoing research in generating and understanding verbal references to entities in the robot's environment. The paper focuses on features of spatial entities that are commonly expressed as vague scalar predicates in natural language, such as, e.g., size. The paper proposes an approach for characterizing such features in terms of properties and distributions over their values. This leads to a basic notion of prototypicality of property-values. Using this notion, the paper shows how different types of contextual standards can be defined, which determine the contextually appropriate use of a vague scalar predicate in linguistically describing a feature of a spatial entity. The approach goes beyond existing work in that it allows for a variety of contextual standards (in class, across classes, across instances) in describing features as vague scalar predicates, and by ultimately basing these standards in models of the robot's perceptual experience.**

I. Introduction

Robotic assistants are no longer simply a product of our imagination. Advances in related fields in robotics, AI and computer vision make it nowadays possible to develop highly autonomous robotic systems. Robots are already able to perform a wide variety of tasks while exhibiting a high degree of adaptivity to new environments.

When it comes to building "talking robots," though, we still face major challenges. These arise primarily from the variability and unpredictability of the kinds of dialogues they need to engage in. This sets talking robots apart from more standard dialogue systems, the development of which has matured over the last couple of years. Computer- or web-based conversational agents can usually talk about confined objective domains, like for example soccer, or pop star trivia [1]. Robots on the other hand have to talk about the environments they share with their users. And unlike location aware personal navigational assistants (e.g., [2]), which usually combine external localization (e.g., through RFID or GPS) with existing maps and information bases, these robots rely on autonomously built maps and knowledge acquired through their own interaction with their environment.

So how could a robot talk about what it knows? A robot's knowledge base contains facts about "things" in the world. These facts include properties of entities (e.g., color, size, age, name, etc.) and relations between entities (e.g., location, ownership). Some of this knowledge might be explicitly represented, some of it might be inferable when necessary, and some of it might only be implicitly accessible.

In this paper, we address the question of how a robot can verbalize what it knows about certain *properties* of things. We focus on those properties that correspond to a scalar feature space (e.g., size, length), and on their contextually appropriate verbalization. We address this issue from two viewpoints. For one, we propose a method for acquiring and evaluating categorical models for such properties in terms of a robot's perceptual capabilities. Secondly we investigate how these models can serve as a basis for situated natural language generation and comprehension.

The rest of the paper is organized as follows. Section II presents related work that serves as background for this report. In Section III we introduce the assumptions underlying our approach. In the subsequent section we then give the details on the acquired categorical models (Section IV) and on the verbalization methods (Section V). We conclude in Section VI.

II. Background

Given that a robot autonomously acquires knowledge about the world, how can such properties be appropriately verbalized for human-robot communication? Conversational robots and their users interact in *situated dialogues*. Such situated language typically involves things and persons in the environment or facts that are relevant for the current task at hand [3]. It relies on situationally and contextually appropriate expressions. In the case of a conversational robot these expressions are generated from the robot's own knowledge base. Such knowledge bases, often being hybrid models that contain symbolic and probabilistic layers [4] are not immediately useful for spoken human-robot interaction.

Moreover, both the robot's and the human's perceptual abilities are limited and may be subject to noise or incomplete observability. Therefore unnecessary precision in verbalizing known properties of things should be avoided. Instead of referring to "the 18m^2-large office," *vague* expressions involving gradable adjectives (such as "the largest office") should be used [5], [6], [7]. This research primarily focuses on instances – in a given visual setting. Our approach moves beyond this, by considering how scalar properties can be modeled as probabilistic distributions over their values – and then use these distributions to construct contextual standards. This makes it possible to consider distributions solely across observed instances (like [5]), and also across

H. Zender is with the Language Technology Lab, German Research Center for Artificial Intelligence (DFKI), Saabrücken, Germany, zender@dfki.de

A. Pronobis is with the Centre for Autonomous Systems, Royal Institute of Technology (KTH), Stockholm, Sweden, pronobis@csc.kth.se

instances within a class (considering values to be *prototypical* values within a class), and across classes. Within-class and across-class contextual standards are not considered (nor immediately possible) in [5]. They are, however, necessary to generate contextually appropriate verbalizations using comparatives. For example, consider the average office to have 8m^2. Talking about two offices, with *office1* measuring 12m^2 and *office2* 18m^2, it would be more appropriate to talk about *office1* as "the smaller office," not as "the small office." The reason being that it is still bigger than the average office. These ideas are based on insights in categorization and prototypicality originating with Brown [8] and Rosch [9].

The approach presented in this paper is consistent with the structure of the spatial knowledge representation developed in the CogX EU project [10]. The representation divides spatial knowledge into low level categorical models based on robot's sensory information, discrete, bottom-up models representing a map of the robot's environment as well as conceptual models attaching semantics to the low-level representations. Moreover, the method is compatible with the existing approaches to topological mapping (e.g. [4], [11]) and place categorization (e.g. [12], [13]).

Finally, the approach takes us beyond earlier methods for relating language to "the world." For example, Steels et al (cf. e.g. [14]) propose *semiotic networks*. A semiotic network is an associative network in which a perceptual layer is connected with a category layer, which in turn associates categories with words. A robot is able to acquire such networks online, establishing nodes within layers, and associations within and across layers. The resulting network captures how words can be related to categories – but does so in an absolute way. We take this a step further. Predications ("words" in a semiotic network) are not determined directly through association with a category. We perform an intermediary interpretation step, which establishes how best to express a property (a "category") within a given context. Such expression may vary, obviously; but it is a variation not possible on the current formulation of semiotic networks. (Similar observations can be made for approaches proposed by Roy et al.)

III. The Approach

In our approach, we assume that the robot is able to perceive the world through its sensors and internally grounds its spatial knowledge upon *features* extracted from the sensory input. Moreover, we distinguish between three separate layers of the knowledge representation: the *spatial layer*, the *categorical layer*, and the *linguistic layer*. The spatial layer represents the knowledge about the environment in which the robot operates, i.e., its world map. We assume that the environment, can be segmented into *areas*, where each area corresponds to a single spatial unit of certain semantics, e.g., a room. As a result, the map stored in the spatial layer consists of a finite number of models, each representing a single area in terms of the observed feature values. This is consistent with the mapping framework presented in [4], which builds on the concepts of discrete places and scenes expressed in terms of arbitrary, possibly complex features and local spatial relations.

The categorical layer contains categorical models grouping the robot's sensory information expressed in terms of feature values. The knowledge represented in this layer is not specific to any particular location in the robot's environment. Instead, it represents a general knowledge about the world at the sensory level. The categorical models stored in this layer give rise to certain *properties* of the spatial units (areas). These properties can either be continuous or discrete and usually correspond to human concepts, such as size, shape or type of a room. The categorical layer could practically be implemented by using a set of classifier models trained in a supervised manner as in case of the existing place categorization approaches [12], [13].

The linguistic layer, finally, contains specialized algorithms for turning such properties into logical *predicates*, which then are realized as natural language expressions. It is important to note that the categorical models do not maintain linguistic labels for properties of the spatial units. The linguistic layer maintains *interpretation functions*, which query the categorical models for relevant information whenever needed. This explicit interpretation step guarantees the formation of contextually appropriate semantic *predicates* for verbalization.

In the following, we present concise definitions of the terms used. A *feature* f_i is a function that provides a potentially complex interpretation of robot's sensor input. The role of features is to provide a new representation of the sensory input that is less sensitive to noise and usually more compact. The quality of a given feature observation is a *feature value*, which is positioned with respect to one or multiple *dimensions* $\mathcal{F}_i \in \mathbb{R}^n$. Thus, each function f_i maps sensory observations $\mathcal{S}$ onto a certain range $\mathcal{F}_i$:

$$f_i : \mathcal{S} \rightarrow \mathcal{F}_i \tag{1}$$

The values of all features correspond to a single point in the so-called *feature space* $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \times \mathcal{F}_\mathcal{N}$.

A *property* p_i of an entity $x \in \mathcal{U}$ is a function that maps entities onto a certain property value $v \in \mathcal{P}_i$, e.g., the property of having a given size, or being of a given type.

$$p_i : \mathcal{U} \rightarrow \mathcal{P}_i \tag{2}$$

$$size : \mathcal{U} \rightarrow \mathbb{R}^+ \tag{3}$$

$$type : \mathcal{U} \rightarrow \{\mathsf{Room}, \mathsf{Office}, \mathsf{Kitchen}, \mathsf{Corridor}, \ldots\} \tag{4}$$

If there additionally is a (partial) order specified over the values of $\mathcal{P}_i$, we speak of *scalar* properties. Entities can hence be sorted in a partial order with respect to such a scale.

(1) $size(\mathsf{office1}) = 12m^2$
$size(\mathsf{office2}) = 18m^2$
$size(\mathsf{office3}) = 7m^2$

(2) $\mathsf{office3} \leq_{size} \mathsf{office1} \leq_{size} \mathsf{office2}$

In our approach, the properties of entities are determined and updated on the basis of spatio-temporally integrated feature value observations. Due to the uncertainty and noise involved in robotic perception, our approach represents properties as probability distributions over a range of values instead of a crisp numerical value, cf. Section IV. The scalar order of entities is determined dynamically when needed, cf. Section V.

Logical *predicates* express propositions about entities. Such propositions can be either true or false or non-sensical. Predicates are the basis for forming semantic representations that can then be realized as natural language descriptions. The following list contains examples for predicates along with natural language paraphrases:

(3) $\mathrm{Size}(\mathsf{office1}, 12m^2)$: 'The size of office1 is $12\mathrm{m}^2$.'

(4) $\mathrm{Type}(\mathsf{fido}, \mathsf{Bird})$: 'Fido is a bird.'

(5) $\mathrm{Small}(\mathsf{fido})$: 'Fido is small.'

(6) $\mathrm{Bigger}(\mathsf{fido}, \mathsf{dido})$: 'Fido is bigger than Dido.'

There is a close link between properties and predicates. The important distinction is that properties encode the agent's knowledge about entities in the world, while predicates express propositions about entities. The truth value of a proposition is determined by evaluating an *interpretation function* against a given knowledge base. In this paper, we want to focus on *vague scalar predicates* that correspond to *gradable adjectives*, like the ones in Examples 5 and 6. When verbalizing the robot's knowledge, only true propositions are to be considered. The truth value of a vague predicate cannot be established in absolute terms. Their meaning is highly context-dependent [6], and moreover needs to take into account a given *standard* [5]. A vague predicate must hence be interpreted against a precise property with respect to a given *context* c.

Following Kennedy, a predicate φ that corresponds to a gradable adjective "can then be analyzed as a function that induces a tripartite partitioning of its (ordered) domain into: (i) a positive extension $pos_c(\varphi)$, which contains objects above some point in the ordering (...), (ii) a negative extension $neg_c(\varphi)$, which contains objects below some point in the ordering (...), and (iii) an 'extension gap' $gap_c(\varphi)$, which contains objects that fall within an indeterminate middle (...)" [15].

$$[\![\varphi(x)]\!]_c = \begin{cases} T & \text{iff} \quad x \in pos_c(\varphi) \\ F & \text{iff} \quad x \in neg_c(\varphi) \\ undef & \text{iff} \quad x \in gap_c(\varphi) \end{cases}$$

[15] further discusses the partitioning of the domain into $pos_c(\varphi)$, $neg_c(\varphi)$ and $gap_c(\varphi)$ with respect to a comparison class. We will however focus on another aspect, which is left out in [15], namely the context-dependent transformation from precise properties to vague predicates, and the determination of a comparison class.

As explained earlier, a scalar property is a derivation from one or many feature observations. Several vague predicates can correspond to different dimensions in the feature space, such as, e.g., "large" and "small", which correspond to the opposite poles of the size scale. The challenge lies thus in defining appropriate functions $pos_c(\varphi)$, $neg_c(\varphi)$ and $gap_c(\varphi)$ for constructing the positive, negative, and indeterminate extensions respectively of a predicate φ from an underlying property p.

We will proceed with illustrating how autonomously acquired categorical models represent features; how these models are used in the task of determining properties of entities in the world; and finally how a *context dependent* interpretation of these properties yields a situationally appropriate vague expression to describe an entity.

IV. The Spatial Categorical Models

As already mentioned, in our approach, the representation of spatial knowledge is divided into three separate layers. Here, we focus on the spatial and categorical layers. The spatial layer maintains models representing areas in the environment in terms of the observed feature values. The dependency between the area and the feature values likely to be observed in that area is captured by a probability distribution $p_{A,F}(a, f)$, where the random variable A represents an area, and the random variable F the values of all features. Similarly, the categorical layer encodes the dependency between the observed values of area properties and the observed values of features as a distribution $p_{P,F}(p, f)$. In the rest of this section, we show how the knowledge represented in the spatial and categorical layers is acquired and used to perform inferences about the properties of areas in the environment.

A. Model Acquisition

Despite the fact that the models stored in both the spatial and categorical layers represent the knowledge in terms of values of features extracted from the sensory input, they differ in the way that knowledge is acquired. The models representing the instances of areas that the robot visited, stored in the spatial layer, are built as the robot explores the environment, in an unsupervised fashion. The continuous space is segmented and models are build separately for each of the segments. On the other hand, the categorical models need to encode human concepts corresponding to the properties of areas. These concepts must be transferred to the robot during a supervised learning stage. A learning algorithm is employed and provided with sets of training sensory data annotated by a human with ground truth indicating the values of properties of areas where the data were acquired. The training data are acquired while the robot is guided by a user through several different environments. Moreover, one of the available robotic databases, such as [16], can be used to provide the initial spatial knowledge for training. The task of the algorithm is to build universal models encoding the correspondence between values of area properties and values of features extracted from the sensory data acquired in the area. At the same time, during the guided tour, the robot builds models of the areas stored within the spatial layer.

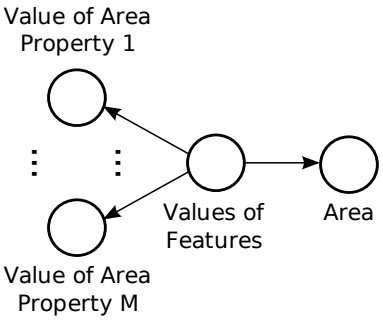


Fig. 1: Graphical model representing the dependencies between values of features, areas and values of area properties.

B. Inferences About Properties of Areas

In order to generate referring expressions, we need to integrate the knowledge represented in the spatial and categorical layers and obtain dependencies between areas and values of properties for those areas. Let's take the example of a size of an area. In such case, the categorical models are trained using sensory data annotated with the correct size of the areas where the data were acquired. As a result, a model is be build that maps the features extracted from the sensory data into the values of the size property. Below, we show how those models could be integrated with the models of the visited areas.

We assume the that the dependency between observed values of features, areas and values of area properties can be expressed through the graphical model presented in Figure 1. The model allows to express the joint probability distribution $p(f, a, p_1, \ldots, p_M)$ as follows:

$$\begin{aligned} p(f, a, p_1, \ldots, p_M) &= p(p_1|f) \cdot \ldots \cdot p(p_M|f) \cdot \\ & \quad p(a|f) \cdot p(f) \end{aligned} \tag{5}$$

In other words, random variables representing the area and the area properties are independent given the values of all features extracted from the sensory input. Such formulation, allows to make inferences about properties of areas by integrating the models of properties in the categorical layer with the models of areas in the spatial layer. For instance, in order to obtain a distribution over the values of a property p_i for an area a, we use the formula given below:

$$\begin{aligned} p(p_i|a) &= \int_{-\infty}^{\infty} p(f, p_i|a)\ df \\ &= \int_{-\infty}^{\infty} p(p_i|f)\ p(f|a)\ df \end{aligned} \tag{6}$$

Similarly, to obtain a distribution over areas described by a certain value of a property p_i, we calculate

$$\begin{aligned} p(a|p_i) &= \int_{-\infty}^{\infty} p(f, a|p_i)\ df \\ &= \int_{-\infty}^{\infty} p(f|p_i)\ p(a|f)\ df \end{aligned} \tag{7}$$

In practice, the integration can be performed as the robot explores the environment and builds the models of areas. In some cases, the models of areas might be based on different features than those required to perform the integration. In such cases, the integration is performed with the features calculated from immediate observations only for the area where the robot is currently located.

V. Verbalization

In this section we are addressing the question how an autonomous robot can verbalize what it knows about its environment such that it can be correctly understood by human hearers. Here, we especially focus on verbalizing spatial properties of areas in the known environment. Consider the following set of examples (with an emphasis on the italicized parts), which illustrate two separate but related phenomena. In Example 7 the task is to produce a unique Referring Expression, while Example 8 concerns categorical descriptions.

(7) "The toolbox is in *the largest office*."
"The kitchen is *the smallest room* on the third floor."

(8) "Restrooms are *small rooms*."
"Meeting rooms are usually *larger than offices*."

Although the focus here is on verbalization, it is important to note that this work is embedded in a larger dialogue system for situated spoken human-robot interaction [17]. Our approach is thus in principle bi-directional (like [18], [19]). The same mechanisms can be used both for natural language generation and comprehension. While this is true and conceivable for generating and resolving referring expressions, a "real" understanding of human descriptions and explanations would require feeding back the conveyed facts into the robot's categorical models. This is still an open issue and remains as a possibility for future work.

A. Referring Expressions

A Referring Expression (RE) is a (complex) noun phrase (NP) that contains enough information to identify an *intended referent*, while avoiding ambiguities with *potential distractors* [20]. [19] illustrates how an appropriate context for references in *large-scale space* can be constructed, and how referring expressions to entities (including areas) in large-scale space can be generated by expressing spatial relations between such entities.

Here, we assume a similar structure of the spatial knowledge base in order to construct an appropriate context. We furthermore make use of the same GRE algorithm in order to narrow down the distractor set through the inclusion of navigational information in complex referring expressions. The contribution of this work lies in allowing a GRE (and also RRE) algorithm to include properties, such as, e.g., size, to further qualify the intended referent by ruling out remaining distractors.

Although, strictly speaking, such properties could be verbalized by expressing a numerical value (e.g., "the room with size 18m^2" or "the 12m^2 large office"), such a level of detail is commonly avoided in human-human dialogues. Not only is such specific information not easy to verify with human

perceptual capabilities, it might as well be hard to determine given a robot's knowledge base. As discussed previously (Section IV), the robot's representation of a property usually is not a crisp number, but rather a probability distribution over a range of values.

Previous existing work propose the use of *vague scalar predicates* as "qualitative linguistic expressions of quantitative information" [5]. Size is widely used as an example for a property that needs to be expressed by vague predicates, namely the adjectives "small" and "large", as well as their comparative (i.e., "smaller", "larger") and superlative (i.e., "smallest", "largest") forms [6].

B. Vague Scalar Predicates

As said previously, one way of avoiding unnecessary precision in verbal interaction is to make use of vague expressions. One intricacy with vague expressions is that they are highly context dependent. For instance, what is called a "large restroom" in one situation could be appropriately described as "the smallest room on the second floor" under different circumstances.[1] A crucial notion is the scale of the underlying property. The applicability of different vague predicates corresponding to that property is then determined on the basis of the referent's position on that scale in comparison to where the distractors are on that scale. A simple ordering of the referent and its distractors on such a scale can give rise to a predication that consists of a superlative form.

$$r <_{size} distractor{-}set \rightarrow \mathrm{Smallest}(r) \tag{8}$$

$$r >_{size} distractor{-}set \rightarrow \mathrm{Largest}(r) \tag{9}$$

A stronger proposition is expressed by the corresponding positive forms. The applicability such predicates is furthermore determined by the position of the referent on that scale with respect to a given standard [5], [15].

$$r \ll_{size} standard \rightarrow \mathrm{Small}(r) \tag{10}$$

$$r \gg_{size} standard \rightarrow \mathrm{Large}(r) \tag{11}$$

In the following we'll explain how these conditionals can be established in our probabilistic approach.

C. Evaluating Against a Set of Distractors

A typical example could be finding places that correspond to "the largest kitchen". In such case, we could define a new random variable

$$L = \begin{cases} 1 & - & \text{the value of the size property for an area} \\ & & \text{is larger than the value of the size property} \\ & & \text{for other areas being kitchens} \\ 0 & - & \text{otherwise} \end{cases}$$

[1]The task of re-identifying "the large restroom" later, however, is a task of correctly retrieving the previous mention and its referent from dialogue history or from a long-term memory. This issue is beyond the scope of this paper.

Then, similarly as for properties, we can calculate $p_{L|A}(l|a)$ as follows:

$$\begin{aligned} p_{L|A}(1|a) &= \int_0^\infty p_{P_S|A}(s_1|a) \\ &\quad \int_0^{s_1} \sum_{a_2} p_{P_S,A|P_T}(s_2, a_2|\text{kitchen})\ ds_2\ ds_1 \\ p_{L|A}(0|a) &= 1 - p_{L|A}(1|a) \end{aligned} \tag{12}$$

where P_S is a random variable representing the size property and P_T is a random variable representing the type property. The probability distribution $p_{P_S,A|P_T}(s_2, a_2|\text{kitchen})$ can be factored as follows:

$$\begin{aligned} p_{P_S,A|P_T}(s_2, a_2|\text{kitchen}) &\propto p_{P_S,P_T|A}(s_2, \text{kitchen}|a_2) \cdot \\ &\quad p_A(a_2) \end{aligned} \tag{13}$$

and the prior $p_A(a_2)$ can be used to specify the context, i.e. which other areas should be taken into consideration, e.g. all areas except the area a.

Now, in order to obtain $p_{A|L,P_T}(a|1, \text{kitchen})$, we use the Bayes rule:

$$\begin{aligned} p_{A|L,P_T}(a|1, \text{kitchen}) &\propto p_{L|A,P_T}(1|a, \text{kitchen}) \cdot \\ &\quad p_{P_T|A}(\text{kitchen}|a). \end{aligned} \tag{14}$$

The probability $p_{L|A,P_T}(1|a, \text{kitchen})$ can be calculated as shown in Eq. 12, by replacing the factor $p_{P_S|A}(s_1|a)$ with $p_{P_S|A,P_T}(s_1|a, \text{kitchen})$.

D. Evaluating Against a Standard

A purely contextual standard (as in [5]) might be appropriate for abstract entities for which there is neither an objective standard nor a standard based on prior experience. For spatial entities, however, people have expectations and standards based on their previous knowledge. Still there exists a contextual bias to which people can adapt their expectations under varying circumstances.

We propose a standard that takes into account both the given context and abstract world knowledge. We establish an evaluation standard on the basis of a prototypical property value. The prototypical quality is determined by averaging across the property values of a given set of entities. We hence propose an extensional definition of prototypicality.

An a priori prototypical standard can be computed after the dedicated learning step (cf. Section IV). Later on, this standard is modified by also taking into account all new instances of a given class. This allows our system to gradually adjust its standard to the environment in which it operates. By this *self-extension* we ensure the production of contextually appropriate expressions.

As shown in the previous section, we can define a new random variable

$$L = \begin{cases} 1 & - & \text{the value of the size property for an area} \\ & & \text{is larger than the expected value of the size} \\ & & \text{property for areas being kitchens} \\ 0 & - & \text{otherwise} \end{cases}$$

Then, again, we can calculate $p_{L|A}(l|a)$ as follows:

$$\begin{aligned} p_{L|A}(1|a) &= \int_{E[P_S|P_T=\text{kitchen}]}^{\infty} p_{P_S|A}(s_1|a) \\ p_{L|A}(0|a) &= 1 - p_{L|A}(1|a) \end{aligned}$$

These equations provide an interpretation for the applicability of the stronger positive predications. The expression "the small kitchen" is hence only generated if the referent is not only the smallest kitchen in the context, but also if it generally belongs to the class of small kitchens as defined by being considerably smaller than a prototypically-sized kitchen.

VI. Conclusions and Future Work

This work presents a robot that autonomously acquires a notion of prototypical appearance and intra-class variation. It starts with a model that is learned off-line. It can use this model right away for verbalizing descriptions and referring expressions. It also constantly extends its models and its instance knowledge and thus shapes its verbalization to better reflect its experience in the operating environment.

Acknowledgments

This work was supported by the EU FP7 ICT Project "CogX" (FP7-ICT-215181).

References

[1] F. Xu, P. Adolphs, H. Uszkoreit, X. Cheng, and H. Li, "Gossip galore: A conversational web agent for collecting and sharing pop trivia," in *Proceedings of ICAART 2009 – First International Conference on Agents and Artificial Intelligence*, J. Filipe, A. Fred, and B. Sharp, Eds., Porto, Portugal, January 2009.

[2] I. Aslan, M. Schwalm, J. Baus, A. Krüger, and T. Schwartz, "Acquisition of spatial knowledge in location aware mobile pedestrian navigation systems," in *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, Helsinki, Finland, 2006, pp. 105–108.

[3] D. K. Byron, "Understanding referring expressions in situated language – some challenges for real-world agents," in *Proceedings of the First International Workshop on Language Understanding and Agents for the Real World*, Hokkaido University, Japan, 2003, pp. 80–87.

[4] A. Pronobis, K. Sjöö, A. Aydemir, A. N. Bishop, and P. Jensfelt, "A framework for robust cognitive spatial mapping," in *Proceedings of the 14th International Conference on Advanced Robotics (ICAR 2009)*, Munich, Germany, June 2009.

[5] D. DeVault and M. Stone, "Interpreting vague utterances in context," in *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, August 2004.

[6] K. van Deemter, "Generating referring expressions that involve gradable properties," *Computational Linguistics*, vol. 32, no. 2, 2006.

[7] C. Kennedy, "Vagueness and grammar: The semantics of relative and absolute gradable adjectives," *Linguistics and Philosophy*, vol. 30, no. 1, pp. 1–45, February 2007.

[8] R. Brown, "How shall a thing be called?" *Psychological Review*, vol. 65, no. 1, pp. 14–21, 1958.

[9] E. Rosch, "Principles of categorization," in *Cognition and Categorization*, E. Rosch and B. Lloyd, Eds. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1978, pp. 27–48.

[10] "Semantic cognitive spatial representation," documents reporting ongoing research.

[11] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, June 2008.

[12] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research, Special Issue on Robot Vision*, 2009, (Submitted).

[13] O. M. Mozos, R. Triebel, P. Jensfelt, A. Rottmann, and W. Burgard, "Supervised semantic labeling of places using information extracted from sensor data," *Robotics and Autonomous Systems*, vol. 55, no. 5, 2007.

[14] L. Steels, "Semiotic dynamics for embodied agents," *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 32–38, 2006.

[15] C. Kennedy, "Gradable adjectives denote measure functions, not partial functions," *Studies in the Linguistic Sciences*, vol. 29, no. 1, pp. 65–80, 1999.

[16] A. Pronobis and B. Caputo, "COLD: CoSy Localization Database," *The International Journal of Robotics Research*, vol. 5, no. 28, May 2009.

[17] G. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, I. Kruijff-Korbayová, and N. Hawes, "Situated dialogue processing for human-robot interaction," in *Cognitive Systems*, H. Christensen, G. Kruijff, and J. Wyatt, Eds. Springer, 2009, to appear.

[18] H. Zender, G.-J. M. Kruijff, and I. Kruijff-Korbayová, "A situated context model for resolution and generation of referring expressions," in *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*. Athens, Greece: Association for Computational Linguistics, March 2009, pp. 126–129. [Online]. Available: http://www.aclweb.org/anthology/W09-0622

[19] ——, "Situated resolution and generation of spatial referring expressions for robotic assistants," in *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, CA, USA, July 2009, pp. 1604–1609.

[20] R. Dale and E. Reiter, "Computational interpretations of the Gricean Maxims in the generation of referring expressions," *Cognitive Science*, vol. 19, no. 2, pp. 233–263, 1995. [Online]. Available: citeseer.ist.psu.edu/dale94computational.html

Verbalization of Ontological Knowledge for Communication about Properties and Gaps (preliminary report)

Hendrik Zender and Geert-Jan M. Kruijff

Abstract— **The paper reports preliminary research on verbalizing a robot's knowledge about an instance I of a particular concept C. This covers both what a robot knows, and what it does not (yet) know about the instance. The paper considers a "gap" to be that information the robot misses to establish a given property P for I, knowing that that property typically applies to instances of C. The paper proposes a method for determining which properties are classifiable as gaps for an instance relative to a concept. This method operates on the TBox and ABox of an ontology. It provides a general method for determining gaps, and is not specific to situated dialogue. The paper shows how the resulting characterization of available and missing knowledge about I relative to C can then be verbalized, following up an approach recently presented in [1]. The paper illustrates the method on an example involving spatial entities, and discusses further research on extending the method.**

I. Introduction

A robot typically does not come equipped with all there is to know about the world it operates in. More often than not, it is uncertain about what it perceives, or how to understand what it perceives. It needs to learn more. Which is exactly the point of the argument for continuous learning.

But there is more to this. A passive strategy, waiting until suitable learning examples present themselves *deus ex machina*, is unlikely to help the robot much. The world presents a robot with a wealth of perceptual information – and at the same time, each perception is unique. Experience is sparse, making it necessary for the robot to employ an active strategy in learning. It needs to figure out what it needs to learn. Which is why a robot needs introspection.

On introspection, the robot determines for a particular context what it knows, and what it doesn't know. Intuitively, "what it doesn't know" represents a (potential) gap in the robot's knowledge. The basic idea in the CogX project is to use such gaps to drive further learning. The robot can actively engage with the environment, or enter into dialogue with other agents, to fill in its gaps. Which raises the question how a robot could identify what it doesn't know.

In this paper, we present a method for determining a specific type of gaps. We are interested in the following question: Given an individual I, and a concept C in an ontology, what information about I is lacking to establish it as an instance of C? We assume that we can express "information about I" as a list of properties. Lacking information then is a list of those properties $P^?$ that together with a list of known properties P^+ for I, would provide the necessary and sufficient facts to establish I as an instance of C.

H. Zender and G.J. Kruijff are with the Language Technology Lab, German Research Center for Artificial Intelligence (DFKI), Saabrücken, Germany {zender, gj}@dfki.de

The basic idea behind the method is to use query mechanisms for checking whether an individual I fulfills the definitions for being an instance of a given concept. This basic idea is similar to the type of slot-filling we find in information state-based dialogue management [2]. Where the methods diverge is that the method we discuss here provides a natural way for integrating with uncertainty in (perceptual) categorization. Given a distribution over possible concepts $C_1, ..., C_n$, the method can query the ontology for fulfillment of I against each of these concepts. The result is a set of property lists $P_i^?$, one for each concept C_i. Which provides us with the possibility to do more than just slot-filling. In a manner akin to generating or resolving referring expressions [3], [4], we can determine what minimal set of properties P^δwould provide a way for optimally dividing $C_1, ..., C_n$ into two subsets of concept descriptions that are mutually exclusive along P^δ. This is a strategy we use in meta-learning to actively select samples to divide a category search space [5]. In our setting it provides the possibility to select a "gap" the answer to which can help to efficiently narrow down the scope of possible concepts for I.

We return to this idea at the end of the paper. What interests us here is the use of gaps $P^?$ and a description of the concept to verbalize the robot's knowledge about I. Such verbalization plays a fundamental role in situated dialogue between a human and a robot that wants to learn more. By telling the human what it does and doesn't know, the robot makes its internal belief state transparent to the human. And this sets up the background for another important aspect of dialogue for learning, namely scaffolding [6]. By telling the human what it doesn't know, the robot indicates what it *would like* to know.

An overview of the paper is as follows. Section II discusses Schütte's algorithm for verbalizing conceptual structures. The verbalization method we use in this paper is based on that algorithm. Furthermore, we briefly outline the scenario in which we work. Section III describes description logics, and their use in formulating ontologies. Section IV explains the different aspects of our proposed approach in terms of verbalization, introspection, and knowledge gap generation. We close the paper with a discussion of future research in Section V.

II. Background

Schütte [1] presents an approach for verbalizing concept descriptions from ontologies. In this work, we generate concept and individual descriptions and use the same mechanisms to identify such missing pieces of information that can be discovered by an agent's knowledge gathering behaviors. We will call these pieces *knowledge gaps*. The domain we are interested in in this work is a spatial knowledge base for an autonomous robot, more specifically the robot's *conceptual spatial map* of its environment, cf. [7]. The verbalization mechanisms are part of a dialogue system for such a robot [8].

III. Ontology-Based Knowledge Representation

Description Logics (DL) based ontologies make a distinction between a conceptual *taxonomy* of concepts (the *TBox* $\mathcal{T}$, for terminological knowledge) and the knowledge about individuals in the domain of discourse (the *ABox* $\mathcal{A}$, for assertional knowledge) [9]. Additionally DL-ontologies contain a set of roles that can hold between individuals, and which are defined over concepts. While some call this the *RBox*, we will assume that role definitions and role restrictions that are used in concept definitions belong to the TBox. Those roles that represent relations between individuals are part of the ABox. Another common name for concept is class. This gives rise to a more extensional perspective – in which a concept can be represented as the set of its member individuals.

An important distinction which we will later get back to is the distinction between *atomic concepts* and *concept descriptions* [10]. Atomic concepts can be *defined* in terms of complex concepts, which are expressed by other concepts, concept constructors and role restrictions, cf. [10], [11] for a more complete account. Here is an example of such a *concept definition* in our robotic spatial map domain (TBox $\mathcal{T}_r$, cf. Figure 1), which defines kitchens as all those rooms that contain at least one kitchen object:

(1) $\mathsf{Kitchen} \equiv \mathsf{Room} \sqcap \exists \mathsf{hasObject}.\mathsf{KitchenObject} \in \mathcal{T}_r$

The task of DL reasoners is to perform certain kinds of inferences in both the TBox and the ABox. The most basic TBox inference – and the one that is relevant for this work – is *subsumption* checking between concepts. This inference turns a set of concept definitions into a hierarchical taxonomy in which concepts are related with a subclass/superclass relation. Given the example above, a DL reasoner could infer that Kitchen is a subclass of Room.

(2) $\models_{\mathcal{T}_r} \mathsf{Kitchen} \sqsubseteq \mathsf{Room}$

In the ABox a DL reasoner establishes class membership of individuals, the so-called *instance checking* mechanism [10]. Continuing our example, we could assert the following facts about our domain:

(3) The example ABox $\mathcal{A}_{ex}$:
$\mathsf{Room}(\mathsf{AREA1})$

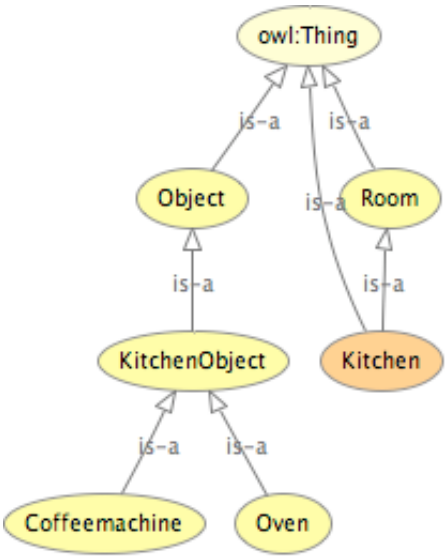


Fig. 1: Visualization of the named-class hierarchy of the example TBox $\mathcal{T}_r$. `owl:Thing` is the OWL equivalent of the top level concept $\top$ in abstract DL formalisms.

$\mathsf{Oven}(\mathsf{OBJ1})$
$\mathsf{hasObject}(\mathsf{AREA1}, \mathsf{OBJ1})$

The reasoner would then infer that OBJ1 is also an instance of KitchenObject and hence AREA1 is an instance of Kitchen[1]:

(4) $\mathcal{A}_{ex} \models_{\mathcal{T}_r} \mathsf{KitchenObject}(\mathsf{OBJ1})$
$\mathcal{A}_{ex} \models_{\mathcal{T}_r} \mathsf{Kitchen}(\mathsf{AREA1})$

At the core of our system is an OWL-DL ontology[2] that represents a robotic *conceptual spatial map* [12], which represents knowledge about places and objects in the environment, as well as relations between them. For the present work, we are using the "Jena" reasoning framework[3] with its built-in OWL reasoning and rule inference facilities. Internally, Jena stores the facts of the ABox and the TBox of the ontology reasoner as RDF[4] triples. The knowledge base can be queried through SPARQL[5] queries. We will later use the abstract DL syntax (see above) and the concrete OWL/RDF syntax interchangeably wherever one is more appropriate.

IV. Verbalization and Introspection

In terms of verbalization and knowledge introspection, DL-based ontologies afford a number of interesting tasks; verbalizing conceptual knowledge, i.e., turning TBox definitions into natural language descriptions [1] being one. Another opportunity for verbalization is to talk about individuals

[1] Of course, an OWL-DL reasoner would establish the full type hierarchy for both individuals along the transitive subsumption axis (cf. Figure 1. This is left implicit here for ease of reading.

[2] `http://www.w3.org/TR/owl-guide/`

[3] `http://jena.sourceforge.net`

[4] `http://www.w3.org/RDF`

[5] `http://www.w3.org/TR/rdf-sparql-query`

in a knowledge base, e.g., generating referring expressions to ABox instances [3]. A task which is closely related to the aforementioned tasks is knowledge introspection, put differently, determining *gaps* in an agent's knowledge.

A. Verbalizing Ontological Knowledge

Ontologies encode knowledge of concepts and their relationships in a specific domain. They are designed to support different inference mechanisms and in order to describe intensional and extensional knowledge about the involved concepts. Ontologies thus contain many concepts that don't have a clear one-to-one correspondence with a lexical item. They hence typically don't straightforwardly afford generating natural language descriptions of concepts and instances. Following Schütte [1], we annotate those concepts in the TBox that correspond to words in natural language.

(5) Definition of the AnnotationProperty lexicalWord, and annotation of the concept Kitchen in $\mathcal{T}_r$:

```
<owl:DatatypeProperty rdf:ID="lexicalWord">
  <rdf:type rdf:resource=
    "&owl;AnnotationProperty"/>
  <rdfs:range rdf:resource="&xsd;string"/>
</owl:DatatypeProperty>

<owl:Class rdf:ID="Kitchen">
  <rdfs:subClassOf rdf:resource="#Room"/>
  <lexicalWord rdf:datatype="&xsd;string">
    @p:entity(kitchen)
  </lexicalWord>
</owl:Class>
```

His approach relies on the distinction between atomic concepts and concept definitions, cf. Section III, which is reflected in the choice of the subject and the subject complement in the introductory sentence.

(6) "A kitchen is a room that has at least one kitchen object, like an oven or a coffee machine."

Atomic concepts are named classes whereas concept definitions are expressed as complex restrictions over roles (i.e., relations). Schütte clusters individual restrictions to property groups. He then iterates over these groups in order to generate *messages* for verbalization. What it essentially does is it generates a list of properties P_C that constitute necessary and/or sufficient conditions for C.

This approach can be extended for verbalizing ABox individuals. This task is related to the task of generating referring expressions (GRE) to the individual I. The difference is that a GRE algorithm iterates over a fixed list of properties in order to (a) find out whether a property holds for I and (b) whether it does not hold for a non-empty subset $D \in \mathcal{A}$ of remaining *distractors* – in other words: a referring expression should include only true information about the intended referent that helps distinguish the referent from potential distractors in the context – cf. [13] for a more complete account of the matter.

Now, in order to verbalize a description of an individual I, of course, only true statements about I should be included. The restriction to discriminatory information, however, must be left out. Querying the ontology for all properties that hold for I yields a list P_I^+ that can be verbalized using a mixture between concept lexicalization (see above) and referring expression generation for all individuals that are involved as relatees in P_I^+. In the following we will illustrate how P_I^+ and P_C together can be used in formalizing absent information.

B. Identifying Knowledge Gaps

Under an open-world assumption every individual I in the ABox *might* be an instance of every concept C. The mere ignorance of facts that would support such an inference is not interpreted as negative information. Only positive inferences are drawn by a reasoner that operates with open-world semantics. Everything else is considered a lack of knowledge. However, equating this with *knowledge gaps*, as relevant in the present work as well as in the CogX project as a whole, is impractical. Many of the concepts in a typical ontology in our domain (for instance [7], [12]) serve a predominantly structural purpose (i.e., they give structure to the ordered concept taxonomy), while offering little information that is directly linked to phenomena that are observable by the robot. Furthermore considering the typical size of such an ontology (e.g., > 200 concepts and > 10 individuals in a running system like [14]), a purely combinatorial approach to establishing gaps would lead to a massive over-generation that would be intractable for the robot. Within the project CogX we are thus identifying *knowledge gaps* as absent knowledge which can be gathered and established by the robot itself. Our notion of knowledge gap is hence inherently linked to the notion of *gap filling*.

In the present work, we make use of a crucial distinction in DL-based knowledge representations and formalisms. We distinguish individuals into *asserted* and *inferred* instances of a concept. This dichotomy corresponds to the difference between user-given knowledge and self-acquired knowledge of the robot. Presence of both kinds of knowledge is a typical characteristic of a robot's knowledge base after it was given a guided tour by its human user through their shared environment [15]. After such a tour, a robot might have acquired a spatial map whose units correspond to Room instances in the knowledge base, augmented with the user-given information that one of these rooms is a kitchen. The respective individual (say, e.g., AREA5) is what we call an asserted instance of the concept Kitchen. Put differently, during the tour the ABox $\mathcal{A}_{tour}$ is filled with the following facts:

(7) Part of $\mathcal{A}_{tour}$:
Room(AREA1), Room(AREA2),
Room(AREA3), Room(AREA4),
Room(AREA5), Kitchen(AREA5),
Room(AREA6)

AREA5 is thus *asserted* to be an instance of the named concept Kitchen, that is, it satisfies the left-hand side of the concept definition (1) above. At the same time $\mathcal{A}_{tour}$ does

not contain enough facts to satisfy the concept description given on the right-hand side. The reasoner cannot prove that the individual is an *inferred* instance of that concept. The asserted-inferred dichotomy thus provides a first interpretation of a relevant knowledge gap: the presence of facts that can be subsumed by a right-hand side of a concept definition would allow the reasoner's instance checking mechanism to infer a fact. Of course, the reasoner is *a priori* indifferent to whether a piece of information is asserted or inferred, as long as it can be assumed to be true. Such a case, on the other hand, provides an interesting opportunity for a robot to gather knowledge in a goal-directed way. The rationale behind this is the assumption that there exist objective facts that (a) lead the human user to make such an assertive statement, and (b) can be verified by the robot, such as, e.g., the presence of an object that is typical for kitchens.

Such a knowledge gap can be determined through introspection. The reasoning behind this is a sort of abduction: given that the robot knows $C(I)$, it can assume that it must be possible to satisfy a set of facts $F_{C'} = F_1, \ldots, F_n$ that can be subsumed by the right-hand side C' of the concept definition $C \equiv C'$ such that these facts give rise to the inference $C'(I)$.

(8) $\mathcal{A} \cup (\bigcup_{i=1}^{n} F_i) \models_{\mathcal{T}} C'(I)$.

To get back to our earlier descriptions, for every individual I and for all concepts $C_1, \ldots, C_n$ that I instantiates – $C_i(I) \in \mathcal{A}$ – one needs to determine the properties P_{C_i} that correspond to C_i'. The next step is then to check the known properties P_I^+ of I parallel to the verbalization task (cf. Section IV-A). For each $C_i(I)$ the list of lacking properties $P_i^?(I)$ can then be calculated as follows.

(9) $P_{C_i}^?(I) = P_{C_i} \backslash P_I^+$

C. Verbalizing Knowledge Gaps

Once lacking properties have been identified through introspection, it is possible to verbalize this missing information. The method is similar to the verbalization approaches presented above. Verbalization of knowledge gaps can be useful for eliciting new verbal information from the user in tutoring settings. However, it would require a concise interaction model in which to embed such verbalizations. The exact choice of words must of course also be carefully matched to the intended scenario. In this report we focus on *informative* verbalizations, which are not supposed to elicit human feedback. We rather propose a way of verbalizing knowledge gaps that can accompany an autonomous knowledge gathering behavior of an autonomous robot. For example consider a robot "gopher" that has just been taken on a guided tour through a user's apartment. At the end of the tour, the robot is left with a knowledge base like $\mathcal{A}_{tour}$ in example Example (7). The robot then decides to acquaint itself with the environment "on his own hook." The first plan it comes up with is to navigate to the kitchen and look out for important objects, such as the oven, the microwave, the coffee machine etc. In order to not intimidate its user with unpredictable behavior, the robot is programmed to inform her about its plans in order to establish *transparency*.

(10) "The kitchen is supposed to have typical objects like an oven or a coffee machine. I will go and check."

After informing the user, the robot turns around and heads for the kitchen to find and locate those objects. It is important to note that from a behavioral point of view, such an informative message only makes sense if the robot will then also execute the respective action. In architectural terms this means that the robot's planning module must first be presented with a possible goal, then it needs to check which action steps could have the desired outcome. Only if there is a sequence of action steps that yields the goal state, the planner can then decide to try to achieve the goal, executing one action at a time. As a first step, the planner should then initiate the verbalization action in order to inform the user about its plans. We will illustrate how such a *gap filling behavior* can be initiated in the next section (Section IV-D. Diverging from the above order of plan formation and instead following the chronology of the observable behavior, we will first have a look at how to generate and verbalize such an informative message.

Every non-empty set of lacking properties $P_i^?(I) \neq \emptyset$ qualifies for such a verbalization task $V(P_i^?(I))$. First the robot should make clear which entity in the world it will talk about. This is done by using a referring expression RE_I to the respective individual I as subject of the generated sentence. The rest of the informative utterance can then take one of two forms, depending on whether RE_I already contains the lexicalization $L(C_i)$ of the concept C_i in question. The most important part is the verbalization $V(P_i^?)$ of the lacking properties $P_i^?$. This is achieved using a modified version of Schütte's algorithm, which gets only a subset of properties for verbalization $P_i^? \subseteq P_{C_i}$.

(11) if $L(C_i) \in RE_I$:
$V(P_i^?(I)) = RE_I \circ$ "is supposed to" $\circ\, V(P_i^?)$

(12) otherwise:
$V(P_i^?(I)) = RE_I \circ$ "is a" $\circ\, L(C_i) \circ$ ", which is supposed to" $\circ\, V(P_i^?)$

In our example above, this individual I is AREA5. A referring expression RE_I "the kitchen" is then generated using our existing GRE algorithms [3], [13]. Since RE_I already contains the concept C_i in question, we can avoid generating the tautological "the kitchen is a kitchen, which is supposed to have typical objects like (…)" and instead produce a short informative sentence (cf. Example (10)).

D. Initiating Gap Filling Behavior

We start from the assumption that only those "blank spots" in an agent's knowledge should be considered as proper knowledge gaps for which there exist knowledge gathering actions that can potentially fill those gaps. Usually, these actions are provided by different modules of a robotic cognitive architecture. An introspective mechanism that is to present opportunities for knowledge gathering actions needs hence to be informed from the outside about such possible actions. This can be done by enforcing that each module

registers the kinds of actions and reasoning facilities it can provide with a central planning and motivation module [14].

We can thus postulate that the module containing the ontology reasoner and the introspection procedures be informed about the kinds of facts that can be established through knowledge gathering actions. One such action is *active visual search* (AVS), in which the robot efficiently locates one or more distinct objects in its environment. AVS benefits from a restricted search space, both in terms of spatial extent and number of object classes. An abductive reasoning over ontological knowledge can provide hypotheses for AVS, such as searching a given spatial area (e.g., "the kitchen") for a limited set of objects (e.g., coffee machines, ovens, microwaves, etc.). Just as ontology verbalization requires an additional layer of annotation, the ontology must contain information about which objects an AVS behavior can detect. In our case, those concepts that represent objects which can be detected visually are subsumed by the concept VisuallyDetectable.

(13) $\{\mathsf{VisuallyDetectable} \sqsubseteq \top,$
$\mathsf{Coffeemachine} \sqsubseteq \mathsf{VisuallyDetectable},$
$\mathsf{Oven} \sqsubseteq \mathsf{VisuallyDetectable},$
$\mathsf{Microwave} \sqsubseteq \mathsf{VisuallyDetectable}\} \subset \mathcal{T}_r$

The semantics of the concept VisuallyDetectable is that AVS can populate the ontology with individuals that instantiate this concept. The task of presenting knowledge gaps that can be filled by AVS now consists of identifying the set of facts $F_{C'}$ (as defined above) that are part of the post-condition of an AVS action. AVS can be defined as an action that takes as parameters a spatial location I_{loc} (e.g., an instance of Room) and a set of objects (object concepts, that is) $C_{obj} = C_1, \ldots, C_n \subseteq \mathsf{VisuallyDetectable}$ to search for in that location (e.g., {Oven, Coffeemachine}). The post-conditions of AVS can then be represented as a set of facts F_p.

(14) $F_p(I_{loc}, C_{obj}) = \{\mathsf{hasObject}(I_{loc}\, I_{obj} \sqcap C(I_{obj}) | C \in C_{obj}\}$

As introduced above, $P^?_{C_i}(I)$ corresponds to all potential lacking properties to establish $C_i(I)$. The intersection of all possible facts F_p that can be produced by a knowledge gathering action and $P^?_{C_i}(I)$ then yields the set of relevant facts $F_{C'_i}$ that could give rise to an inference $C_i(I)$. It is this $F_{C'_i}$, which is then presented to the planner as a potential goal state – thus denoting a *fillable knowledge gap*.

V. Conclusions and Future Work

The paper discussed ongoing research on developing methods for (a) introspecting a robot's ontological knowledge, (b) determining gaps in knowledge about a specific individual I relative to one or more concepts $C_1, ..., C_n$, and (c) verbalizing what the robot does and does not know about I relative to these concepts. The paper operated with a limited notion of "gap," defined as a property of a known concept C but unknown for I if instantiated as C. The proposed method used queries on an ontology to establish the conditions under which I could be an inferred instance of C, and determined gaps $P^?$ from the extent to which I did not yet fulfill these conditions. Verbalization then combined concept description with the known properties of I, P^+ and the gaps $P^?$, extending the approach proposed by Schütte [1].

There are several directions for future research we intend to follow. These concern the introspection mechanisms themselves, and the subsequent use of introspection in carrying out a situated dialogue for learning more about the environment.

a) Uncertainty over categories and properties: We are not considering an individual I, nor the ontologies against which we want to interpret I, in logical isolation. Concepts their instances are anchored in the perceptual and proprioceptive experience of a robot, over time and space. As uncertainty is inherent to a robot's experience, we need a way to deal with 'that' – uncertainty in the actual concepts an individual can instantiate, uncertainty about what properties can be recognized for the individual, and uncertainty about what values any of the recognized properties may take. As we already indicated in Section I, there are natural ways in which we can extend our method to deal with various sources of uncertainty. Given a set of alternative concepts $C_1, ..., C_n$ for I, ranked by (un)certainty, the method can retrieve a multiset of unknown properties, one $P^?_i$ for each concept C_i $(1 \leq i \leq n)$. We assume that each $P^?_i$ is finite, and each property $p^?_h \in P^?_i$ has a finite, discrete (or discretizable) range. Then, given u the number of unique properties in $P^?_1, ..., P^?_n$ we have $\mathcal{O}(u * n * (n-1))$ comparisons between concepts to establish which properties are shared between $C_1, ..., C_n$. For $Shared = \{p_1, ..., p_m\}$ the set of shared properties we can subsequently determine, how each property (by presence) would help split the set of concepts into evenly balanced subsets. We can directly determine this by a linear computation over the findings in the comparison matrix constructed in the previous step. Given a cost function over observing particular properties, and the uncertainty in having (or not having) observed that property for I, we want to investigate how ranking gaps in order of cost/uncertainty can help us establish what would be the most suitable sequence for querying the environment or a human about these gaps, to establish the correct concept for I.

b) Weighted abduction for lowest-cost proofs of categorical identity: A cost-based ranking over gaps provides a direct connection with cost-based planning or inferencing for dealing with gaps. Kruijff & Janiček [16] present a form of weighted abduction based on [17], [18], [19], among others extending it with a notion of assertion alike [20]. This kind of abduction helps us to establish the lowest-cost proof for making an update to the robot's belief model. Weights in this form of abduction represent uncertainty in knowledge [21]. We can use this to reflect uncertainty in an observation (the cost of making the right assumption) or the actual cost of making an observation (as per above). Kruijff & Janiček introduce assertions to identify propositions that are included in an abductive proof, but which are in need

of future validation. We intend to investigate how we can see establishing concept membership of an individual I as a weighted abductive proof. Given the costs of "deciding" between alternative concepts, and given the (un)certainties for the observations about I, which concept C would establish the lowest-cost proof for I instantiating C? Using a form of inference like abduction would also enable us to take into account any logical structure over the interrelations between properties for a concept (as often explored in inferential forms of knowledge discovery).

c) From verbalization to information requests: Another interesting use of a ranking over gaps that are discriminative between potentially applicable concepts, is to drive verbalization and dialogue planning for information requests. Realized gaps provide a focus – they are the properties we are after when asking for more information. This helps us to structure a dialogue for getting more information. We would like to investigate how the abductive view on continual collaborative activity, explored by Kruijff & Janiček, can provide a basis for comprehending, deliberating, and producing such (sub-)dialogues for information requests to help resolve gaps.

Finally, we will need to generalize from the notion of gap used in this paper, to more general notions. In this paper we only focus on understanding an individual at the ABox level in an ontology. We are not dealing with extending existing concepts in the TBox, or even establishing new concept classes. It currently remains an open question within CogX how to define such more general gaps, how to ground them in perceptual and proprioceptive models (i.e. in "experience"), and how to use them in self-extension.

ACKNOWLEDGMENTS

This work was supported by the EU FP7 ICT Project "CogX" (FP7-ICT-215181).

REFERENCES

[1] N. Schütte, "Generating natural language descriptions of ontology concepts," in *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, Athens, Greece, March 2009, pp. 106–109.

[2] D. Traum and S. Larsson, "The information state approach to dialogue management," in *Current and New Directions in Discourse and Dialogue*, J. van Kuppevelt and R. Smith, Eds. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2003.

[3] H. Zender, G.-J. M. Kruijff, and I. Kruijff-Korbayová, "Situated resolution and generation of spatial referring expressions for robotic assistants," in *Proceedings of the Twenty-first International Joint Conference ,on Artificial Intelligence (IJCAI-09)*. Pasadena, CA, USA: AAAI Press, July 2009.

[4] R. Dale and E. Reiter, "Computational interpretations of the Gricean Maxims in the generation of referring expressions," *Cognitive Science*, vol. 19, no. 2, pp. 233–263, 1995. [Online]. Available: citeseer.ist.psu.edu/dale94computational.html

[5] S. Roa and G. Kruijff, "Curiosity-driven acquisition of sensorimotor concepts using memory-based active learning," in *Proceedings of the 2008 IEEE International Conference on Robotics and Biomimetics (ROBIO 2008)*, 2008.

[6] A. Thomaz, "Socially guided machine learning," Ph.D. dissertation, Massachusetts Institute of Technology, May 2006.

[7] H. Zender, O. M. Mozos, P. Jensfelt, G.-J. M. Kruijff, and W. Burgard, "Conceptual spatial representations for indoor mobile robots," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 493–502, June 2008.

[8] G. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, H. Zender, and I. Kruijff-Korbayová, "Situated dialogue processing for human-robot interaction," in *Cognitive Systems*, H. Christensen, G. Kruijff, and J. Wyatt, Eds. Springer Verlag, 2009, available at `http://www.cognitivesystems.org/cosybook`.

[9] F. Baader, "Description logic terminology," in *The Description Logic Handbook: Theory, Implementation, and Applications*, F. Baader, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds. Cambridge, UK; New York, NY, USA: Cambridge University Press, 2003, ch. Appendix 1.

[10] D. Nardi and R. J. Brachman, "An introduction to description logics," in *The Description Logic Handbook: Theory, Implementation, and Applications*, F. Baader, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds. Cambridge, UK; New York, NY, USA: Cambridge University Press, 2003, ch. 1.

[11] F. Baader and W. Nutt, "Basic description logics," in *The Description Logic Handbook: Theory, Implementation, and Applications*, F. Baader, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds. Cambridge, UK; New York, NY, USA: Cambridge University Press, 2003, ch. 2.

[12] H. Zender and G. Kruijff, "Multi-layered conceptual spatial mapping for autonomous mobile robots," in *Control Mechanisms for Spatial Knowledge Processing in Cognitive / Intelligent Systems*, ser. AAAI Spring Symposium 2007, March 2007.

[13] ——, "Towards generating referring expressions in a mobile robot scenario," in *Language and Robots: Proceedings of the Symposium*, Aveiro, Portugal, December 2007, pp. 101–106.

[14] N. Hawes, H. Zender, K. Sjöö, M. Brenner, G.-J. M. Kruijff, and P. Jensfelt, "Planning and acting with an integrated sense of space," in *Proceedings of the 1st International Workshop on Hybrid Control of Autonomous Systems – Integrating Learning, Deliberation and Reactive Control (HYCAS)*, Pasadena, CA, USA, July 2009, pp. 25–32.

[15] E. A. Topp, H. Hüttenrauch, H. Christensen, and K. Severinson Eklundh, "Bringing together human and robotic environment representations – a pilot study," in *Proc. ofIROS-2006*, Beijing, China, October 2006.

[16] G. Kruijff and M. Janiček, "Abduction for clarification in situated dialogue," EU FP7 CogX project, Tech. Rep., 2009, Deliverable DR6.1.

[17] M. Stone and R. Thomason, "Context in abductive interpretation," in *Proceedings of EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue*, 2002.

[18] ——, "Coordinating understanding and generation in an abductive approach to interpretation," in *Proceedings of DIABRUCK 2003: 7th workshop on the semantics and pragmatics of dialogue*, 2003.

[19] R. Thomason, M. Stone, and D. DeVault, "Enlightened update: A computational architecture for presupposition and other pragmatic phenomena," in *Presupposition Accommodation*, D. Byron, C. Roberts, and S. Schwenter, Eds., (to appear).

[20] M. Brenner and B. Nebel, "Continual planning and acting in dynamic multiagent environments," *Journal of Autonomous Agents and Multiagent Systems*, 2008.

[21] J. Hobbs, "Abduction in natural language understanding," in *Handbook of Pragmatics*, L. Horn and G. Ward, Eds. Blackwell, 2004, pp. 724–741.

Phrasing Questions

Geert-Jan M. Kruijff
German Research Center
for Artificial Intelligence (DFKI GmbH)
Saarbrücken, Germany
gj@dfki.de

Michael Brenner
Institute for Computer Science
Albert-Ludwigs-Universität
Freiburg, Germany
brenner@informatik.uni-freiburg.de

Abstract

In a constructive learning setting, a robot builds up beliefs about the world by interacting – interacting with the world, and with other agents. Asking questions is key in such a setting. It provides a mechanism for interactively exploring possibilities, to extend and explain the robot's beliefs. The paper focuses on how to linguistically phrase questions in dialogue. How well the point of a question gets across depends on how it is put. It needs to be effective in making transparent the agent's intentions and beliefs behind raising the question, and in helping to scaffold the dialogue such that the desired answers can be obtained. The paper proposes an algorithm for deciding what to include in formulating a question. Its formulation is based on the idea of considering transparency and scaffolding as referential aspects of a question.

Introduction

Robots are slowly making their entry into "the real world." And it is slowly becoming an accepted fact of life that we cannot possibly provide such robots will all there is to know, out-of-the-box. So they need to learn. The point of socially guided (machine) learning (Thomaz 2006) is that some of that learning can be done effectively through social interaction with other agents in the environment.

This paper focuses on how a robot should phrase its questions, considering a social learning setting in which situated dialogue is the main interactive modality (Kruijff et al. 2006a; Jacobsson et al. 2007). The robot and a human use spoken dialogue to discuss different aspects of the environment. We consider learning to be driven by the robot's own, perceived learning needs. This requires dialogue to be mixed-initiative. Both the human and the robot can take the initiative in driving this "show-and-tell-then-ask" dialogue. Questions play a fundamental role in such dialogues. Assuming a robot has the ability to raise issues in need of clarification or learning for any modality, (e.g. (Kruijff, Brenner, and Hawes 2008)), the problem thus becomes how to properly *phrase* a question.

Typically, a question is represented as an abstraction over the argument of a predicate. For example, assuming $?x.P(x)$ to indicate that a question regards a parameter x of some predicate $P(x)$, a question about the color of a ball could be phrased as $?x.(ball(y) \wedge has{-}color(y,x))$. However, more aspects need to be taken into account, for a question to be posed in such a way that the addressee is likely to understand the question and provide a suitable answer (Ginzburg 1995b).

First of all, the phrasing needs to make *transparent* how a question arises from an agent's beliefs, what beliefs – and what gaps in an agent's beliefs – it refers to. It should make clear *what a question is about*. Furthermore, there is a reason behind raising the question. The agent has a specific goal, it intends to obtain a particular kind of answer. Not just any answer will do. Raising a question also needs to set up, *scaffold*, the right context for answering it. This is the *why* of a question, pointing to how the agent would like to see the question *resolved*.

An example in (Kruijff et al. 2006b; 2007b) provides an interesting illustration.[1] The robot is capable of figuring out when it might have mistakenly classified a particular passage in the environment as a door. At the point where it realizes this, it asks, "Is there a door here?" Unfortunately, the place where it asks this is not related to the location "here" refers to. To anyone but a developer-acting-as-user it is not transparent what the "here" means. This often leads to the user giving the wrong answer, namely "yes this room has a door" rather than, "no, there is no door between the trash bin and the table." The way the question was phrased lacked both in transparency (location reference) and in scaffolding (specific location, not the room as such).

The paper presents an approach to generating a content representation for a question. These representations reflect what is being asked after, in reference to beliefs (aboutness, transparency) and intentions (resolvedness, scaffolding). The approach explicitly regards transparency and scaffolding as *referential qualities* of a question. This way their referential nature in the larger dialogue- and situated context can be considered. Following out that idea, the approach bases its content determination algorithm on Dale & Reiter's incremental algorithm for generating referring expressions (Dale and Reiter 1995), in combination with algo-

[1] See also the video at the CoSy website's Explorer page, at http://cosy.dfki.de/www/media/explorer.y2.html.

rithms for referential context determination (Zender, Kruijff, and Kruijff-Korbayová 2009; Paraboni, van Deemter, and Masthoff 2007).

Central to the approach is establishing the information pertaining to the question. A description logic-like formalism is used to represent such information, as a conceptual structure in which propositions have ontological sorts and unique indices, and can be related through named relations. A question can then be represented as a structure in which we are querying one or more aspects of such a representation (Ginzburg 1995b; Kruijff, Brenner, and Hawes 2008). The formalism allows everything to be queried: relations, propositions, sorts. Around the formulation of a question we construct a nucleus, comprising the situation (the "facts") and the beliefs that have led up to the question, the question itself, and the goal content which would resolve the question. The question nucleus integrates Ginzburg's notions of aboutness, and (potential) resolvedness.

Based on the question nucleus, the algorithm starts by determining to what extend the different aspects are covered by the (dialogue) common ground between the robot and the human. For this, contextual references are resolved in a dialogue context model (Kruijff et al. 2007a), and it is established how these can be related to inferences over domain knowledge and instances (Kruijff et al. 2007b). The question nucleus is extended with these connections – or rather, with indications of the information structure or informativity of individual content – so that it includes an explicit notion of what is shared, and what is privately held information (cf. (Lochbaum, Grosz, and Sidner 1999; Grosz and Kraus 1999)).

The algorithm next decides what aspects of a question nucleus to include in the content for phrasing the question. For each aspect of the nucleus (facts, beliefs, question, goals) the algorithm uses the informativity of the aspect's content, in conjunction with similarly related but contrasting content in the dialogue context model, to determine whether to include it. Essentially, new or contrastive content will be considered, whereas salient "old" information will not. The form in which the content will be included is determined by content-specific algorithms for generating referring expressions (e.g. (Kelleher and Kruijff 2006; Zender, Kruijff, and Kruijff-Korbayová 2009)). The decisions to include particular content can be weighted according to a comprehensibility ranking as e.g. in (Krahmer, van Erk, and Verleg 2003).

The contributions the approach aims for are, briefly, as follows. Purver and Ginzburg develop an account for generating questions in a dialogue context (Purver, Ginzburg, and Healey 2003; Purver 2004). Their focus was, however, on clarification for the purpose of dialogue grounding. A similar observation can be made for recent work in HRI (Li, Wrede, and Sagerer 2006), We are more interested in formulating questions regarding issues in building up situation awareness, including the acquisition of new ways of understanding situations (cf. also (Kruijff, Brenner, and Hawes 2008)). In issue-based (or information state-based) dialogue systems (Larsson 2002), the problem of how to phrase a question is greatly simplified because the task domain is fixed. There is little need for paying attention to transparency or scaffolding, as it can be assumed the user understands the task domain.

An overview of the paper is as follows.The paper starts with a discussion of basic issues in modeling questions and their semantics, based on (Ginzburg 1995b). Then the approach is presented. The approach starts from the assumption that a question is a dialogue, not just a single utterance. Discussed is how the content plan for such a question dialogue can be determined, providing definitions, representation, and algorithms. The paper ends with a discussion of how the approach could be integrated, evaluated, and points for further research.

Background

What is a question? Ginzburg (1995b) discusses a variety of linguistic approaches. All of them aim to provide an invariant characterization of the semantics of a question. Broadly, they have proposed the following aspects as crucial to that definition.

First, several approaches propose to see a question as an n-ary relation. The relation puts together the question with one or more contributions pertaining to answering it. The point here is to take into account the fact that a question can be discussed over several turns in a dialogue. Second, there is a sense of *aboutness* to a question. Each question can be associated with a collection of propositions, which are –intuitively– related to the question. And, finally, each question can be considered to be associated with a (possibly complex) proposition which provides an *exhaustive answer*. In other words, an exhaustive answer resolves the question.

Ginzburg suggests that all these aspects together make up a characterization of a question – not just one of them, as most approaches suggest. Furthermore, these aspects are to be understood as being *relative*. What a question is about, and how it can be resolved, should be understood relative to an agent's *goal* and *belief/knowledge state* (cf. also (Ginzburg 1995a)). The following example illustrates this.

(1) Context: a robot drives around campus, and is about to enter the DFKI building.
   a. Janitor: Do you know where you are?
      Robot: DFKI.
   b. Janitor believes the robot knows where it is.

(2) Context: a robot drives around the DFKI building, to get a cup of coffee.
   a. Janitor: Do you know where you are?
      Robot: DFKI.
   b. The janitor is not convinced the robot really knows where it is.

What counts as an answer to a question may thus vary across contexts. What a question is thus cannot be reduced to an analysis of just what counts as its answers. Instead, Ginzburg starts with setting up an ontology in which questions, propositions and facts are considered as equal citizens. This makes it possible to consider a question *in relation to*

possible answers for it. The ontology is defined using situation theoretic constructs, which we will adopt throughout this paper. (All definitions as per (Ginzburg 1995a; 1995b).)

Definition 1 (SOA, Situation, Fact)**.** A SOA (State Of Affairs) describes possible ways an actual situation might be. SOAs are either *basic*, or built up from basic ones using algebraic operations. A *basic SOA* is an atomic possibility, written as $\langle R, f : i\rangle$ with R a relation, f a mapping assigning entities to the argument roles of R, and i is a polarity i.e. $i \in \{+, -\}$. A situation s supports the factuality of a SOA σ iff $s \models \sigma$. The SOA σ is then considered a *fact* in s. To enable complex SOAs, SOAs can be structured as a Heyting algebra under a partial order '$\rightarrow$', which is closed under arbitrary meets ($\bigwedge$) and joins ($\bigvee$). Situations and SOAs together form a SOA-algebra:

1. If $s \models \sigma$ and $\sigma \rightarrow \tau$ then $\sigma \models \tau$
2. $s \not\models 0$, $s \models 1$ (FALSE,TRUE)
3. If Σ is any finite set of SOAs, then $s \models \bigwedge \Sigma$ iff $s \models \sigma$ for each $\sigma \in \Sigma$
4. If Σ is any finite set of SOAs, then $s \models \bigvee \Sigma$ iff $s \models \sigma$ for at least one $\sigma \in \Sigma$

Finally, an application operator is defined, to allow for variable assignment (and reduction):
$\lambda x.\langle R, a : b, c : x : +\rangle | x \mapsto d| = \langle R, a : b, c : d : +\rangle$ □

Using Definition 1, we can now consider a proposition to be an assertion about the truth of a possibility relative to a situation.

Definition 2 (Proposition)**.** A proposition p is a relational entity, asserting a truth regarding a SOA τ in a particular situation s: $p = (s : \tau)$. A proposition $p = (s : \tau)$ is TRUE iff τ is a *fact* of s, denoted as $s \models \tau$. □

Before defining what a question is, the notions of *resolvedness* and *aboutness* need to be defined. Resolvedness, or rather the broader concept of *potentially resolving* a question, is defined as follows. The definition distinguishes whether a (possibly complex) fact resolves a question depending on whether the question is *polar*, asking for the truth of an assertion (e.g. "Is the ball red?"), or *factive*, asking after a value (e.g. "What color is the ball?").

Definition 3 (Resolvedness conditions)**.** A SOA τ *potentially resolves* a question q if either

1. τ positively-resolves q (for 'polarity p': any information that *entails* p; for a factive question: any information that entails that the extension of the queried predicate is non-empty)
2. τ negatively-resolves q (for 'polarity p': any information that *entails* $\neg p$; for a factive question: any information that entails that the extension of the queried predicate is empty)

□

We will leave the notion of *aboutness* for the moment. Essentially, Ginzburg (1995a; 1995b) defines this as a collection of SOAs which can be associated with the content of a question q, with a SOA being about q if it subsumes the fact that q is either positively or negatively resolved. (For subsumption, recall Definition 1.)

Ginzburg's definition of what a question is then works out as follows.

Definition 4 (Question)**.** A question is an entity $(s?\mu)$ constructed from a situation s and an n-ary abstract SOA $\mu = \lambda x_1, ..., x_n \sigma(x_1, ..., x_n)$ $(n \geq 0)$:

1. μ constitutes an underspecified SOA from which the class of SOAs that are *about* q can be characterized.
2. Those SOAs which are facts of s and informationally subsume a level determined by μ constitute a class of SOAs that *potentially* resolve q.

□

The definition includes references to the relational character of a question (the abstract), and the notions of aboutness (intuitively, the space within which we are looking for an answer) and of resolvedness (the space of possible answers we are looking for, one of which will -hopefully- establish itself as fact). Finally, we already indicated above that resolvedness is an agent-relative notion. Ginzburg suggests to do so using Definition 3 as follows.

Definition 5 (Agent-relative resolvedness)**.** A fact τ *resolves* a question $(s?\mu)$ relative to a mental situation ms iff

1. Semantic condition: τ is a fact of s that potentially resolves μ
2. Agent relativisation: $\tau \implies_{ms} Goal - content(ms)$, i.e. τ entails the goal represented in the mental situation ms relative to the inferential capabilities encoded in ms.

□

Approach

The previous section presented a formal (but relatively abstract) notion of what a question is. It made clear that a question is more than a predicate with an open variable, or (alternatively) just another way of characterizing a set of propositions that would serve as exhaustive answer. Instead, a question is a relational structure, tying into a larger context. For one, this "context" provides a set of beliefs (SOAs, in Ginzburg's terms), a background within which potential answers are sought. An agent's goals help motivate to focus which beliefs are associated with the question. Another point about this "context" is that a question isn't just a single utterance, or just forming a unit with an utterance that answers it. There is a dialogue context in which this question is phrased. The question itself, and whatever utterances contribute to help clarify, refine and answer that question, may (though need not) refer to content already established in that context.

Phrasing a question, in other words, means we need to provide the possibility for such contextual factors to influence how the content of a question is determined. Once the

agent has determined that it needs to raise a question, and about what (e.g. cf. (Kruijff, Brenner, and Hawes 2008) for questions in situated forms of learning), it needs to establish how best to communicate the question. In this paper, we suggest to do this as follows. We will begin by further explication of the notion of question, using a structure we term the *question nucleus*. The question nucleus captures more explicitly the relation between beliefs and intentions that are active in a current context, and how they determine the space of possible answers (or complexes of those). Then, we sketch several algorithms. The first group of algorithms concern *context determination*. Intuitively, these algorithms determine what beliefs and potential answers form the relevant background for the question. The background specifies what can be assumed to be known, (and can thus be referred to or even silently assumed), both in terms of content and intentions in the the dialogue- and situated context. How a question is to be phrased relies on what it needs to explicate relative to that background, to effectively communicate it. This is then finally done by the *content determination* algorithm. The result of this algorithm is a logical form, expressed in a (decidable) description logic. The logical form specifies the core content for the question, which a content planner subsequently can turn into one or more fully-fledged utterances.

The following definition defines more precisely what we mean by a logical form, based on (Blackburn 2000; Baldridge and Kruijff 2002). We will use the same formalism to describe SOAs (cf. Definition 1).

Definition 6 (Logical forms)**.** A logical form is a formula ϕ built up using a sorted description logic. For a set of propositions $PROP = \{p, ...\}$, an inventory of ontological sorts $SORT = \{s, ...\}$, and a set of modal relations $MOD = \{R, ...\}$, $\phi = p \mid i : s \mid \psi \wedge \psi' \mid \langle R \rangle \psi \mid @_{i:s}\psi$. The construction $i : s$ identifies a nominal (or index) with ontological sort s. The at-operator construction $@_{i:s}\psi$ specifies that a formula ψ holds at a possible world uniquely referred to by i, and which has ontological sort s. □

A standard Kripke-style model-based semantics can be defined for this language (Blackburn 2000). Intuitively, this language makes it possible to build up relational structures, in which propositions can be assigned ontological sorts, and referred to by using i as indices. For example, $@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle (c1 : color \wedge \textbf{red})$ means we have a "ball" entity, which we can uniquely refer to as $b1$, and which has a (referable) color property. (An alternative, equal way of viewing this formula is as a conjunction of elementary predications: $@_{b1:entity}\textbf{ball} \wedge @_{b1:entity}\langle Property \rangle c1 : color \wedge @_{c1:color}\textbf{red}$.)

Question nucleus

We start by defining the notion of *question nucleus*. The function of a question nucleus is twofold. First, it should capture the question's background in terms of associated beliefs and intentions, and what space of expected answers these give rise to. An expected answer is naturally only as specific (or unspecific) as is inferable on the basis of what the agent knows.

Definition 7 (Expected answer)**.** An expected answer a for a question q is a proposition $a = (s : \tau)$, with τ potentially resolving q as per Definition 3. τ is a logical formula (Definition 6) which can be underspecified, both regarding the employed ontological sorts, and arguments. □

Effectively, assuming that the agent has a collection of ontologies which provide a subsumption structure ($a \sqsupset b$ meaning a subsumes b, i.e. b is more specific), an expected answer can be said to define a "level" of specifity (Definition 4) according to subsumption. Following up on the ball example, assume the agent has an ontology which defines $material - property \sqsupset \{color, shape\}$. An expected answer to a question, what particular shape the ball has, would take the form $@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle (s1 : shape))$. All the proposition specifies is that there is an identifiable shape. If the question would be about any, or some unknown, property of the ball, an expected answer could be phrased as $@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle (m1 : material - property))$. Using the available ontological structure, and relational structure between formulas, we can formulate expected answers at any level of specifity without requiring the agent to already know the answer (cf. also (Kruijff, Brenner, and Hawes 2008)).

Definition 8 (Question nucleus)**.** A *question nucleus* is a structure $qNucleus = \{r, BL, XP, AS\}$ with:

1. A referent r relative to which the question q (part of XP) is phrased.
2. BL (*Beliefs*) is a set of private and shared beliefs, about agent intentions and facts in the current context (cf. (Lochbaum, Grosz, and Sidner 1999; Grosz and Kraus 1999)).
3. XP (*Execution Plan*) is a continual plan with an execution record (Brenner and Nebel 2008) for resolving a question $q = (s?\mu)$.
4. AS (*Answer Structure*) is a finite $\sqsupset$-structure over propositions $p_1, ...$ which potentially resolve q, and which are implied by BL.

The beliefs BL specify what the agent knows about r, what the agent presumes to be shared knowledge about r, and what the agent presumes other agents could know about r. BL is based on the dialogue leading up to the question, any previous actions involving r, and a domain model of agent competences (Brenner and Kruijff-Korbayová 2008). XP makes explicit that phrasing a question constitutes a dialogue, with an associated plan for communicating the question and a record for how far the question has been fully answered. This record maintains which aspects (elementary predications) of the question are still open ("under discussion," similar to the Question-Under-Discussion construct of (Ginzburg 1995b)). The AS is a set of propositions, relating those propositions to the aspect(s) of the question they would potentially resolve (and thus to the execution record in XP). AS is based on propositions implied by BL (relative to r, q) and is $\sqsupset$-structured according to ontological structure. □

Contextually determining aboutness

Asking a question starts with the agent having determined what it is it needs to know about some referent r, e.g. an area in the environment, an object – or, more specifically, relations or properties. (To allow for group referents, we will consider r to be a *set*.) Next the question nucleus is built up, starting with the beliefs about the question, BL.

We adopt the approach to belief modeling described in (Brenner and Kruijff-Korbayová 2008). Beliefs are formulated as relational structures with *multi-valued state variables* (MVSVs). These state variables are used for several purposes. First, they can indicate domain values, as illustrated by the sorted indices in the examples above. The color $c1$ would be a Property-type state variable of the entity $b1$, and could take domain values in the range of that ontological sort. Important is that the absence of a value for an MVSV is interpreted as *ignorance*, not as falsehood: $@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle(s1 : shape))$ means the agent does not know what shape the ball has, not that it has no shape (as per a closed-world assumption). In a similar way, state variables are used for expressing *private beliefs*, and mutual or *shared beliefs* (Lochbaum, Grosz, and Sidner 1999; Grosz and Kraus 1999). A private belief of agent a_1 about content ϕ is expressed as $(K\{a_1\}\phi)$ whereas a mutual belief, held by several agents, is expressed as $(K\{a_1, a_2, ...\}\phi)$. Secondly, MSVSs can be quantified over, for example using the ? to express a question: $?s1.@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle(s1 : shape))$ represents a question regarding the shape of the referent $b1$.

As an agent perceives the environment, we assume it builds up beliefs about the instances it perceives, and what relations can be observed or inferred to hold between them. For example, see (Brenner et al. 2007) for a robot manipulating objects in a local visual scene, or (Kruijff et al. 2007b) for a robot exploring an indoor environment. Furthermore, we assume that the agent's planning domains include models of agent capabilities – what another agent is capable of doing, including talking (and answering questions!) about particular aspects of the environment (Brenner and Kruijff-Korbayová 2008). Finally, if the agent has been engaged in a dialogue with another agent, and discussed the referent-in-question r before, we assume that the (agreed-upon) content discussed so far constitutes shared beliefs, held by all agents involved.

Algorithm 1 : Determine(BL) (*sketch*)

Require: `BELs` is a set of private and mutual beliefs the agent holds, (including beliefs about capabilities); r is the referent (set) in question

```
BL = ∅
for b ∈ BELs do
  if b includes a MVSV m ∈ r then
    BL = BL ∪ b
  end if
end for

return BL
```

Algorithm 1 sketches the basis of the algorithm for establishing BL. Those beliefs are gathered which refer explicitly to the referent the question is about. Note that BL may end up being empty. This means that r has not been talked about, nor does the agent know whether another agent could actually offer it an answer to what it would like to know more about.

Contextually determining resolvedness

The beliefs BL about the referent in question r state what the agent already believes about r (privately, or shared), and what it believes about another agent's capabilities. Next, these beliefs need to be structured such that potentially resolving answers can be derived. We assume that we can make use of the ontological sorts, and the structuring over these sorts provided by domain ontologies, to organize beliefs. The organization we are after first of all relates a belief to a potentially resolving answer, by combining it (inferentially) with the ?-quantified, ontologically sorted MVSVs in the question to yields a partially or completely reduced logical form (Definition 1). Secondly, the organization relates beliefs by (sortal) subsumption over the potentially resolving answers they generate.

For example, consider a question about the color of a ball: $?c1.@_{b1:entity}(\textbf{ball} \wedge \langle Property \rangle(c1 : color))$. Let us assume the robot holds several beliefs with regard to $b1$, and the variable $c1$. A robot learning more about visual properties of objects through interaction with a human tutor (Jacobsson et al. 2007) typically holds at least beliefs about what the tutor is capable of telling it. Thus, assume the robot believes the tutor can tell it about material properties, colors, and shapes. Using `tell-val` (*tell value* action) we can model these beliefs as $(K\ \{a_1\}\ tell - val(a_2, m : material - property)$, $(K\ \{a_1\}\ tell - val(a_2, c : color)$. The variables m, b are existentially bound in these beliefs. Using the inference that $material - property \sqsupset color$ and introducing bound variables m', c' for m and c respectively, the beliefs can be combined with the question to yield the potentially resolving propositions $c' : color, m' : material - property$. Furthermore, subsumption yields $m' : material - property \sqsupset c' : color$. Thus, by combining the beliefs with what the agent already knows, it can expect to know something it doesn't yet know by asking a question. And by making use of the way its knowledge is ontologically structured, it can determine how precise that answer is likely to be.

Algorithm 2 provides a first sketch of the algorithm for establishing AS. (In the current version, propositional content and additional relational structure pertaining to m in the context of b is not yet included into AS.)

Content determination

Finally, once the beliefs about q and the potentially resolving answers for q have been established, we can turn to determining the exact content for communicating q. The purpose of content determination is to establish what, how much, should be communicated for the agent to get an appropriate answer – how much content it needs to communicate to ensure proper scaffolding and transparency. For example,

Algorithm 2 : Determine(AS) (*sketch*)

Require: BL is a set of beliefs relative to r, q is a question about r, and ONT is a collection of ontologies supporting subsumption inferences on sorts used in BL and q.

AS = $\emptyset$ (empty subsumption)
for b $\in$ BLs **do**
  $\phi = \top$
  for MVSV $m \in r$ existentially bound in b **do**
    introduce a bound variable m'
    $\phi = \phi \wedge m' : sort(MVSV)$
  end for
  AS = AS $\sqcup \phi$, under $\sqsupset$
end for

return AS

consider again the question about the color of the ball. How the question should be phrased, depends on whether e.g. the ball has already been talked about, what goals are involved (are we learning how this ball looks like, or how objects roll?), etc. Example 3 provides some illustrations.

(3) Asking about the color of a single ball on a table ...

a. If the robot is not sure whether the other agent knows about colors:
"Could you tell me about the color of this ball?"

b. If the robot believes the other agent knows about colors:
" Could you tell me what color this ball is?"

c. If the robot is not sure whether asking about color is relevant to the current goal:
"I would like to know more about the color of this ball. Could you tell me what it is?"

d. If the ball is under discussion, and asking for color is relevant:
"What's the color?"

Example 3 particularly illustrates how scaffolding and transparency come into play. We connect these terms explicitly to the question nucleus. We see scaffolding primarily as appropriately embedding a question into an intentional setting, relating to AS and the extent to which available beliefs lead to specific (potentially resolving) answers. Transparency relates to the referential setting of the question nucleus, relating r to BL in the sense of what the agent can already assume to be mutually known about the referent under discussion. Planning the question as a dialogue, then, means determining relevant beliefs, and the information status of relevant content. Relevant beliefs are those which are associated with maximally specific, potentially resolving answer(s). A distinction needs to be made between private and mutual believes, particularly as beliefs about competences are first and foremost private beliefs. Furthermore, it should be determined whether these beliefs fit into the current intentional context. (For the purposes of the current paper, we will consider learning goals only, and consider them to specify what ontological sorts the agent is trying to learn.) Information status regards whether content, pertaining to r, can be assumed to be mutually known – most notably, whether r is mutually known (i.e. mutually identifiable in context).

Algorithm 3 : Content determination (*sketch*)

Require: BL is a set of beliefs relative to r, q is a question about r, ONT is a collection of ontologies supporting subsumption inferences on sorts used in BL and q, AS is a structure over potentially resolving answers

RelBL = $\emptyset$
for a $\in$ AS **do**
  if a is maximally specific, i.e. there is no a' s.t. a $\sqsupset$ a' **then**
    RelBL = RelBL $\cup$ { b }, for b yielding a
  end if
end for
MutualRelBL = mutual beliefs in RelBL
ScaffoldingBL = $\emptyset$
TransparencyBL = $\emptyset$
for MVSV m in q **do**
  if there is a $b \in$ MutualRelBL associated to m **then**
    TransparencyBL = TransparencyBL $\cup$ { b }
  else
    ScaffoldingBL = ScaffoldingBL $\cup$ { beliefs associated to most specific answers for m }
  end if
end for
return ScaffoldingBL, TransparencyBL

Algorithm 3 first determines what beliefs are relevant to achieve a maximally specific answer, and which of these beliefs are mutual. How much scaffolding needs to be done depends on whether these mutual beliefs imply all potentially resolving answers to the questioned MVSVs in r. If not, the algorithm backs off by constructing a belief set which needs to be communicated for appropriate scaffolding. The basis for transparency is formed by the mutual beliefs about r.

On the basis of these sets of beliefs, and q itself, the communication of q can be planned. We do not provide an in-depth discussion of dialogue- and content-planning here, for space (and time) reasons. We refer the interested reader to (Brenner and Kruijff-Korbayová 2008; Kruijff et al. 2009). In brief, beliefs in the scaffolding set are specified as assertions (Brenner and Nebel 2008). The plan for communicating the question starts by verifying these assertions, and then raises the question itself. It is a matter for content fusion whether such verification can be done in conjunction with the question itself (Example 3, a–b) or as preceding utterances (Example 3, c). For the realization of the question, the transparency beliefs are used to determine information status. Content planning then turns information status into decisions about how to refer to r and the asked-after properties – e.g. using pronominal reference (Example 3, c) or even omitting explicit reference, by eliding any mention of r (Example 3, d).

Conclusions

The approach presented in this paper is still under development. The key technologies it is based on (planning, motivation, dialogue processing, and ontological inferencing) are already available in the system architecture the approach will be integrated into. We will describe the full integration, with working examples, in a full version of this paper. We will then also consider how this approach can be applied in related settings, such as performance requests.

We are currently considering various alternative ways to evaluate the approach. User experiments are just one option here. The problem is that an approach as presented here, and the overall architecture it will be integrated into, present a large parameter space. Consequently, it is difficult to ensure a controlled setting for a user experiment – and, only a very limited part of the parameter space can be effectively explored. An alternative way we are therefore currently considering is to use techniques from language evolution. In simulations we would like to explore what the effects of different parameter settings would be on how agents are able to communicate, and what this consequently means for measurable parameters such as learning performance. Examples of such experiments can be found in (Ginzburg and Macura 2006).

There remain for the moment plenty of open issues to be investigated further – this paper really only provides a first description of the approach we are developing. It does aim to make clear how notions such as scaffolding and transparency can be folded into a characterization of how a system can phrase a question – seeing a question, in fact, as a subdialogue to be planned, not just a single utterance paired with a possible answer. Basic issues remain in the construction of the various belief sets, and the associated structures over potentially resolving answers. Although an "unweighted" approach as followed here will work for most simple scenarios, it remains to be seen whether associating *costs* with beliefs (and assuming them, in a plan for communicating a dialogue) could provide a more adaptive, scalable approach in the long run. Furthermore, the current formulation of the construction of the answer structure AS (Algorithm 2) does not cover polar questions (though this is an easy extension).

Acknowledgments

This work was supported by the EU FP7 IST Project "CogX" (FP7-IST-215181).

References

Baldridge, J., and Kruijff, G. 2002. Coupling CCG and hybrid logic dependency semantics. In *Proc. ACL 2002*, 319–326.

Blackburn, P. 2000. Representation, reasoning, and relational structures: a hybrid logic manifesto. *Logic Journal of the IGPL* 8(3):339–625.

Brenner, M., and Kruijff-Korbayová, I. 2008. A continual multiagent planning approach to situated dialogue. In *Proceedings of the LONDIAL (The 12th SEMDIAL Workshop on Semantics and Pragmatics of Dialogue)*.

Brenner, M., and Nebel, B. 2008. Continual planning and acting in dynamic multiagent environments. *Journal of Autonomous Agents and Multiagent Systems*.

Brenner, M.; Hawes, N.; Kelleher, J.; and Wyatt, J. 2007. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*.

Dale, R., and Reiter, E. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2):233–263.

Ginzburg, J., and Macura, Z. 2006. Lexical acquisition with and without metacommunication. In Lyon, C.; Nehaniv, C.; and Cangelosi, A., eds., *The Emergence of Communication and Language*. Springer Verlag. 287–301.

Ginzburg, J. 1995a. Resolving questions, I. *Linguistics and Philosophy* 18(5):459–527.

Ginzburg, J. 1995b. The semantics of interrogatives. In Lappin, S., ed., *Handbook of Contemporary Semantic Theory*. Blackwell.

Grosz, B., and Kraus, S. 1999. The evolution of shared plans. In Rao, A., and Wooldridge, M., eds., *Foundations and Theories of Rational Agency*. Springer. 227–262.

Jacobsson, H.; Hawes, N.; Skocaj, D.; and Kruijff, G. 2007. Interactive learning and cross-modal binding – a combined approach. In *Language and Robots: Proceedings of the Symposium*, 1pp–1pp.

Kelleher, J., and Kruijff, G. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 1041–1048.

Krahmer, E.; van Erk, S.; and Verleg, A. 2003. Graph-based generation of referring expressions. *Computational Linguistics* 29(1):53–72.

Kruijff, G.; Kelleher, J.; Berginc, G.; and Leonardis, A. 2006a. Structural descriptions in human-assisted robot visual learning. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI'06)*.

Kruijff, G.; Zender, H.; Jensfelt, P.; and Christensen, H. 2006b. Clarification dialogues in human-augmented mapping. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI'06)*.

Kruijff, G.; Lison, P.; Benjamin, T.; Jacobsson, H.; and Hawes, N. 2007a. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Language and Robots: Proceedings from the Symposium (LangRo'2007)*.

Kruijff, G.; Zender, H.; Jensfelt, P.; and Christensen, H. 2007b. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems* 4(2).

Kruijff, G.; Lison, P.; Benjamin, T.; Jacobsson, H.; Zender, H.; and Kruijff-Korbayová, I. 2009. Situated dialogue processing for human-robot interaction. In Christensen, H.; Kruijff, G.; and

Wyatt, J., eds., *Cognitive Systems*. Available at http://www.cognitivesystems.org/cosybook.

Kruijff, G.; Brenner, M.; and Hawes, N. 2008. Continual planning for cross-modal situated clarification in human-robot interaction. In *Proceedings of the 17th International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*.

Larsson, S. 2002. *Issue-Based Dialogue Management*. Phd thesis, Department of Linguistics, Göteborg University, Göteborg, Sweden.

Li, S.; Wrede, B.; and Sagerer, G. 2006. A computational model of multi-modal grounding. In *Proc. ACL SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL 2006*, 153–160.

Lochbaum, K.; Grosz, B.; and Sidner, C. 1999. Discourse structure and intention recognition. In Dale, R.; Moisl, H.; ; and Somers, H., eds., *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. New York: Marcel Dekker.

Paraboni, I.; van Deemter, K.; and Masthoff, J. 2007. Generating referring expressions: Making referents easy to identify. *Computational Linguistics* 33(2):229–254.

Purver, M.; Ginzburg, J.; and Healey, P. 2003. On the means for clarification in dialogue. In Smith, R., and van Kuppevelt, J., eds., *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*. Kluwer Academic Publishers. 235–255.

Purver, M. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. Dissertation, King's College, University of London.

Thomaz, A. L. 2006. *Socially Guided Machine Learning*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Zender, H.; Kruijff, G.; and Kruijff-Korbayová, I. 2009. A situated context model for resolution and generation of referring expressions. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 126–129.

Continual Collaborative Planning for Situated Interaction

Michael Brenner
Albert-Ludwigs-Universität
Freiburg, Germany

Geert-Jan Kruijff
DFKI GmbH
Saarbrücken, Germany

Ivana Kruijff-Korbayová
DFKI GmbH
Saarbrücken, Germany

Nick Hawes
School of Computer Science
University of Birmingham, UK

Abstract

When several agents are situated in a common environment they usually interact both verbally and physically. Human-Robot Interaction (HRI) is a prototypical case of such situated interaction. It requires agents to closely integrate dialogue with behavior planning, physical action execution, and perception. The paper describes a framework called Continual Collaborative Planning (CCP) and its application to HRI. CCP enables agents to autonomously plan and realise situated interaction that intelligently interleaves planning, acting, and communicating. The paper analyses the behavior and efficiency of CCP agents in simulation, and on two robot implementations.

1 Introduction

When agents try to jointly solve a task in a shared environment, they typically interact with each other and the environment in a variety of ways. They see, they say, they act. These modes of interactions are closely tied. Dialogue is typically about the situation, about plans to be executed, things that happened before. That makes it possible for these modes to complement each other, and, where possible, to substitute each other. For example, a physical action can serve as communicative feedback, to acknowledge that an instruction was understood. The problem is how to bring such a close coupling about.

The paper investigates this problem in the context of cognitive architectures for robot assistants. These robots can sense their environment, they build up models of where things are and what you can do there, and then use these models to talk to humans about things they (the robots) should do. To address the problem we propose to use a framework called *Continual Collaborative Planning* (CCP). CCP makes it possible to interleave planning, sensing, acting, and interacting (DesJardins et al., 1999) – and to make explicit to what extent planned actions and interactions are contingent on what the robot knows right now. As the execution of a plan unfolds, these contingencies (modeled as "assertions") can trigger revision of the plan (i.e. the robot's goals), or a further detailed of the plan (making use of knowledge the robot has acquired more recently). This dynamics yields a pro-active behavior in the robot: It actively tries to gather information to establish how to balance its goals and their execution against the dynamics of a collaborative setting.

The paper describes the framework, and shows experiments with an actual implementation on two robot platforms (one for object manipulation, one for an office assistant), and in a simulated environment (MAPSIM). The experiments focus on the effects of dynamic sensing and acting on the nature of interaction between a robot and a human. Figure 1 shows a typical example (MAPSIM). The experiments quantify how, in domains of different complexity, the dynamics arising from the continual nature of interleaving sensing, acting and interacting provide for more task-efficient behavior than if no such dynamic interleaving were possible.

Overview. §2 introduces the multiagent planning formalism, and discusses its suitability for dialogue planning. §3 presents the CCP algorithm. §4 presents the simulation environment MAPSIM and quantitative experiments we have run in it to evaluate our approach. §5 discusses the implementation of CCP in the robot systems. We discuss related work in §6 and conclude.

2 Representation

When planning for situated interaction an agent needs to represent its physical environment as well as other agents, it must reason about about their (mu-

(1) Anne: 'Please bring me the beer, R2D2.'
(2) R2D2: 'Okay, I'll try, Anne.'
(3) R2D2 moves to the hall.
(4) R2D2: 'Where is the beer, Bill?'
(5) Bill: 'The beer is in the kitchen, R2D2.'
(6) R2D2: 'Thanks.'
(7) R2D2 moves to the living room.
(8) R2D2: 'Please open the kitchen door, Anne.'
(9) Anne opens the kitchen door.
(10) R2D2: 'Thanks.'
(11) R2D2 moves to the kitchen.
(12) R2D2 grasps the beer.
(13) R2D2 moves to the living room.
(14) R2D2 gives Anne the beer.
(15) Anne: 'Thanks for bringing me the beer, R2D2.'

Figure 1: Mixed-initiative interaction involving acting, communication and sensing between three artificial CCP agents in MAPSIM (*Explorer* domain). Sensing actions have been omitted for reasons of space.

(1) Bill goes home.
(2) Bill: "Please bake the pizza, Oven."
(3) Oven: "Okay."
(4) Oven bakes the pizza.
(5) Oven: "I have finished baking the pizza, Bill."
(6) Bill: "Thanks for baking the pizza, Oven."
(7) Bill: "Please bring me the pizza, R2D2."
(8) R2D2: "Okay."
(9) R2D2 brings Bill the pizza.
(10) Bill: "Thanks for bringing me the pizza, R2D2."
(11) Bill eats the pizza.

Figure 2: Dialogue between three artificial agents in MAPSIM (*Pizza* domain).

tual) beliefs as well as their physical, communicative and perceptual capabilities. One language for modelling these elements is the multiagent planning language MAPL (Brenner and Nebel, 2009). Here, we will describe MAPL informally and discuss its suitability for situated interaction; for details of MAPL's semantics see (Brenner and Nebel, 2009).

MAPL is a multiagent variant of PDDL (Planning Domain Definition Language), the de-facto standard language for classical planning (Fox and Long, 2003). One important extension in MAPL is the use of multi-valued state variables (MVSVs) instead of propositions. For example, a state variable *colour(ball)* would have exactly one of its possible *domain* values *red, yellow,* or *blue* compared to the three semantically unrelated propositions *(colour ball red), (colour ball yellow), (colour ball blue)*, all of which could be true in a given STRIPS state. MVSVs have successfully been used in classical planning in recent years (Helmert, 2006), but they also provide distinctive benefits when used for interaction planning. Firstly, MVSVs can be used to model *knowledge* and *ignorance* of agents by adding a special constant *unknown* to the domain of each MVSV. This concept can also be extended to beliefs about other agents' beliefs and mutual beliefs which are modeled by so-called **belief state variables**. Secondly, *wh-questions* can be modeled as queries about MVSVs in our model (see below). Thirdly, algorithms for generating and interpreting *referring expressions* rely on the mutual exclusivity between feature values, as expressed in the MVSV representation.

MAPL **actions** are similar to those of PDDL. In MAPL, every action has a **controlling agent** who executes the action and controls when it is done. Agents are assumed to be autonomous when executing actions, i. e. there is no external component synchronising or scheduling actions by different agent. As a consequence an action will only be executed if, in addition to its preconditions being satisfied, the controlling agent *knows* that they hold. Implicitly, all MAPL actions are extended with such **knowledge preconditions**. Similarly, there are implicit **commitment preconditions**, intuitively describing the fact that an agent will only execute actions if he has agreed to do so.

Three different ways to affect the beliefs of agents, e. g., for satisfying knowledge preconditions, can be modelled in a MAPL domain : sensing, copresence (joint sensing), and communication. All three are MAPL actions with knowledge effects. **Sensor models** describe circumstances when the current value of a state variable can be perceived. **Copresence models** are multiagent sensor models that induce mutual belief about the perceived state variable (Clark and Marshall, 1981). Informally, agents are copresent when they are in a common situation where they can not only perceive the same things but also each other. Individual and joint sensing are important for dialogue because they help *avoiding* it: an agent does not need to ask for what

he sees himself, and he does not need to verbalize what he assumes to be perceived by the other agents as well. Communicative acts currently come in two forms: (i) **Declarative statements** are actions that, similarly to sensory actions, can change the belief state of another agent in specific circumstances. Line 5 of Fig. 2 shows an example of an agent explicitly providing another one with factual information. (ii) **Questions**, **commands** and **acknowledgments** do not have to be modelled explicitly, but are derived from a MAPL domain automatically. They are used in CCP as discussed in Sect. 3.

MAPL **goals** correspond to PDDL goal formulae. However, MAPL has two additional goal-like constructs: **Temporary subgoals** (TSGs) are mandatory, but not necessarily permanent goals, i. e. they must be satisfied by the plan at some point, but may be violated in the final state. **Assertions**, on the other hand, describe *optional* "landmarks", i. e. TSGs that may helpful in achieving specific effects in later phases of the continual planning processes, which cannot be fully planned for yet because of missing information (Brenner and Nebel, 2009). For example, the MAPL domain used to create the simulation in Fig. 1 contains an assertion stating that, informally speaking, to get something one must first know where it is.

MAPL plans differ from PDDL plans in being only *partially ordered*. This is inevitable since we assume that there is no central executive which could guarantee a totally ordered execution. We use the term **asynchronous plans** since MAPL plans also allow for *concurrent* occurrence of actions. Fig. 3 shows an example. An asynchronous plan that guarantees that the implied knowledge preconditions will be satisfied during execution (e. g. by explicitly naming the perceptions to be made and speech acts to be used) is called **self-synchronizing plan** because it "explains" how the agents can coordinate their behavior during execution.

It is often impossible for a group of situated agents to jointly commit to a self-synchronizing plan prior to beginning its execution. As an example, line 1 of Fig. 2 shows how an agent must start executing its individual multiagent plan (i. e. a plan for a group of agents but to which no other agent has committed yet) in order to even get the chance to negotiate the plan with the others: In this scenario, Bill must physically move first because he can only communicate with his household appliances "at home". This is modeled explicitly in the MAPL domain by means of a so-called **communication precondition** that the planner has to satisfy if agents should engage in dialogue. In future work, we will also use communication preconditions to model dialogue-specific requirements like *attention* (Grosz and Sidner, 1986) and *engagement* (Sidner et al., 2005).

3 Continual Collaborative Planning

Continual Collaborative Planning (CCP) agents switch between planning, partial plan execution, monitoring, plan adaptation and communication. Alg. 1 gives a high-level description of the CCP algorithm. Since the current state of the algorithm not only depends on what the agent has been doing, but also on the messages received from others, CCP is specified as a Distributed Algorithm (Lynch, 1996).

Algorithm 1 CCP AGENT(S, G)

$P = \emptyset$
Received no message:
  if S satisfies G do
    return "goal reached"
  else
    $P =$ MONITORINGANDREPLANNING(S, G, P)
  if $P = \emptyset$ then
    return "cannot achieve goal G"
  else
    $(S, P) =$ EXECUTIONANDSTATEESTIMATION(S, P)
Received (tell-val vx) **from** agent a:
  add $v \doteq x$ to S
Received request(e) **from** agent a:
  $sg =$ TRANSLATEREQUESTTOGOAL(e)
  $P =$ MONITORINGANDREPLANNING$(S, G \cup sg, \emptyset)$
  if $P = \emptyset$ then
    send "cannot execute request e" **to** a
  else
    add sg to G as temporary subgoal

We will first discuss the base case when no communication has taken place yet, i. e. the CCP agent has neither sent nor received any messages yet. Roughly speaking, the agent alternates between (re-)planning and acting in this case. The two phases are detailed in Algs. 2 and 3. Alg. 2 shows how a new planning phase is triggered: the agent *monitors* whether his current plan has become invalid due to unexpected (external) events or changes in his goals. If this is the case, the agent adapts its plan by replanning those parts that are no longer executable. In or-

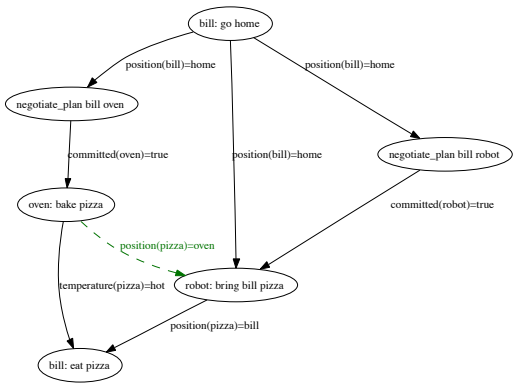


Figure 3: Bill's initial plan for getting pizza.

der to exploit the power of state-of-the-art planning systems, Alg. 2 uses an unspecified classical planner PLANNER to (re-)plan for the obsolete or missing parts of the old plan. The details of this process are irrelevant for the purpose of this paper; it results in an asynchronous plan that specifies actions for (possibly) several agents and the causal and temporal relation between them that is necessary to achieve the planning agent's goal.

Algorithm 2 MONITORINGANDREPLANNING(S, G, P)

```
if res(S,P) ⊉ G
   REMOVEOBSOLETESUFFIXGRAPH(P)
   P' = PLANNER(A, res(S,P), G)
   P = CONCAT(P, P')
return P
```

Fig. 3 shows such an asynchronous plan for the *pizza* scenario of Fig. 2, created with Alg. 2. Note that this plan contains special *negotiation* actions; they will be the triggers for task-orientated subdialogues in a later phase of CCP. The planning algorithm enforces such negotiation actions to be included in a plan whenever this plan includes actions or subplans to be executed not by the planning agent, but by another agent who is not yet committed to this plan. Thus CCP ensures that a (sub-)dialogue will take place that either secures the other agent's commitment or triggers replanning. Note how, in turn, the need for negotiation has forced the planner to include a physical action (Bill's moving home) into the plan in order to satisfy the above communication precondition.

As soon as a CCP agent has found (or repaired) a valid plan it enters the execution phase, described in Alg. 3. First, an action, e, on the first level of the plan, i. e. one whose preconditions are satisfied in the current state, is chosen non-deterministically. If the action is controlled by the CCP agent himself, it is executed. If not, the planning agent tries to determine whether the action was executed by its controlling agent. In both cases, the CCP agent will try to update its knowledge about the world state based on the expected effects and the actual perceptions made (FUSE function).

Algorithm 3 EXECUTIONANDSTATEESTIMATION(S, P)

```
e = choose a first-level event from P
if e ='negotiate_plan with agent a'
   r = SELECTBESTREQUEST(P, a)
   send request(r) to a
else if agt(e) = self then
   EXECUTE(e)
S' = app(S, e)
exp = EXPECTEDPERCEPTIONS(S', A^s)
perc = GETSENSORDATA()
if perc ⊇ exp or exp = ∅ then
   remove e from P
S = FUSE(S', perc)
return (S,P)
```

The most important case for *verbal* interaction is the one where the action chosen to be executed is *negotiate_plan*. This means that a CCP agent A is now in a situation where he is able communicate with another agent B who he intends to collaborate with, i. e. A's plan includes at least one action controlled by B, that B has not yet committed to. In this case, A will send a *request* to B. However, if a plan contains several actions by another agent, i. e. a whole subplan, it is often best not to request execution of the actions individually, but to ask for the end result or, respectively, the final action in the subplan. In other situations it may even be reasonable to request the achievement of subplans that include more than one agent. CCP does not stipulate a specific implementation of SELECTBESTREQUEST; the standard version, REQUESTSUBPLAN, selects the longest possible subplan using only one agent.

When an agent receives a request, Alg. 1 first tests for its individual achievability, i.e., regardless of other goals. If it can in principle be achieved, it is adopted. Accepted requests[1] are adopted as *tem-*

[1]For space reasons, we have omitted the treatment of rejected requests and failed action execution from this presentation. Essentially, agents keep "black lists" that prevent repeti-

porary subgoals (TSGs). This means that they must only be achieved temporarily and do not have to hold any more when the agent's main goal is achieved. The adoption of requests as TSGs is a crucial element of CCP that, to the best of our knowledge, has not been described in other Continual Planning approaches: in addition to repeatedly revising their beliefs about the world, CCP agents also perform continual *goal revision*. In the simplest case, this leads to information-seeking *subdialogues*, as in lines 4–6 of Fig. 1. But newly adopted TSGs also explain why agents engage in subdialogues that mix communicative and physical actions (as in lines 8–10 of the same example).

4 Situated Interaction in Simulation

Studying situated interaction requires environments where agents can physically or virtually act and interact. The same is true for studying continual planning: it needs environments where agents can not only plan, but also execute, monitor and revise their plans. To be able to investigate situated interaction across many application domains and algorithmic variants, we have developed MAPSIM, a *simulation generator* that automatically transforms MAPL domains into multiagent simulations. MAPSIM parses and analyses a MAPL domain and turns it into perception, action, and communication models for CCP agents. During the simulation, MAPSIM maintains and updates the global world state, it uses the sensor models to compute individual and joint perceptions of agents, and it executes MAPL speech acts by passing them on from the sender to the addressee. In other words, MAPSIM interprets the planning domain as an *executable model* of the environment.

MAPSIM and the CCP agents described in this paper have been implemented in Python, integrated with a planning engine in C as a subsolver. The base planner currently used in our implementation is a slightly modified version of Axioms-FF (Thiebaux et al., 2003). MAPSIM includes a basic verbalisation module, called the *reporter* agent, which observes all physical and communicative events in the simulation and verbalises them using a simple recursive template engine. The examples througout the paper were created with the MAPSIM reporter agent. For our robot implementation (cf. next section), we use a specific low-level dialogue planner.

tion of unpromising behaviour.

Task	Turns	Agents	Calls	t Avg	t Tot	Total
prob1	7	2	3	0.02	0.06	1.17
prob2	10	2	5	0.02	0.12	1.67
prob3	13	2	8	0.03	0.25	2.36
prob4	15	3	11	0.05	0.6	4.79
prob5	23	2	15	0.09	1.18	7.34
prob6	27	3	33	0.13	3.71	16.84

Table 1: Experiments in the *Explorer* domain: problem complexity and runtimes (in secs).

For studying whether CCP is suitable for situated interaction and for measuring its performance and scaling behaviour, we conducted several experiments in the *Explorer* domain. This is a simplified service robotics domains (cf., Figure 1 and §5) in which tasks (consisting of a number of agents, interconnected rooms and objects) can become demanding quite quickly. MAPSIM was run on a 2 GHz AMD Athlon with 1 GB RAM. The CCP implementation used was the same that also runs on our robot platform, but sensing and communication was computed and routed by MAPSIM.

Table 1 provides some general information about the experiments and the performance of our CCP implementation. In each task, several autonomous agents interacted in a common simulated environment, automatically synthesized from a short (about 150 lines) MAPL description of the Explorer domain (cf. Figure 1). The knowledge and goals varied over tasks, leading to increasingly complex interactions, roughly measurable by the number of turns it took the agents to achieve their goals. Individual plans are computed quite fast (t Avg) and even in sum (t Tot) are largely dominated by other (yet unoptimised) parts of the simulation (Total). This speed should be competitive with approaches based on plan libraries[2] , while providing the agents with the additional flexibility to react to new situations truly autonomously by finding previously unseen plans on their own.

Table 2 explains in which situations agents change

[2]The planning problems in these tasks usually consist of several hundred distinct facts and instantiated actions, which may lead to enormous search spaces (size may be exponential in the number of facts) and an even greater number of plans. It is not obvious how a system based on a necessarily limited plan librariy can choose an appropriate plan for any given situation.

Task	Reuse	Replanning	Expansion	Goal Change
prob1	9	1	0	1
prob2	10	3	1	2
prob3	10	6	1	3
prob4	20	8	1	3
prob5	16	11	2	5
prob6	20	26	1	5

Table 2: Plan reuse vs. Replanning (and causes for plan invalidation).

Task	Speech Acts	Phys. Actions	Sensings	Wait
prob1	3	4	6	4
prob2	6	4	5	4
prob3	8	5	8	4
prob4	8	7	12	15
prob5	14	9	9	7
prob6	14	13	14	25

Table 3: Different types of actions and their distribution in *Explorer* experiments.

their plans during situated interactions with CCP. Note first that plan reuse usually dominates replanning. However, complex interactions with several agents are so dynamic that plans become obsolete quickly. A major factor here is the adaptation of an agent's *goals* during an interaction (due to requests and questions by others or due to perceptions "activating" conditional goals). Such goal changes (and the resultant replanning) should not be considered as planning failures; instead they enable engagement in previously unforeseen subdialogues and changes changes in initiative (as, e.g., in turns 3 and 6 of Figure 1). Expansion of MAPL assertions, i.e. detailed planning for previously postponed subproblems, only seldom is the cause for replanning. Still, assertions are the major tool that make agents behave proactively in situations where they need to get help or information from others or the environment: without assertions, all tasks except for prob1 become unsolvable for the agents, because they cannot work around their initial lack of knowledge or joint commitment (e.g. turns 3–6 of Figure 1 where the R2D2 postpones detailed planning for Anne's goal until is has got more information from a third agent, Bill). If replanning is neither caused by assertions nor goal changes, it is due to changes in the environment caused by other agents that render the previous plan invalid.

Table 3 shows the distribution of action types throughout interactions. It is interesting that this fairly even distribution and the continual, seamless switching between actions (as can be witnessed in Figure 1), is not enforced anywhere in the CCP algorithm. It arises from the needs of the agents to achieve their individual goals by means of acting, seeing, and communicating. Note also that physical behaviour and sensing sometimes substitute communicative actions (e.g. in turn 8 of Figure 1). Waiting, i.e. not acting deliberately or for want of a plan, occurs in CCP, whenever the next actions in a plan are to be executed not by the agent itself, but by another (cf. Algorithm 3). This "passing" behaviour also leads to some simple, if limited, *turn-taking* mechanism (detailed discussion is beyond the scope of this paper.)

5 Robot Implementations

We have also implemented CCP in two real robotic systems, one for interactive table-top manipulation and one for interactive exploration of an indoor environment. The first robot, called the PlayMate, is stationary and uses a 6-DOF Katana arm to manipulate objects on a table. It can interactively learn about properties of the objects and play small games with a human user standing nearby. The second robot system, the Explorer, is based a mobile platform, an ActivMedia PeopleBot. It interactively builds up an understanding of the spatial and functional organisation of an indoor environment, using automated mapping techniques, and can perform basic tasks for a human user, e. g. finding a object and carrying it back to the human (if a second human puts the object on the robot).

The two robot systems differ significantly in both their hardware and their capabilities. However, they use a common architecture schema, CAS (Hawes et al., 2007), for integrating varying subarchitectures (SAs). In CAS, all SAs are active in parallel, and all operate on SA-specific representations (as is necessary for robust and efficient task-specific processing). These disparate representations are unified by a *binding SA*, which performs abstraction and cross-modal information fusion on the information from the other SAs (Jacobsson et al., 2008), yet stores information about the origins and original representations in a so-called Address-Variable Map (AVM).

The representations produced by the binding SA are sufficiently close to MAPL to enable planning with CCP. The PlayMate and the Explorer robot employ the same planning domains for planning their behaviour that are used for the respective MAPSIM simulations, i. e. the robots continually produced and revise MAPL plans exactly as in the simulation. However, actions are executed not in simulation, but in the real world, by SAs controlling the actuators and sensors of the robot. Each SA whose behaviour is modelled in the planning domain provides a so-called *action dispatcher* which, using the AVM, translates MAPL commands and arguments back into its own representation.

For human-robot interaction it is most important that the robots can *communicate* naturally with humans. Here, the verbalisation engine of MAPSIM is not sufficient. Instead, CCP is integrated with a specialised SA for situated dialogue processing (Kruijff et al., 2007). In this architecture, CCP does the high-level pragmatic reasoning, i.e. it determines how and why to use communication for achieving a (possibly non-communicative) goal. Linguistic situation-appropriate realisation of the planned "verbal behaviour", as well as the situation-aware interpretation of speech acts by other agents, are handled by the specialised dialogue component. CCP provides the dialogue component with a high-level (communicative) goal and additional context information, so that the dialogue component can turn it into an utterance (or even subdialogue) that communicates the provided content.

6 Related Work

The aim of this work is to bring together ideas from Planning, Dialogue, and Multiagent Systems research. We can only discuss a few prototypical inspirations from those fields here.

Our work is close in spirit to models of collaborative *dialogue*, most notably those based on BDI models of collaboration, such as the SharedPlans formalism (Grosz and Kraus, 1996; Lochbaum, 1998; Rich et al., 2001). Similar to (Blaylock et al., 2003) we emphasise the importance of integrating collaborative planning and collaborative plan execution. In contrast to all aforementioned approaches, CCP relies on first-principles planning rather than pre-defined plan libraries (see Section 7). The explicit reasoning about perception and copresence in CCP can explain how agents try to bring about knowledge conditions for joint behaviour, as specified, e. g. in the SharedPlans literature.

Distributed Continual Planning was advocated as a new paradigm for planning in dynamic multiagent environments by desJardins and colleagues (DesJardins et al., 1999). To the best our knowledge, ours is the first principled attempt to apply DCP to situated dialogue and HRI, and also the first DCP approach describing deliberative *goal revision* as part of a DCP algorithm. Planning with sensing has often been described in the planning literature (see, e. g. (Petrick and Bacchus, 2002) for a modern example), but mostly in the context of conditional, rather than continual planning, and, to the best of our knowledge, without including the concept of *copresence* (Lewis, 1969). The form of active execution monitoring used in CCP is also related to the *attentional state* of (Grosz and Sidner, 1986).

Most of the aforementioned work relies on *hierarchical* action and plan representations. In contrast, CCP could be said to decompose planning problems *over time* by actively postponing subproblem solving. The two approaches are not mutually exclusive and CCP could be adapted to hierarchical planning.

Collagen (Rich et al., 2001) is a system for building collaborative interface agents that is based on (Grosz and Sidner, 1986; Grosz and Sidner, 1990), which is domain-independent and has been used for various applications. Collagen's methods for representing the discourse state and doing plan recognition are much more sophisticated than CCP currently. However, Collagen does not (yet) include a first-principles planner, but relies on plan libraries and domain-specific code plug-ins (Rich and Sidner, 2007). It would be interesting to investigate whether CCP can be integrated with Collagen.

Similarly, the most prominent representative of the information-state-update approach to dialogue modeling, GoDiS (Traum and Larsson, 2003), has complementary rather than competing main strengths: GoDiS has a more elaborate repertoire of dialogue moves and can produce more sophisticated dialogue behavior than CCP and MAPSIM, but it uses static plans, and it is not clear how it would combine communication with physical action.

7 Conclusion

This paper has presented a new algorithmic model for situated interaction, Continual Collaborative Planning (CCP). We have shown how mixed-initiative interaction among agents interleaving physical actions, sensing, and communication occurs naturally during CCP. An early empirical investigation has shown that CCP agents can compute, execute and revise situation-specific plans for interaction in real-time.

CCP is a fairly simple distributed algorithm. Most reasoning about necessary actions is encapsuled in the planning algorithm and enforced by the semantics of MAPL. As a result, many desirable features of situated interaction arise without being explicit in the algorithm, as by-products of goal-directed planning, e.g. switching between physical and verbal behaviour or the initiation of subdialogues. Likewise, actions without purpose are silently dropped or substitued, e.g. in the case of implicit acknowledgments. Finally, the causal structure of plans (and histories of plan changes in the case of CCP) permits introspection and verbalisation of past and future planning processes, which is also crucial for our and similar HRI applications.

Planning requires a declarative model of actions, which in the real world, in particular for communication, is often hard to define. (This is probably the main reason for planning-based approaches to interaction having lost relevance in the last decade.) CCP tries to accomodate for that by relying on executing monitoring as much as on prediction, and by enabling the agent to actively suspend the planning process. For situated interaction, this leads to particularly dynamic, mixed-initiative behaviour because CCP agents will rather engage in dialogue or sensing instead of trying to plan with an incomplete model.

Many extensions to CCP are possible and planned for the future, e. g. the inclusion of goal and plan recognition techniques (for proactively supporting collaborators or preventing execution conflicts) and reasoning about the presuppositions underlying the speech acts of others. We will experiment with these variants of CCP in simulation first, then evaluate them in HRI on our robot systems.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement No. 215181, CogX.

References

N. Blaylock, J. Allen, and G. Ferguson. 2003. Managing communicative intentions with collaborative problem solving. In *Current and New Directions in Dialogue*. Kluwer.

M. Brenner and B. Nebel. 2009. Continual planning and acting in dynamic multiagent environments. *Journal of Autonomous Agents and Multiagent Systems*. to appear.

H. H. Clark and C. R. Marshall. 1981. Definite reference and mutual knowledge. In *Elements of discourse understanding*. Cambridge University Press.

M. DesJardins, E. Durfee, Jr. C. Ortiz, and M. Wolverton. 1999. A survey of research in distributed, continual planning. *The AI Magazine*.

M. Fox and D. Long. 2003. PDDL 2.1: an extension to PDDL for expressing temporal planning domains. *JAIR*.

B. J. Grosz and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence*, 86.

B. J. Grosz and C. L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3).

B. Grosz and C. Sidner. 1990. Plans for discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, MA.

N. Hawes, A. Sloman, J. Wyatt, M. Zillich, H. Jacobsson, G. J. Kruijff, M. Brenner, G. Berginc, and D. Skočaj. 2007. Towards an integrated robot with multiple cognitive functions. In *Proc. AAAI '07*, Vancouver, Canada.

M. Helmert. 2006. The Fast Downward planning system. *JAIR*, 26:191–246.

H. Jacobsson, N. Hawes, G. J. Kruijff, and J. Wyatt. 2008. Crossmodal content binding in information-processing architectures. In *HRI-08*.

G. J. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes. 2007. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Language and Robots: Proceedings of the Symposium*, Aveiro, Portugal, December.

D. Lewis. 1969. *Convention. A Philosophical Study*. Harvard University Press, Cambridge, Massachusetts.

K. E. Lochbaum. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*.

N. Lynch. 1996. *Distributed Algorithms*. Morgan Kaufmann, San Francisco, CA.

R. Petrick and F. Bacchus. 2002. A knowledge-based approach to planning with incomplete information and sensing. In *Proc. AIPS-02*.

C. Rich and C. L. Sidner. 2007. Generating, recognizing and communicating intentions in human-computer collaboration. In *AAAI Spring Symposium on Intentions in Intelligent Systems, Stanford, CA*.

C. Rich, C. L. Sidner, and Neal Lesh. 2001. Collagen: applying collaborative discourse theory to human-computer interaction. *The AI Magazine*, 22(4).

C. L. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich. 2005. Explorations in engagement for humans and robots. *AIJ*.

S. Thiebaux, J. Hoffmann, and B. Nebel. 2003. In defense of axioms in PDDL. In *Proc. IJCAI*.

D. Traum and S. Larsson. 2003. The information state approach to dialogue management. In *Current and New Directions in Discourse and Dialogue*. Kluwer.

Abduction for clarification in situated dialogue

Geert-Jan M. Kruijff and Miroslav Janíček

DFKI GmbH, Saarbrücken, Germany
{gj, miroslav.janicek}@dfki.de

Abstract

A robot can use situated dialogue with a human, in an effort to learn more about the world it finds itself in. When asking the human for more information, it needs to be clear to the human, what the robot is talking about. The robot needs to make transparent what it would like to know more about, what it does know (or doesn't), and what it is after. Otherwise, the human is less likely to provide a useful answer to the robot. They need to establish a common ground in. The paper presents ongoing research on developing an approach for comprehending and producing (sub-)dialogues for clarifying or requesting information about the world in which establishing common ground in beliefs, intentions, and attention plays an explicit role. The approach is based on Stone & Thomason's abductive framework [42–44]. This framework integrates intention, attentional state, and dynamic interpretation to abductively derive an explanation on what assumptions and intentions communicated content can be interpreted as updating a belief context. The approach extends the framework of Stone & Thomason with assertions, to provide an explicit notion of checkpoint, and a more explicit form of multi-agent beliefs [7]. The approach uses these notions to formulate clarification as a continual process of comprehension and production set in dialogue as a collaborative activity.

1 Introduction

Robots need to continuously learn. They do not always know everything, or understand everything. So they can ask: we make it possible for a robot to communicate with humans, to learn more. For such dialogue to be effective, the human and the robot need to form a mutual understanding of what is being talked about, and why. Recent theories focus on how this mutual understanding can come about through alignment [33,14]. Agents align how they communicate content, what they pay attention to, and what they intend to do next. They base this alignment on how they

[1] The research reported here was performed in the EU FP7 IP "CogX: Cognitive Systems that Self-Understand and Self-Extend" (ICT-215181); http://cogx.eu

perceive each other's perspective on the world – situatedness, attention, intention, capabilities, current cognitive and emotional state (cf. e.g. [39,12,40]).

This works out reasonably well as long as we can assume a more or less common way of "looking" at things. Even when humans normally differ in what they know, can, and intend to do, there is typically a common categorical framework in which they can try to characterize the world, to arrive at a common ground. But this is where a problem arises in communication between a human, and a robot that continuously learns. Because that robot is not just learning more about instances in the world, possibly using a predefined human-like ontology. Ultimately, it will also be forming category systems for thinking about these instances, based in patterns that arise from its own ways of perceiving reality. And those may well be substantially different from how humans see things.

Which is why mechanisms for clarification, and information requests in general, are necessary for situated dialogue. Humans and robots interacting with each other need to be able to ask when something is not clear. As Clark already indicated, this covers a broad range of possible unclarities. Clarification typically covers misunderstanding at the purely linguistic level, but extends all the way to not being able to understand how an utterance relates to the (situated) context it pertains to. In human-robot interaction (HRI), and dialogue systems in general, clarification has primarily focused on linguistic misunderstanding [46,34,27,36,35,28]. Relatively few HRI systems extended clarification to also include misunderstanding or lack of understanding relative to a situated context [45,26,25]. Trafton et al. [45] deal with ambiguities in, or absence of, suitable object antecedent for references. Kruijff et al. [26,25] more focus on clarification-as-information request, using situated dialogue for a robot to obtain more information about particular aspects of the environment. In this paper we describe a computational approach to clarification which aims to deal with the continuum indicated by Clark.

The approach is based on an extension of Stone & Thomason's abductive framework [42–44]. In this framework, comprehension and production of dialogue are based in the construction of an abductive proof. Abduction reasons towards an explanation consisting of a consistent context update and possible changes to attentional state. The explanation is based on factual assumptions, observations, and inferred intentions – all included at a context-sensitive cost. The resulting framework thus places belief context, attentional state, and intention on a par. This is in idea comparable to other intentional approaches to dialogue and discourse, like Grosz & Sidner's [20]. Stone & Thomason's approach arguably provides more flexibility [44] in that aspects such as reference resolution are dynamically determined through proof, rather than being constrained by hierarchical composition of a context model. (This particularly applies to a comparison with approaches such as SDRT [3].) For comprehension an abductive proof provides the conditions under which an agent can update her belief model and attentional model with the content for a communicated utterance, and her task model using the inferred intentions un-

derlying the utterance. For production an abductive proof provides the conditions for executing a plan to achieve an intended context- and attentional state update in another agent.

The approach we present here extends Stone & Thomason's framework in several ways. We expand Stone & Thomason's context [42] to incorporate the types of situated multi-agent beliefs and tasks the robot reasons with in understanding collaboration, and the world as such. Furthermore, we make Stone & Thomason's notion of "checkpoints" more explicit. Stone & Thomason propose to use checkpoints, a communicative means to establish whether assumptions are in fact warranted [44]. Checkpoints introduce a relation between the construction of an explanation, and acting on it. This suggests a similarity to the construction of a plan and the monitoring of its execution, as found in continuous multi-agent planning [7]. Brenner & Nebel [7] introduce a notion of assertion for continual planning. An assertion poses the availability of future observations, to enable the construction of a continual plan including actions based on such an assertion. Upon execution, assertions are checked and are points for possible revision or extension of a plan.

We propose to use a similar notion. In an abductive proof, we can include assumptions, observations, and actions at varying costs to infer an explanation. They all contribute facts or outcomes from which further inferences can be drawn. An assertion is a statement whose truth we need to assume, but which we cannot prove or disprove on the current set of beliefs of the agent. Marking assertions turns these statements in an abductive proof into points that warrant explicit verification – i.e. they act as checkpoints. Checkpoints make explicit how acting upon an abductive proof turns into a form of execution that is similar to continual plan execution. The notions of assertion and checkpoint provide the approach with a fundamental way for dealing with clarification. While constructing a proof, the inclusion of an assertion can itself trigger a clarification process to verify its validity – or it can be used to assert the positive outcome of a clarification process itself. The former case concerns assumption-turned-assertion, whereas the latter case regards an assertion about a future observation.

Taken together, the approach treats clarification as a process. A proof constructs how the content of a request is based in private and shared beliefs and intentions, and in an attentional state. It links the request and its expected answers, modeling the latter as assertion. This turns the outcome of a request into an explicit checkpoint, and with that, clarification into a process of requesting, answering, and verifying the answer. Should the verification of the answer fail, a proof can be expanded or reconstructed. The resulting approach formulates clarification as a continual process of comprehension and production set in dialogue as a collaborative activity [44].

An overview of the paper is as follows. §2 discusses Clark's notion of grounding, and computational approaches of that notion. §3 presents a formalization of the

approach, with §4 discussing a prototype implementation and integration of the approach into a cognitive system. The paper ends with a discussion of a wide range of examples, and conclusions.

2 Background

In this section we look into the concept of common ground, the process of adding to the common ground – *grounding* and its computational modeling. We also examine the relation between grounding and clarification and briefly touch the topic of surface realizations of clarification requests.

2.1 Common ground and grounding

Following Clark ([10,11]), we define common ground as a set of mutual, common or joint knowledge, beliefs and suppositions of a group of agents C.

Definition 1 (Common ground (shared basis)) p *is* common ground *for members of community* C *if and only if*

(1) every member of C *has information that basis* b *holds (*b *is referred to as* shared basis*),*
(2) b *indicates to every member of* C *that every member of* C *has information that* b *holds,*
(3) b *indicates to members of* C *that* p*.*

Clark distinguishes two broad sources of shared bases: *communal* and *personal*. Communal bases are based on a group member's membership in cultural communities (such as nationality or the accent of English the member speaks) and imply the features, facts and beliefs about the individual commonly ascribed to him/her by other agents. Personal bases originate from joint personal experiences within the group. Personal bases comprise both joint perceptual experiences about the situation and joint actions.

In order to add to common ground through joint actions, agents have to look for evidence of shared bases in signals that other agents display.

In joint actions, the agents act together to achieve a common goal. Since this can only be successful if the agents have common ground, the agents actively work on its establishment. This process is called *grounding*. Common ground needs to be established *well enough for current purposes* at all levels of communication. The definitions of the levels vary across theories (see e.g. [41,46,10,2]).

In Clark's view, it has to be established at all levels of what he calls *joint action ladder*, the following hierarchy of causally related actions:

Level	**Speaker A's actions**	**Hearer B's actions**
1	executing behavior t	attending to behavior t
2	presenting signal s	identifying signal s
3	signalling that p	recognizing that p
4	proposing joint project w	considering the proposal of w

Lower levels provide evidence to the upper ones and are the required for the upper levels' completion.

2.1.1 Closures

In general, agents performing an action require evidence that they have succeeded in performing it. The optimal evidence usually isn't the strongest, most economical and most timely evidence possible, because that may be too costly. Clark formulates these principles of closures:

- *principle of least effort* – all things being equal, agents try to minimize their effort in doing what they intend to do.
- *principle of opportunistic closure* – agents consider an action complete just as soon as they have evidence sufficient for current purposes that it is complete.
- *principle of holistic evidence* – evidence that an agent succeeded on a whole action is also evidence that the agent succeeded on each of its parts.
- *principle of joint closure* – agents try to establish shared basis for the mutual belief that they have succeeded well enough for the current purposes.

2.1.2 Contributions

In a conversation, *contributions* to the conversation are seen by Clark as joint actions aimed towards the successful understanding of the displayed (uttered) signals. The contributions, according to the joint closure principle, are divided into two phases:

- the *presentation phase*, where A presents the content to B, the underlying assumption being that if B gives sufficient evidence e of its acceptance, A can believe that B understands what A has meant;
- the *acceptance phase*, where B gives evidence e' that she understands the message. B's action is based on the assumption that this evidence is required by A to believe that B has understood.

The evidence required in a dialogue can be classified as follows:

- *continued attention*, showing that the hearer is satisfied with the speaker's presentation;
- *initiation of the relevant next contribution*; [2]
- *acknowledgment*, by saying "OK" or the like;
- *demonstration* of what B has understood the message meant;
- verbatim *display* of all or a part of A's presentation.

Note that every acceptance except for the *continued attention* is also a contribution to the dialogue (in the opposite direction) and therefore also has a presentation phase.

2.2 Computational models of grounding

The descriptive and off-line nature of Clark's theory of grounding disqualifies it from direct use in dialogue systems. In practical applications, a more computationally-oriented model is needed.

In non-situated dialogue systems, models of linguistic grounding have been studied in detail, see e.g. [46,27,35]. In human-robot interaction, however, they have only been studied to a limited degree. [28] present a basic model for Clark-style communicative grounding in human-robot interaction, adopting an approach very similar to [46]. [26] discuss an approach to clarification in human-augmented mapping, making use of the Question-Under-Discussion mechanism of [16,17,27].

In this section, we sketch out David Traum's finite-state model ([46]) and the situated multi-modal model of Li et al. ([28]) and relate them to Clark's contribution model.

2.2.1 Traum's finite-state model

The Grounding Acts model, introduced by David Traum (see [46,47]), is a computational, prescriptive and non-situated on-line reformulation of Clark's theory of contributions. Its cornerstones are *grounding acts*, interpretations of *discourse units* with a specific function towards grounding of the units to effectively model the the presentation and acceptance phases of the contribution model.

The grounding acts, distinguished by whether they are performed by the initiator I of the discourse unit or the responder R, are the following:

- $\mathrm{initiate}$ – initiate a new discourse unit;

[2] Which, of course, implies the hearer's understanding of the speaker's contribution.

- $\mathrm{continue}$ – continue the discourse unit, i.e. further specify its content (in Clark and Schaefer's terminology);
- $\mathrm{acknowledge}$ – signal understanding of the discourse unit to the other party; [3]
- repair – correct a misunderstanding in the discourse unit's content;
- $\mathrm{request\_repair}$ – signal a lack of understanding;
- $\mathrm{request\_acknowledgment}$ – signal a request of acknowledgment by the other party;
- cancel – abort the process of the discourse unit's grounding.

Traum avoids treating each grounding act as a Clark-style presentation, eliminating the need for each of them to have its own acceptance phase. This allows him to devise a *finite-state* model: each grounding act changes the grounding state of the underlying discourse unit, possibly to the same state. There are seven grounding states, the initial state is S, the desired, grounded, state is F:

- *state* S – initial state
- *state* 1 – acknowledgment by R required for grounding
- *state* 2 – need for repair by I and acknowledgment by R
- *state* 3 – need for acknowledgment by I
- *state* 4 – need for repair by R and acknowledgment by I
- *state* F – grounded
- *state* D – dead state (grounding either abandoned or failed)

From the viewpoint of Clark's theory, once a discourse unit is grounded, i.e. its grounding state is F, its acceptance phase is successfully finished.

The Grounding Acts model does not require any particular size of the discourse units to be grounded. However, the size of the units determines what behavior can be modeled. In [47], Traum indicates that most work has been done with intonation phrases as grounding units, but that smaller as well as larger units have been used.

2.2.2 Stack-based situated model of Li et al.

Li et al. ([28]) build on Traum's work and present a situated multi-modal grounding model for human-robot interaction.

Their model uses *exchanges* ([9]) as grounding units. Exchange is a pair of "contributions" – speech acts initiated by the two agents engaged in the dialogue. The first act of the exchange is called Presentation, the second is called Acceptance. Note that Li et al. use the term "contribution" in a different way than Clark does: for Clark, a contribution comprises *both* acceptance and presentation.

[3] Traum uses "acknowledgment" to cover the entire spectrum of positive signals of understanding that the hearer issues.

Similarly to Traum, instead of treating each Acceptance as a new presentation as Clark does, Li et al. explicitly model the embedding of (clarification) subdialogues by organizing exchanges in a stack operated by a push-down automaton augmented so as to allow transitions trigger actions and push or pop a variable number of exchanges in one step. They identify four relations of the grounding status of individual exchanges stored in the stack to the grounding of the stack as a whole:

- default – the Presentation defines a new account independent to the previous ones;
- support – specified when an agent is unable to provide an Acceptance for the given Presentation. In such case a new exchange is initiated to support the grounding of the given Presentation. (For instance, by saying "What?");
- correct – a correction of the previous Presentation;
- delete – abandon the grounding process of the previous Presentation.

The (Clark's) acceptance phase is successfully finished when there is an Acceptance (Li et al's) available. The Acceptance may also be an implicit one, as in Clark's *continued attention*. Grounded units are immediately removed from the stack.

The contributions are structured as *interaction units* (IUs) that consist of two layers: conversational and intention layer. The conversation layer component comprises verbal and non-verbal realizations of the intention represented in the intention layer. This allows the handling of deictic expressions such as "this box" in a systematic way and therefore modeling a multi-modal dialogue.

For each contribution received from the other dialogue participant, it is first attempted to determine the underlying intention by examining the verbal content of the IU; if that fails, the non-verbal component is examined. If the intention is recognized and found conforming to the current joint goal, a new acceptance contribution is generated and the presentation is deemed to be grounded. If the intention is not recognized, a new contribution with the relation support or correct is pushed onto the stack.

2.3 Clarification as a means of grounding

To overcome a breakdown in communication, people relatively often use clarification (sub)dialogues rather than starting the whole conversation again as that would not waste their resources (following the principle of least effort in Section 2.1).

Clarification, corresponding to Traum's concept of "repair" and the support/correct grounding relation in Li et al.'s approach mentioned above, can take many forms – be that relative to the form of the utterance, its meaning and the level of understanding that has been achieved. In the following, we examine the range of clarification

requests that may be raised, e.g. by a robot in a human-robot interaction scenario, depending on the source of communication breakdown represented as the achieved level in Clark's joint action ladder and give an example classification of their surface forms based on a corpus study.

A communication breakdown may occur at any level of the joint action ladder, which corresponds to different levels of understanding and interpretation of the utterance. The following classification is based on [38] and extended by the notion of visual recognition problems:

Breakdowns on level 1 (execution / attention) are typically *channel problems*, where there is a lack of contact between the parties [38].

Examples:

- The hearer (B) doesn't notice that the speaker(A) is talking to her;
- A can't talk to B since B is looking somewhere else.

Breakdowns on level 2 (presentation / identification) can arise due to *acoustic problems* [38]. A part of the speech signal was not heard or not recognized or there is an uncertainty in the word recognition such as multiple similarly ranked hypotheses in the output of the recognition.

- "Pick up the red $\langle$*noise*$\rangle$" or "$\langle$*noise*$\rangle$ the red ball" or "$\langle$*noise*$\rangle$";
- "Move the {cup | cub}".

Breakdowns on level 3 (signaling / recognition) arise when we fail to process a signal – whether linguistically, or as part of maintaining situation awareness. Linguistically, we distinguish *lexical problems* [38], when the meaning of one or more words in the utterance is not known, and *parsing problems* [38]. A parsing problem regards an ambiguity in the syntactic structure of the utterance that results in ambiguity of its semantic interpretation.

- The hearer knows the word "football" but does not have any concept or visual image associated with it (lexical problem).
- "bring the mug on top of the shelf" which can either concern a mug on the shelf or a mug somewhere else that should be put onto a shelf (parsing problem)
- "the mug with a flower" which can either refer to a mug with a flower in ir or a mug with a flower painted on it (parsing problem).

Similarly, we might face a problem in processing signals in other modalities. A *perceptual recognition problem* arises when the agent is uncertain about what it

perceives, or when it does not recognize it. (Note that we understand perception here in a broad sense – not just visual perception, but also e.g. laser range-finder based perception and classification.)

- A robot sees an object, but is unable to recognize it as a mug;
- The robot can't see a door in the room.

Another possibility is that the agent may fail in *reference resolution* [38]. There is an ambiguity or failure in the resolution of the intended referent of a natural language expression. This may concern the referents of NPs and deictic expressions (with or without gestures), as well as action references. This problem can thus go beyond the realm of dialogue, as an agent may fail to ground a reference in her world model(s).

- "the red ball" when there are multiple red balls on the table, or when there is none;
- "the supermarket" when there are multiple (or no) supermarkets in the relevant region;
- "this" + pointing when it's not clear which object is meant;
- "Monday the first" when the month has not been specified.

Finally, a *belief conflict* can arise [38]. There may be a problem with the validation of a proposition against the agent's beliefs. (This might perhaps be seen as just another case of reference resolution problem.) Presupposition failures, apart from referential presuppositions which are handled above, would also belong here.

- "You'll need a visa" when the hearer doesn't think so ([37]);
- "There are three balls on the table" when the robot can only see two;
- "This is a ball" when the robot is perceiving a square object;
- "Put the ball on the table" presupposes that the ball is not on the table yet; if it is, a conflict arises.

Breakdowns on level 4 (proposal / consideration) arise as problems in *intention recognition or -evaluation* [38]. It is not clear why the utterance with that interpretation has been uttered or why an action in the physical world has taken place, i.e., how it is relevant to the interaction, what contribution it makes to the goal(s), if there are any.

- "This is a phone" when one is playing the color game;
- "Do you know where the printer is?" or "Go to the printer room" when the current task is to get a cup of coffee;
- if the human places an object in front of the robot without saying anything, the robot may not know what is expected from it, or when the human is moving objects, the robot may not know why s/he's doing it.

3 Approach

In this section we discuss an abductive approach to comprehension and production of clarification dialogues. The approach is based on Stone & Thomason [42–44]. We extend their approach with *assertions* (§3.2), and a more explicit notion of *multi-agent beliefs* (§3.3). Assertions introduce explicit checkpoints in a proof. At these checkpoints, an assertion needs to be verified against the current context. Should this verification fail, a proof can be expanded to overcome this failure (identifying which actions to undertake), or require reconstruction. This lends the approach a continual nature [7], making abduction for clarification part of a larger, continual form of collaborative activity (§3.4).

3.1 Abduction: Stone & Thomason's approach

Stone & Thomason propose a contextualized form of weighted abduction, for producing and comprehending dialogue [42,43]. The abductive inference is set within the broader model of collaborative activity [44]. This model makes explicit how the proofs for comprehension and production interact with actual steps for acting upon these proofs. Below, we first focus on the definitions for Stone & Thomason's form of abduction. In subsequent sections we provide definitions of the formal extensions we propose.

Definition 2 (Abductive modal inference [42]) *Contextual reasoning is phrased as a modal logic. Modal operators of the form* $[c]$ *are associated with contexts. A distinguished operator* $\Box$ *specifies an axiom to be true in all contexts.* A *schematizes atomic formulas;* $\top$ *identifies the always-true atom.*

If κ *is a sequence of context operators of the form* $[c_0]...[c_n]$ *(possibly empty) then the notation* $\kappa(\phi)$ *is used to name the formula* $[c_0]...[c_n]\phi$*.* $\kappa \circ \kappa'$ *denotes the concatenation of* κ *and* κ'*. Goals are modalized atomic formulas, clauses* P *are modalized Horn clauses whose antecedent and conclusion formulas may themselves be modalized:*

$$\begin{aligned} G ::= & \kappa(A) | \top \\ P ::= & \kappa'(\kappa_1(A_1)...\kappa_m(A_m) \rightarrow k''(H)) | \\ & \Box\kappa'(\kappa_1(A_1)...\kappa_m(A_m) \rightarrow k''(H)) \end{aligned}$$

During proof construction each subgoal is associated with an assumability function f_j *to indicate assumption costs:*

$$(\Box)\kappa'(G_1/f_1. \wedge .. \wedge G_m/f_m \rightarrow k''(Q))$$

An abductive proof for a query Q *is a sequence of lists for whose initial element is*

$Q[unsolved]$, in which transformation rules (truth, assumption, resolution, factoring) transform this list successively into a final list in which none of the elements are marked as unsolved. The proof determines an answer to the query as a pair $\langle Q_0, \mathbb{D} \rangle$ consisting of an instantiation Q_0 of Q and a set of postulated assumptions Δ . Δ is a multiset consisting of a formula $\kappa(A)$ for each element of the form $\kappa : A[assumed]$ in the terminal list. □

The transformation rules for abductive proof construction are defined as follows.

Definition 3 (Transformation rules [42]) *Given a state in the construction of an abductive proof, represented as a list $L = Q_1, ..., Q_n$ with $Q_i = \kappa : A$ the left-most query marked as* unsolved*. This state can then be transformed using one of the following* transformation rules*:*

Truth *If A is $\top$, derive a new state exactly like L except that the label of Q_i is* resolved *rather than* unsolved*.*

Assumption *Derive a new state exactly like L except that the label of Q_i is* assumed *rather than* unsolved*.*

Resolution *Select a clause R of the form P or $\Box P$, where P is*
$\kappa'(\kappa_1(A_1)...\kappa_m(A_m) \rightarrow k''(H))$
with its variables renamed, so that it has no variables in common with L. Suppose H and A are unifiable with a most general unifier $\sigma, \kappa = \kappa^* \circ \kappa' \circ \kappa$", and unless R is $\Box P$ then k^* is empty. Then derive the new state:
$Q_1\sigma, ..., Q_{i-1}\sigma,$
$\kappa = \kappa^* \circ \kappa' \circ \kappa_1 : A_1\sigma[unsolved], ..., \kappa = \kappa^* \circ \kappa' \circ \kappa_m : A_m\sigma[unsolved],$
$\kappa : A\sigma[resolved], Q_{i+1}, ..., Q_n\sigma$

Factoring *Suppose some element Q_s describing a query $\kappa : H$ precedes Q_i in L, and H and A are unifiable with most general unifier σ. Then derive a new state suppressing the duplicate proof of Q_i: $Q_1\sigma, ..., Q_{i-1}\sigma, Q_{i+1}\sigma, ..., Q_n\sigma$.*

□

In [43] Stone & Thomason identify four types of context κ: c is a context, i represents old information in c, e represents new information about events, and a is the attentional component. For understanding the utterance 'he left' Stone & Thomason provide the following illustration:

$$
\begin{array}{ll}
a_1\text{:} & he(X) \\
e_1\text{:} & utter(\text{'he left'}, E, P) \\
e_1\text{:} & do(E) \\
i_1\text{:} & leave(P, X) \\
\hline
c_1\text{:} & add{-}info(P, i_1, i_2) \\
c_1\text{:} & put{-}in{-}focus(X, a_1, a_2)
\end{array}
$$

The proof assumes that $he(X)$ can be resolved against the current attentional context with low cost if X is masculine, singular and in focus. The event e_1 is postulated on observation, and the information i_1 is proven given by the grammar, which provides the logical form as a way of expressing P. The inference about the intended effects of communication (i.e. of the intention $do(E)$) is the addition of P to the context (i.e. a context update), and the maintaining of X in focus.

3.2 Assertions in abductive inference

We expand Stone & Thomason's definition of abductive modal inference with a notion of *assertion*. The point of an assertion is to provide an explicit checkpoint in a proof. Brenner & Nebel [7] define a notion of assertion for continual planning, in which an assertion can explicitly trigger replanning or plan expansion. We propose to do something similar. Intuitively, the truth of an assertion depends on the outcome of an action, or on a future observation verifying the asserted proposition. Such verification occurs during the execution of the actions, based on the interpretation of a proof as a plan [44]. Should an assertion not be confirmed, we can revisit the assertion in the proof and check how to adapt the proof – akin to continual planning.

We propose a notion of assertion for abduction, based on *test actions* $\langle F \rangle ?$ [5]. Baldoni et al. [5] specify a test as a proof rule. In this rule, a goal F follows from a state $a_1, ..., a_n$ after steps $\langle F \rangle ?, p_1, ..., p_m$ if we can establish F on $a_1, ..., a_n$ with answer σ and this (also) holds in the final state resulting from executing $p_1, ..., p_m$. Using the notion of context as per Definition 2, a test $\kappa : \langle F \rangle ?$ means we need to be able to verify F in context κ. If we only use axioms A, testing is restricted to observability of facts. By extending Definition 2 with *embedded implications*, we can also make tests range over obtainable outcomes. An embedded implication is an implication $D \rightarrow C$ as part of a Horn clause. [4] show how embedded implications can be included in abduction; (see [8] for a general overview). An embedded implication establishes a *local module*: The clauses D can only be used to prove C. Formulating a test over an embedded implication $\mu : D \rightarrow \langle C \rangle ?$, we make it explicit that we assume the truth of the statement but require its eventual verifi-

cation in μ. (Test actions turn an embedded implication into a *static* module, only promoting the conclusion [4].) Finally, an assertion then is the transformation of a test, into a partial proof which assumes the verification of the test, while at the same time conditioning the obtainability of the proof goal on the tested statements. Intuitively, $\mu : \langle D \rangle?$ within a proof $\Pi[\langle D \rangle?]$ to a goal C turns into $\Pi[D] \rightarrow C \wedge \mu : D$. Should $\mu : D$ not be verifiable, Π is invalidated.

Definition 4 (Test; embedded implication) *Given the language for Horn clauses P in Definition 2. We extend this with* embedded implications $D \supset C$ *and a* test operator $\langle \cdot \rangle?$ *as follows:*

P ::=*(as per Definition 2)*|
a clause of the form R or $\Box R$ where R is:
$\kappa'(\kappa_1(A_1)...\kappa_m(A_m) \rightarrow k''(D \supset C))|$
$\kappa'(\kappa_1(A_1)...\kappa_m(A_m) \rightarrow k''(D \supset \langle C \rangle?))|$
$\kappa'(\kappa_1(A_1)...\kappa_j(\langle A_j \rangle?)...\kappa_m(A_m) \rightarrow k''(H))$
with C atomic or tested, $\langle C' \rangle?$ with C' atomic

A test operator $\langle A \rangle?$ applies to atomic A, and resolves to $A^\top$ (A is true) or $A^\perp$ (A is false, under an open world assumption). □

Definition 5 (Abductive modal inference with assertions) *An* assertion *is the explicit assumption of a positive test result for a test $\langle A \rangle?$ for A atomic. We extend the abductive model inference of Definition 2 with an* assertion transformation rule*:*

Assertion *Select a clause R of the form P or $\Box P$, where P is*
$\kappa'(\kappa_1(A_1)...\kappa_j(\langle A_j \rangle?)...\kappa_m(A_m) \rightarrow k''(H))$.
then derive the new state:
$Q_1, ..., Q_{i-1},$
$\kappa' \circ \kappa_1 : A_1[unsolved], ...,$
$\kappa' \circ \kappa_j : A_j[asserted], ...,$
$\kappa' \circ \kappa_m : A_m[unsolved], Q_{i+1}...Q_n$

The multiset Δ is extended to also include assertions $\kappa' \circ \kappa_j : A_j[asserted]$, besides the assumptions in the proof (Definition 2). A falsified assertion $A^\perp$ can lead to the expansion or reconstruction of a proof. Given a proof as a state $L = Q_1, ..., Q_i, ..., Q_n$ for a goal G, with $Q_i = \kappa : A[asserted]$. If $A^\perp$, a proof expansion *is an abductive proof Π for $A^\top$ as goal, starting from the current conditions including $A^\perp$, s.t. $L = Q_1, ..., Q_{i-1}, \Pi, Q_{i+1}, ..., Q_n$ derives goal G (and Δ for L updated with Δ for Π). Failure to produce a proof expansion leads to a* proof reconstruction*, starting from a context $\kappa : \neg A$.* □

3.3 Multi-agent beliefs

We exploit Stone & Thomason's idea of contextualizing abductive inference [43] to explicitly reason with situated multi-agent beliefs. In our approach, abductive inference reasons over belief models. A belief model represents what an agent believes about the world, about actions are to be performed there, and what she is currently paying attention to. Beliefs and tasks are relativized to one or more agents, and to spatio-temporal "frames," situating beliefs and tasks in time and space.

The logical definition of belief models captures the beliefs, tasks, and attentional state. The structures used in the logical definition are based on the notion of Multi-Variant State Variables (MVSV) used in MAPL to define multi-agent belief models [7]. These state variables are used for several purposes. First, they can indicate domain values, taking values in the range of an ontological sort the variable is defined for. Important is that the absence of a value for an MVSV is interpreted as *ignorance*, not as falsehood (as per a closed-world assumption). In a similar way, state variables are used for expressing *private beliefs*, and mutual or *shared beliefs* [31,19]. A private belief of agent a_1 about content ϕ is (basically) expressed as $(K\{a_1\}\phi)$ whereas a mutual belief, held by several agents, is expressed as $(K\{a_1, a_2, \}\phi)$. Finally, MSVSs can be quantified over.

Definition 6 (Belief model, attentional state) *A* belief model *is a tuple* $\mathbb{B} = \langle \mathcal{A}, \mathcal{S}, \mathcal{K}, \mathcal{T}, \mathcal{F} \rangle$. *$\mathcal{A}$ is a non-empty set of agents. $\mathcal{S}$ is a spatio-temporal model, consisting of a set of spatiotemporal frames and a set of relations defined over these frames (see definitions below). $\mathcal{K}$ is a set of private and/or mutual beliefs [31,19]. A private belief of agent $a_i \in \mathcal{A}$ about content ϕ in spatiotemporal frame $\sigma_k \in \mathcal{S}$ is expressed as $(K\ \sigma_k\ \{a_i\}\phi)$, a private belief of a_i contributed to another agent a_j is written as $(K\ \sigma_k\ \{[a_i]a_j\}\phi)$ whereas a mutual belief, held by several agents $a_i..a_j \in \mathcal{A}$ for a spatiotemporal frame $\sigma_k \in \mathcal{S}$, is expressed as $(K\ \sigma_k\ \{a_i, a_j, \}\phi)$ (after [7]). Every belief ϕ is explicitly indexed, noted as $\phi[\mathbb{I}]$ for an index set $\mathbb{I}$ with indices from the namespace of indices for $\mathbb{B}$, $\mathbb{N}(\mathbb{B})$. The namespace for a $\mathbb{B}$ is a set of indices of content over which beliefs and tasks can be defined. $\mathcal{T}$ is an ordered set of tasks. A task t_a for an agent $a_i \in \mathcal{A}$ in spatiotemporal frame $\sigma_k \in \mathcal{S}$ is represented as $(T\ \sigma_k\{a_i\}t_a)$; a task involving multiple agents $a_i, , a_j$ is represented as $(T\ \sigma_k\{a_i, a_j\}t_a)$. Just like beliefs, tasks are explicitly indexed: $t_a[\mathbb{I}]$. $\mathcal{F}$ is the set of foregrounded beliefs and tasks, for which it holds that $\mathcal{F} \subseteq (\mathcal{K} \cup \mathcal{T})$.* □

Definition 7 (Spatiotemporal calculus) *Let a spatiotemporal frame be a tuple* $stf = \langle S, T, p, \varphi_{a_i} \rangle$ *consisting of a spatial interval S defined as a set of one or more contiguous places in space, T a continuous temporal interval $\langle X^o, X^c \rangle$, and φ_{a_i} a perspective under which S and possibly T are considered relative to an agent a_i. If S, T are presumed valid under every perspective, $\varphi = \top$ and can be omitted from the tuple. p is a time-point, and functions as unique index for the spatiotemporal frame. X^o, X^c are interpreted as points on the (interval) real line such*

that $X^o < X^c$. X^o is called the opening of the interval and X^c the closing of the interval. p is a time-point on the (point) real line. Given $stf_i = \langle S_i, \langle X_i^o, X_i^c \rangle, p_i$ and $stf_j = \langle S_j, \langle X_j^o, X_j^c \rangle, p_j \rangle$, if $p_i \leq p_j$ then $X_i^c \leq X_j^o$. We assume a function $\mathcal{Z}$ that provides a bi-simulation between time-points and temporal intervals [6]. Relations between temporal intervals are defined using a tractable fragment of the Allen interval calculus [1,32] with RCC-8 (cf. e.g. [15]). In addition, the calculus defines the following shorthands:

- *Given a frame F_{now} and $\langle X_{now}^o, X_{now}^c \rangle$ for identifying the present time, and a frame F_i with $\langle X_i^o, X_i^c \rangle$ such that F_{now} during F_i, then F_i is called open relative to F_{now}: $\mathsf{open}(F_i|F_{now})$, or simply $\mathsf{open}(F_i)$.*
- *Given a frame F_{now} and $\langle X_{now}^o, X_{now}^c \rangle$ for identifying the present time, and a frame F_i with $\langle X_i^o, X_i^c \rangle$ such that F_{now} after F_i, then F_i is called closed relative to F_{now}: $\mathsf{closed}(F_i|F_{now})$, or simply $\mathsf{closed}(F_i)$*

□

With these notions we can provide more detail to Stone & Thomason's four types of context κ [43], with c a context, i old information in c, e new information about events, and a the attentional state. We assume a context of type c to range over beliefs and tasks in $\mathbb{B}$, (and by extension so does i), and over general ("□-true-for-all-contexts") rules from a domain model. To make explicit what belief or task a c-type context in a proof step concerns, we adopt the following structure over labels functioning as context.

Definition 8 ($\mathbb{B}$-relative contexts in abductive inference) *Given a query of the form $Q_i = \kappa : A$. We further structure κ as a term following the format of beliefs and tasks (Definition 6). For a modality M (K or T), a spatiotemporal frame σ and a set of agents α, we build a term $M[\sigma\alpha]$: $Q_i = M[\sigma\ \alpha] : A$. Composition $\circ$ over contexts requires this composition to apply to the composed spatiotemporal frames and sets of agents (to whom beliefs and tasks are contributed). Specifically, $M[\sigma\ \alpha] \circ M[\sigma'\ \alpha'] \vdash M[\sigma \circ_\sigma \sigma'\ \alpha \circ_\alpha \alpha']$ under labelled logics that control composition over spatiotemporal frames, and over agent sets. For A holding for all contexts we keep □ (Definition 2).* □

For spatiotemporal frames, we control composition using a spatiotemporal calculus (Definition 7) to infer what relation R can hold between two frames: If $R(X, Y)$ for the intervals represented by σ, σ' then $\sigma \circ_R \sigma'$ in the label (cf. [13]). Similarly, we can define an algebra over agent sets based on a basic logic of attributing multi-agent beliefs.

3.4 Abduction for clarification as a continual collaborative activity

Stone & Thomason place their abductive inference in the context of an algorithm for comprehending and producing collaborative activity [44]. The algorithm is based on the idea that, ultimately, what we try to understand is the *intention* behind an activity. Looking at it from the viewpoint of collaboration, why did the agent do something? The algorithm aims to capture the interplay between action and interaction in collaboration. It defines collaborative activity in terms of collaborative agents taking tacit and public actions. Tacit actions are actions which one agent performs (mentally or physically) without those being observable ("sensed") by the other agent(s). Public actions are observable to all agents involved, and most importantly, help to explicitly further a common ground.

Algorithm 1 presents a definition of Stone & Thomason's algorithm. When trying to comprehend an observed event e, understanding builds an abductive proof for e returning an intention i presumed to underly e and updates to the context c' (including updates to the attentional state). This comprehension process takes into account the available communicative resources r, and the horizon of contextual alternatives $Z(c)$ [44]. The weighted cost-based nature of the abductive inference deals with the uncertainty inherent to such understanding; cf. also [22]. On the other hand, producing actions is based in the possible undertaking of tacit actions set against private beliefs, and then the selection of a message in the resulting context to be communicated publicly to the other agents involved.

Algorithm 1 Collaborative acting [44]

loop {
Perception
$e \leftarrow$ SENSE()
$\langle c', i \rangle \leftarrow$ UNDERSTAND$(r, Z(c), e)$
$c \leftarrow$ UPDATE(c', i)

Determination and Deliberation
$c' \leftarrow$ ACT-TACITLY(p, c)
$m \leftarrow$ SELECT(p, c')
$i \leftarrow$ GENERATE$(r, c', m, Z(c))$

Action
ACT-PUBLICLY$(a(i))$
$c \leftarrow$ UPDATE(c', i)
}

Underlying Algorithm 1 is an assumption that there is a symmetry between comprehension and production. They are assumed to be *coordinated* [43,44]): for a fixed perspective, an utterance will be understood the way the speaker intends it. Stone &

Thomason call this the Principle of Coordination Maintenance. They note that this is a strong assumption [44], as natural communication is able to deal with divergent perspectives and less certainty in action. The use of assertions allows us to lessen this assumption. Certainty can be asserted, but it needs to be verified. Should an assertion turn out to fail, we need to revise.

The Principle of Coordination Maintenance is reflected in the unverified updates made in Algorithm 1. With assertions, it may happen that an update is not warranted. For example both GENERATE and UNDERSTAND may rely on an asserted outcome of a clarification request. Only if this assertion is verified, can we make the update. Otherwise, the underlying proof needs to be expanded or revised.

Verification in a collaborative setting is, in and by itself, another run through the loop in the algorithm. A clarification question is raised (GENERATE), the answer obtained is processed (UNDERSTAND), and we need to verify whether the update made on the basis of the public action initially resulting from raising the question is actually warranted.

Algorithm 2 Continual collaborative acting

$\Sigma^{\pi} = \emptyset$

loop {
Perception
$e \leftarrow$ SENSE()
$\langle c', i, \Pi \rangle \leftarrow$ UNDERSTAND$(r, Z(c) \oplus \Sigma^{\pi}, e)$
$c \leftarrow$ VERIFIABLE-UPDATE(c', i, Π)

Determination and Deliberation
$c' \leftarrow$ ACT-TACITLY(p, c)
$m \leftarrow$ SELECT(p, c')
$\langle i, \Pi \rangle \leftarrow$ GENERATE$(r, c', m, Z(c) \oplus \Sigma^{\pi})$

Action
ACT-PUBLICLY$(a(i))$
$c \leftarrow$ VERIFIABLE-UPDATE(c', i, Π)
}

We alter Algorithm 1 to reflect this need for verification. Algorithm 2 adds a stack $\Sigma^{p}i$ of 'open' proofs, and it turns the UPDATE steps of Algorithm 1 into VERIFIABLE-UPDATE steps. The presence of a proof Π on the stack $\Sigma^{p}i$ indicates it has assertions in its Δ-set (cf. Definitions 2 and 5) that are either not yet verified, or have been falsified. We provide the proof stack as argument to the UNDERSTAND and GENERATE steps. This makes it possible to expand or restructure a proof currently on $\Sigma^{p}i$, using it to determine the next public action to make. (The use of Σ^{π} can be compared

to Ginzburg's Question-Under-Discussion structure.) A VERIFIABLE-UPDATE tests the update to be made, based on $\langle c', i\rangle$ and the underlying proof Π, against Σ^{π}.

Algorithm 3 Algorithmic definition of VERIFIABLE-UPDATE

Given an input proof Π with $\langle c', i\rangle$
if $\Sigma^{\pi} = \emptyset$ **then**
  $\Sigma^{\pi} \leftarrow$ PUSH(Π)
  return $c \leftarrow$ UPDATE(c', i)
else
  $\Pi' \leftarrow$ POP(Σ^{π})
end if
verified = true
for A_i `[asserted]` $\in \Delta$ of Π' **do**
  if $\langle c', i\rangle \wedge A_i$[asserted] $\vdash \bot$ **then**
    verified = false
    $\Delta \leftarrow \Delta[A_i/A^{\bot}\langle c', i\rangle]$
  else
    $\Delta \leftarrow \Delta[A_i/A^{\top}]$
  end if
end for
$\Sigma^{\pi} \leftarrow$ PUSH(Π)
if verified == false **then**
  $\Sigma^{\pi} \leftarrow$ PUSH(Π')
  return $c' \leftarrow$ DOWNDATE(c', Π')
else
  return $c' \leftarrow$ UPDATE(c', Π)
end if

Algorithm 3 outlines the algorithm for VERIFIABLE-UPDATE. If $\Sigma^{p}i$ is empty, we continue to make an update like in Algorithm 1: $c \leftarrow$ UPDATE(c', i). Otherwise, $\langle c', i\rangle$ are tested against the assertions in the Δ-set of the proof Π' popped from Σ^{π}. For each assertion $A_i[asserted]$ in Δ we check whether the intended update would falsify A_i. If so, we mark the assertion in Δ as $A^{\bot}$ and provide the falsifying update, If A_i is not falsified, it is marked as $A^{\top}$. Should we find that one or more assertions in Δ were falsified, we downdate the context with the update of Π'. We store subsequently push Π onto Σ^{π}, and finally Π' (making it the top-proof to be addressed).

This provides for a straightforward model of clarification in collaborative activity. If we specify a clarification in a proof together with an asserted positive outcome, the algorithm for continual collaborative activity can verify whether the answer is indeed subsequently obtained. If not, or if only partially so, further abductive inference can expand or restructure the invalidated proof to continue clarification.

4 Examples and Implementation

In this section we illustrate our approach on the following dialogue:

(1) Human places an object on the table
(2) Robot: "That is a brown object."
(3) Human: "It is a red object."
(4) Robot: "Ok. What kind of object is it?"
(5) Human: "Yes."
(6) Robot: "Aha. But what KIND of object is it?"
(7) Human: "It is a box."

This dialogue illustrates several important phenomena we would like to capture, in relation to common ground, grounding and clarification (§2). In (1) the human places an object on the table. The robot interprets this activity as an intention on behalf of the human to show the robot something. The robot accordingly acknowledges this, by communicating what it understands about the object (2). Next we see the first "conflict." The robot believes the object is brown, which contrasts with the human's belief that it is actually red (§2.3: Level 3, breakdown in recognition, belief conflict). The human corrects the robot (3). The robot accepts this correction (4), and subsequently asks a question after the type of object it's seeing. The human, not being particularly collaborative, replies with "yes" (5). This leads to another temporary breakdown in the dialogue, as a polar response isn't an expected kind of answer to the question the robot just posed (§2.3: Level 4, breakdown in consideration, intention evaluation). The robot repeats the question (6), with a stress on "KIND," to stress the expected answer (and contrast with the initial and rather unhelpful answer the human provided, (5)). Finally, the human provides the desired information (7).

The discussion below relies on a design of a preliminary implementation of weighted abduction, (Definitions 2 and 5), and its integration into the dialogue system and the overall cognitive architecture we are developing. Core to the design is that we maintain belief models (Definition 6) at different working memories. For our current purposes, we assume a belief model for dialogue, and one for a short-term working memory with a-modal content (i.e. "binding" working memory [23]). These two belief models are *synchronized*, in both directions. Following Lison & Kruijff's notion of context-sensitive language processing [30,29], foregrounded a-modal beliefs are provided to the dialogue belief model to represent salient information about the situated context(s) under consideration. We use namespace-information to appropriately keep track of where beliefs have their origin. This enables us to percolate changes to beliefs across multiple belief models, and achieve synchronization in the direction from the dialogue belief model back to the a-modal belief model. For example, the dialogue above illustrates how a robot's belief about visual properties ("this is a brown object") gets corrected through interaction with a human ("no

it is a red object"). Another use of synchronization, resulting from updating the foregrounded beliefs in the dialogue belief model with new beliefs in the a-modal model, is that salient visual referents can thus become available as "given" without having been explicitly introduced into the dialogue context beforehand.

(1) Human places an object on the table, (2) robot replies with "that is a brown object." Vision introduces a representation on the a-modal working memory, identifying a "brown object," which the human a_h has placed on the table. Following the methodology of [23], this introduces a *union* on this working memory. We use the union to create a corresponding belief for the robot a_r:

$$(K\sigma_{now}\{[a_r]a_h@_{v1:thing}(\mathbf{object} \wedge \langle Color\rangle(vb1 : color \wedge \mathbf{brown}))) \qquad (1)$$

Based on its (assumed) high visual salience, the belief in (1) is placed in the foreground set of the belief model on a-modal working memory, $\mathcal{F}_{am}$. This triggers synchronization with the dialogue belief model, ensuring that this belief also becomes foregrounded in the dialogue belief model ($\mathcal{F}_{com}$).

We treat the appearance of the a-modal belief in $\mathcal{F}_{com}$ as a SENSE action, cf. Algorithm 2. This triggers interpretation of that the observation as an act in the setting of a collaborative activity. The interpretation is guided by the fact that in (1) we have the robot *attribute* a belief to a_h. We treat the robot as a *cautious agent*, considering this attribution still as a private belief. As a_h was the one placing the object $v1$, the attribution does make the object available for reference. At the same time it triggers an explicit acknowledgment to make the belief part of the common ground. The following proofs reflects this.

e_1/σ_{now}:	$show(a_h, @_{v1:thing}(\mathbf{object} \wedge \langle Color\rangle(vb1 : color \wedge \mathbf{brown}))), E)$
e_1/σ_{now}:	$do(E)$
$\mathcal{F}_{com}$	$put{-}in{-}focus(v1, a_1, a_2)$
$K[\sigma_{now}\ [a_r]a_h]$:	$@_{v1:thing}(\mathbf{object} \wedge \langle Color\rangle(vb1 : color \wedge \mathbf{brown}))$

(2)

(2) starts with understanding the intention behind the human's showing the object. The proof establishes it as an attempt to update the belief model of the robot, including a new (attributed) belief about the object, and an updated foreground. Given this updated belief model, the next step we take is to generate a public act. Acting as a cautious agent, the goal of this public act is to turn the attributed belief into a shared belief about the object.

$a_2/\mathcal{F}_{com}$:	$that(v1)$
$[asserted]K[\sigma_{now}\ [a_r]a_h]$:	$@_{v1:thing}(\mathbf{object})$
$[asserted]K[\sigma_{now}\ [a_r]a_h]$:	$@_{v1:thing}\langle Color\rangle(vb1 : color \wedge \mathbf{brown}))$
e_2/σ_{now}:	$utter(a_r,$ 'that is a brown object', $E,$
	$@_{v1:thing}(\mathbf{object} \wedge \langle Color\rangle(vb1 : color \wedge \mathbf{brown})))$
e_2/σ_{now}:	$do(E)$
$K[\sigma_{now}\ \{a_r, a_h\}]$:	$@_{v1:thing}(\mathbf{object} \wedge \langle Color\rangle(vb1 : color \wedge \mathbf{brown})))$

(3)

The proof in (3) breaks up the proposition in (1) into its elementary predications. It then explicitly asserts the individual perceived properties ascribed to the object. An utterance "that is a brown object" is then introduced as action to make the ascription public, and yield the desired update of the observation to a shared belief. The use of a deictic pronoun to refer to $v1$ can be assumed at low cost as $v1$ is part of the foreground.

(3) Human indicates the robot is wrong, "it is a red object." (4) the robot complies, "Ok." At this point, the proof stack Σ^{π} contains the proof (3). The assertions about the ascribed properties are yet to be verified. Now the human replies with, "it is a red object." This yields a straightforward proof in the UNDERSTAND step.

$a_2/\mathcal{F}_{com}$:	$it(v1)$
$K[\sigma_{now}\ \{a_r, a_h\}]$:	$@_{v1:thing}(\mathbf{object})$
e_2/σ_{now}:	$utter(a_h,$ 'it is red', $E,$
	$@_{v1:thing}(\mathbf{object} \wedge \langle Color\rangle(vb1 : color \wedge \mathbf{red})))$
e_2/σ_{now}:	$do(E)$
$\mathcal{F}_{com}$	$put{-}in{-}focus(v1, a_2, a_3)$
$K[\sigma_{now}\ \{a_r, a_h\}]$:	$@_{v1:thing}(\mathbf{object} \wedge \langle Color\rangle(vb1 : color \wedge \mathbf{red})))$

(4)

The proof in (4) assumes that there is a shared belief about the object (minimal assumption), and that it can be referred to at low cost as it is (still) in the foreground. The observed utterance yields a semantic representation, which the proof concludes is the desired update relative to the object.

The problem now arises when we try to perform the update. The VERIFIABLE-UPDATE determines that the desired update contradicts an assertion in the proof currently on top of the stack, (3). As a result, we downdate the incorrect belief. We push the proof again onto the stack with a revised Δ-set. The object assertion is verified, the color assertion has been falsified by the intended update that $v1$ has a red color.

Production of the next public action is now guided by the problematic proof that remained on Σ^{π}. This illustrates the continual nature of our approach. The assertion of the color property made it possible for another agent to correct it. This resulted in a retraction of the belief *and* now leads to a step to address the situation. This step is simple. Acting on the assumption that the human is truthful, the robot simply acknowledges the correction, and updates its beliefs accordingly. The proof (5) does so based on expansion of (4) to yield the desired final update.

$$\begin{array}{ll} K[\sigma_{now}\ \{a_r, a_h\}]: & correction(@_{v1:thing}(\mathbf{object} \wedge \langle Color\rangle(vb1 : color \wedge \mathbf{red})), E') \\ e_2/\sigma_{now} & utter(a_r, \text{'ok'}, E', \top) \\ e_2/\sigma_{now} & do(E') \\ \hline K[\sigma_{now}\ \{a_r, a_h\}]: & @_{v1:thing}(\mathbf{object} \wedge \langle Color\rangle(vb1 : color \wedge \mathbf{red}))) \end{array} \tag{5}$$

The revised belief about $v1$ in the dialogue belief model leads to a synchronization with the corresponding belief about $v1$ in the a-modal belief model. Subsequently, this makes it possible for the visual modality in which the belief originated, to use the updated information to correct its categorization models.

(4) the robot asks after object type, (5) to which the human unhelpfully responds with "yes." Next, we assume that vision triggers a request for more information about the type of object we are looking at. Vision provides this trigger to the motivation planner, which in turn sends it to dialogue [21]. The *determination and deliberation* phase of Algorithm 2 handles the trigger by formulating a question: "What kind of object is this?" Following [18,24] this question has an expected answer, and is set against the background of shared beliefs about the object. The shared beliefs about the object, and the fact that $v1$ is (still) part of $\mathcal{F}_{com}$, makes it possible to simply refer to the object using a pronoun. We then produce a proof for a public action based on the idea that if we ask a question, and get a suitable answer, we can make the appropriate update to our belief model. (To keep the proof readable, we abbreviate the following semantic representation for the question to P in the proof.)

```
@b1:ascription(be ^
                <Mood>int ^
                <Tense>pres ^
                <Cop-Restr>(i1:thing ^ it ^
                            <Num>sg) ^
                <Cop-Scope>(k1:thing ^ kind ^
                            <Delimitation>unique ^
                            <Num>sg ^
                            <Quantification>specific ^
                            <Owner>(o1:thing ^ object)) ^
                <Subject>i1:thing ^
                <Wh-Restr>(w1:specifier ^ what ^
                          <Scope>k1:thing))
```

$$
\begin{array}{ll}
K[\sigma_{now}\ \{a_r, a_h\}]: & @_{v1:thing}(\mathbf{object}) \\
a_3/\mathcal{F}_{com}: & it(v1) \\
e_3/\sigma_{now} & utter(a_r, \text{'what kind of object is it'}, E, P) \\
e_3/\sigma_{now} & do(E) \\
[asserted]K[\sigma_{now}\ \{a_h\}] & Answer \sqsubset\ thing \\
[asserted]/e_4 \circ_{if} e_3 & utter(a_h, X, E', P' \models Answer) \\
[asserted]/e_4 \circ_{if} e_3 & do(E') \\
\hline
K[\sigma_{now}\ \{a_r, a_h\}]: & @_{v1:thing}(\mathbf{Answer})
\end{array}
\tag{6}
$$

Proof (6) proceeds by assuming that both agents know about $v1$, and that it is still in the foreground. It assumes a semantic structure for the utterance (provided by content planning), realized as an utterance (using a CCG realizer). The next steps formulate the idea that, once the question is raised, the human a_h responds with an answer. The proof needs to assert that a_h first of all knows the answer, and then provides this answer in an utterance in an event e_4 immediately following ($\circ_{if}$) upon the question. Provided that the answer is a proper answer to the question, we can update the belief both agents have with that information.

The subsequent answer "yes" satisfies the assertion that the user did something. It just wasn't the right answer. A polar statement is not a proper subtype of $thing$ (it is a *marker*). At the same time, it does neither prove nor disprove whether or not the human actually knows the answer. We are thus still left with falsified assertions, an open question, and proof (6) on Σ^{π}.

(6) the robot repeats the question, (7) after which the human provides a correct answer. Earlier, we dealt with falsified assertions by simply adopting a correction that the user provided ((3) – (5)). What makes the current setting different is that

in this case the assertions do not concern the beliefs of the robot a_r, but statements about the beliefs and intentions of the human a_h. The robot is therefore in the position to "correct" – which we do by repeating the action. The repetition of the question leads to a stronger stress on the phrase identifying the expected type of answer (see also the report by Kruijff-Korbayová et al.).

$$
\begin{array}{ll}
K[\sigma_{now}\ \{a_r, a_h\}]: & @_{v1:thing}(\mathbf{object}) \\
a_5/\mathcal{F}_{com}: & it(v1) \\
e_5/\sigma_{now} & utter(a_r, \text{'what kind of object is it'}, E, P) \\
e_5/\sigma_{now} & re-do(E) \\
[asserted]K[\sigma_{now}\ \{a_h\}] & Answer \sqsubset\ thing \\
[asserted]/e_6 \circ_{if} e_5 & utter(a_h, X, E', P' \models Answer) \\
[asserted]/e_6 \circ_{if} e_5 & do(E') \\
\hline
K[\sigma_{now}\ \{a_r, a_h\}]: & @_{v1:thing}(\mathbf{Answer})
\end{array}
\tag{7}
$$

The subsequent answer "it is a box" then satisfies the assertions, and leads to another update on $v1$. This update is handled as per the mechanisms discussed already.

5 Conclusions

The paper discussed an approach to modeling continual collaborative activity, using weighted abduction. The approach is based on earlier work by Stone & Thomason, which it extends with more explicit handling of multi-agent belief models, and the introduction of a notion of assertion based on [7]. The result is an approach in which Stone & Thomason's strong principle on coordination between comprehension and production [44] can be relaxed. Should coordination fail, indicated by failing assertions, then proof revision (expansion or rewriting) can be used as a continual way to redress the problem. The paper exemplified the approach on an extended example from human-robot interaction in a continual visual learning setting. The illustrations were set against the background of a design for integrating the approach into the dialogue system and the larger cognitive architecture, and a preliminary implementation. The examples illustrated how different issues in common ground, clarification, and grounding come together in a coherent framework that looks at dialogue as a continual collaborative activity.

References

[1] J.F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[2] J. Allwood. An activity based approach to pragmatics. In H. Bunt and B. Black, editors, *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*, pages 47–80. John Benjamins, Amsterdam, The Netherlands, 2000.

[3] N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, Cambridge, UK, 2003.

[4] M. Baldoni, L. Giordano, and A. Martelli. A modal extention of logic programming: Modularity, beliefs and hypothetical reasoning. *Journal of Logic and Computation*, 8(5):597–635, 1998.

[5] M. Baldoni, L. Giordano, A. Martelli, and V. Patti. A modal programming language for representing complex actions. In A. Bonner, B. Freitag, and L. Giordano, editors, *Proceedings of the 1998 JICSLP'98 Post-Conference Workshop on Transactions and Change in Logic Databases (DYNAMICS'98)*, pages 1–15, 1998.

[6] P. Blackburn, C. Gardent, and M. de Rijke. Back and forth through time and events. In *Proceedings of the 9th Amsterdam Colloquium*, Amsterdam, The Netherlands, 1993.

[7] M. Brenner and B. Nebel. Continual planning and acting in dynamic multiagent environments. *Journal of Autonomous Agents and Multiagent Systems*, 2008.

[8] M. Bugliesi, E. Lamma, and P. Mello. Modularity in logic programming. *Journal of Logic Programming*, 19, 20:443–502, 1994.

[9] J.E. Cahn and S.E. Brennan. A psychological model of grounding and repair in dialog. In *Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

[10] H. Clark. *Using Language*. Cambridge University Press, Cambridge, UK, 1996.

[11] Herbert H. Clark and Edward F. Schaefer. Contributing to Discourse. *Cognitive Science*, 13(2):259–294, 1989.

[12] K. Fischer. Discourse conditions for spatial perspective taking. In *Proceedings of the Workshop of Spatial Language and Discourse*, Hanse Wissenschaftskolleg, 2005.

[13] D.M. Gabbay. *Labelled Deduction Systems: Volume 1*, volume 33 of *Oxford Logic Guides*. Clarendon Press, 1996.

[14] S. Garrod and M. Pickering. Why is conversation so easy? *Trends in Cognitive Sciences*, 8:8–11, 2004.

[15] A. Gerevini and B. Nebel. Qualitative spatio-temporal reasoning with rcc-8 and allen's interval calculus: Computational complexity. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'02)*, 2002.

[16] J. Ginzburg. Resolving questions, I. *Linguistics and Philosophy*, 18(5):459–527, 1995.

[17] J. Ginzburg. Resolving questions, II. *Linguistics and Philosophy*, 18(6):567–609, 1995.

[18] J. Ginzburg. The semantics of interrogatives. In S. Lappin, editor, *Handbook of Contemporary Semantic Theory*. Blackwell, 1995.

[19] B.J. Grosz and S. Kraus. The evolution of shared plans. In A. Rao and M. Wooldridge, editors, *Foundations and Theories of Rational Agency*, pages 227–262. Springer, 1999.

[20] B.J. Grosz and C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[21] N. Hawes, M. Brenner, and K. Sjöö. Planning as an architectural control mechanism. In *HRI '09*, pages 229–230, New York, NY, USA, 2009. ACM.

[22] J.R. Hobbs. Abduction in natural language understanding. In L. Horn and G. Ward, editors, *Handbook of Pragmatics*, pages 724–741. Blackwell, 2004.

[23] H. Jacobsson, N. Hawes, G.J.M. Kruijff, and J. Wyatt. Crossmodal content binding in information-processing architectures. In *Proceedings of the 3rd Annual Conference on Human-Robot Interaction (HRI'08)*, 2008.

[24] G.J.M. Kruijff and M. Brenner. Phrasing questions. In *Proceedings of the AAAI 2009 Spring Symposium on Agents that Learn from Human Teachers*, Stanford, CA, March 2009.

[25] G.J.M. Kruijff, M. Brenner, and N.A. Hawes. Continual planning for cross-modal situated clarification in human-robot interaction. In *Proceedings of the 17th International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, Munich, Germany, 2008.

[26] G.J.M. Kruijff, H. Zender, P. Jensfelt, and H.I. Christensen. Clarification dialogues in human-augmented mapping. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction (HRI'06)*, 2006.

[27] S. Larsson. *Issue-Based Dialogue Management*. Phd thesis, Department of Linguistics, Göteborg University, Göteborg, Sweden, 2002.

[28] S. Li, B. Wrede, and G. Sagerer. A computational model of multi-modal grounding. In *Proc. ACL SIGdial workshop on discourse and dialog, in conjunction with COLING/ACL 2006*, pages 153–160, 2006.

[29] P. Lison and G.J.M. Kruijff. Efficient parsing of spoken inputs for human-robot interaction. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN09)*, Toyama, Japan, 2009.

[30] P. Lison and G.J.M. Kruijff. An integrated approach to robust processing of situated spoken dialogue. In *Proceedings of the Second International Workshop on the Semantic Representation of Spoken Language (SRSL09)*, Athens, Greece, 2009.

[31] K. Lochbaum, B.J. Grosz, and C.L. Sidner. Discourse structure and intention recognition. In R. Dale, H. Moisl, , and H. Somers, editors, *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. Marcel Dekker, New York, 1999.

[32] B. Nebel and H.J. Bürckert. Reasoning about temporal relations: A maximal tractable subclass of Allen's interval algebra. *Journal of the ACM*, 42(1):43–66, 1995.

[33] M.J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–225, 2004.

[34] M. Poesio and D. Traum. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347, 1997.

[35] M. Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, King's College, University of London, August 2004.

[36] M. Purver, J. Ginzburg, and P. Healey. On the means for clarification in dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*, pages 235–255. Kluwer Academic Publishers, 2003.

[37] V. Rieser and J.D. Moore. Implications for generating clarification requests in task-oriented dialogues. *Ann Arbor*, 100, 2005.

[38] K.J. Rodriguez and D. Schlangen. Form, intonation and function of clarification requests in German task oriented spoken dialogues. In *Proceedings of Catalog'04 (The 8th Workshop on the Semantics and Pragmatics of Dialogue, SemDial04)*, 2004.

[39] A.W. Russell and M.F. Schober. How beliefs about a partner's goals affect referring in goal-discrepant conversations. *Discourse Processes*, 27(1):1–33, 1999.

[40] M.F. Schober. Spatial dialogue between partners with mismatched abilities. In K.R. Coventry, T. Tenbrink, and J.A. Bateman, editors, *Spatial language and dialogue*, pages 23–39. Oxford University Press, 2009.

[41] P. Sgall, E. Hajičová, and J. Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht, the Netherlands, 1986.

[42] M. Stone and R.H. Thomason. Context in abductive interpretation. In *Proceedings of EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue*, 2002.

[43] M. Stone and R.H. Thomason. Coordinating understanding and generation in an abductive approach to interpretation. In *Proceedings of DIABRUCK 2003: 7th workshop on the semantics and pragmatics of dialogue*, 2003.

[44] R.H. Thomason, M. Stone, and D. DeVault. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. In D. Byron, C. Roberts, and S. Schwenter, editors, *Presupposition Accommodation*. to appear.

[45] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, 35(4):460–470, 2005.

[46] D.R. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, Computer Science Department, University of Rochester, December 1994.

[47] D.R. Traum. Computational Models of Grounding in Collaborative Systems. In *working notes of AAAI Fall Symposium on Psychological Models of Communication*, pages 124–131, 1999.

Contextually appropriate intonation of clarification in situated dialogue (preliminary report)

I. Kruijff-Korbayová, R. Meena, and G.J.M. Kruijff*
DFKI GmbH
Saarbrücken, Germany
{ivana.kruijff,rame01,gj}@dfki.de

***Abstract*— We develop an approach for determining contextually appropriate intonation of grounding feedback utterances, in particular clarification requests raised during continuous and cross-modal learning in autonomous robots. Following the analysis in [Ginzburg, 1996], [Purver et al., 2003], [Purver, 2004] on clarifications in human dialogue, we develop strategies for formulating clarification requests in human-robot dialogue. As for intonation, we combine the approaches of [Steedman, 2000a], [Lambrecht and Michaelis, 1998] and [Engdahl, 2006] to intonation assignment based on *information structure*, an underlying partitioning of utterance context the reflects its relation to discourse context. We implement our approach in the CogX system. Empirical verification of our approach comes from psycholinguistic experiments.**

I. Introduction

When in doubt, ask. This paradigm very much applies to autonomous robots which self-understand and self-extend in the environment they find themselves. It is therefore essential for these systems to learn continuously about their surroundings. Moreover, the learning process has to be driven mainly by their own curiosity, rather then some external motivation. During the course of learning or planning actions, a robot might require additional information from a human interlocutor. Spoken dialogue is a means through which a robot can request new information or clarify the knowledge it has acquired about the situated environment.

This ability to self-initiate a dialogue, besides adding autonomy to a robot's behavior, also allows the robot to connect its belief state to that of its interlocutor. This enables the participating agents to perform *grounding*, and arrive at a *common ground* [Clark and Schaefer, 1989]. A robot's grounding *feedback* is one of the means to arrive at a common ground. By employing a variety of feedbacks, such as *acknowledgement* (e.g. 'I see a red box'), *verification request* (e.g. 'Is it a red box?'), *disambiguation request* (e.g. 'Is this box red or brown?') or simple *information request* (e.g. 'What color is this box?') a robot is able to assert or clarify its knowledge about its surroundings.

But is it just the lexical choice that makes up the meaning of an utterance? For example, the answers to the questions in (1a) and (2a) (sentences 4 and 5 taken from [Steedman, 2000a]), contain the same exact string of words, (1b) and (2b). If it were only the lexical choice that governed the semantics of these answer pairs, then both these answers should convey the same meaning, and hence be interchangeable. This, however, is not the case. What is it then, that distinguishes the response in (2b) from that in (1b)?

(1) a. Who proved completeness?

b. (MARCEL) (proved COMPLETENESS).
H* L L+H* LH%

(2) a. What did Marcel prove?

b. (Marcel PROVED) (COMPLETENESS).
L+H* LH% H* LL%.

[Steedman, 2000a], among others, attributes the differences in these answer pairs to their information content, more specifically their *information structure* (IS). In spoken English, IS is realized through intonation. The intonation contours shown under the answers, with the words printed in SMALL CAPITALS indicating the alignment of the most prominent pitch accents (fundamental frequency f_0 peaks) and the brackets indicating the intonation phrases, originate in the pioneering work of [Pierrehumbert, 1980]. The utterances in (1b) and (2b) with their respective intonation contours are not interchangeable without sounding unnatural or altering the meaning of the dialog in the given context.

The task of making the grounding feedback utterances of a conversational robot contextually appropriate, inevitably also involves intonation assignment.

To illustrate some of the issues we aim to address, let us first consider a scenario where a H(uman) presents the R(obot) an object in an "empty" verbal and visual context, i.e., this is the first object presented, there are no other objects in the scene (e.g., an empty table top) and nothing has been said so far. R recognizes the shape and color of the object with some degrees of certainty. To verify its perception, R can produce a verification request, giving H the opportunity to accept, reject or correct all or parts of R's hypothesis:

(3) Is it a red box?

Next, let us assume that R is certain above some threshold ϵ that the object is a box, but not certain enough that it is red. Instead of the more neutral verification request above, R can, for example, provide the following grounding feedback:

(4) It is a box. Is it red?

* Supported by EU FP7 Project 'CogX'(FP7-ICT-215181).

The first utterance provides an *acknowledgment* concerning the more certain recognized property, for the sake of transparency.[1][2] The second utterance verifies the less certain recognized property. Interestingly, this two-fold feedback can also be combined into one utterance:

(5) Is it a RED box?

Intonation plays an important role here: Intuitively, and in line with the existing work on intonation and information structure, accentuation can mark the part(s) of the utterance with the highest need for verification, whereas that assumed to have been correctly recognized, and thus part of the common ground between R and H, can remain unaccented. Conversely, if color is assumed correctly recognized, but shape is uncertain, R can utter the following verification request.

(6) Is it a red BOX?

The intonation countour in (6) will, however, be practically hardly distinguishable from the case where both color and shape are presented for verification:

(7) Is it a RED BOX?

This compact manner of formulating R's feedback, combining acknowledgment and clarification request in one utterance, is reminiscent of the *implicit feedback* strategy often used in dialogue systems, e.g., [Aust et al., 1995], [Larsson, 2002] where parameters extracted from user's input recognized with high confidence are incorporated into the next system prompt, e.g., "When do you want to travel to Paris?" incorporates the recognized destination city into the next prompt asking for the date of travel.

When the verbal and/or visual context is not empty, additional factors influencing accent placement become relevant. Previous research has addressed the contextual factors influencing accent placement. For example, it is widely accepted that in statements accent is assigned to those words that distinguish a referent from relevant alternatives available in the context [Steedman, 2000b], [Steedman and Kruijff-Korbayová, 2003]. This role of accent placement is present in questions, too [Lambrecht and Michaelis, 1998], [Engdahl, 2006], and similar principles hold. The second confirmation request by R in the following example illustrates accent placement w.r.t. previously mentioned objects:

(8) (H presents a red cone)
R: Is it a red CONE?
H: Yes, that's right.
(H presents a blue pyramid)
R: Is it a BLUE cone?
H: No, it's a blue pyramid.

[1]R can of course misrecognize objects and/or their properties, as well as be wrong in assessing its (un)certainty. That is why grounding feedback is essential.

[2]R may make a short pause after the first utterance, to give H the opportunity for positive or negative feedback. If H gives positive or no feedback, R goes on.

The use of contextually inappropriate intonation in situated dialogue might lead to ambiguities and/or mislead the dialogue participants to maintain incongruous belief states. Such situations would undermine the very purpose of spoken dialogue. To avoid such miscommunication, a situated dialogue system needs to use contextual information in the utterance content planning and determination of intonation contour during surface realization. This in turns requires the system to incorporate mechanism to capture, represent and maintain contextual details.

This paper is organized as follows. In Section II we overview the most relevant previous work on clarification in dialogue, and on using information structure to control intonation of system output. In Section III we discuss the contextual factors that we take into account when formulating grounding feedback. In Section IV we describe our approach to partitioning utterances according to their information structure, and the effect this has on intonation. The implementation of our approach is presented in Section V. In VI we sketch the experimental setup in which we will obtain empirical verification of our approach. In Section VII we conclude.

II. BACKGROUND

[Purver et al., 2003] investigated the nature of clarifications in human dialogue, using the BNC dialogue corpus. They chart out the range of possible forms of clarification requests (CR), together with the range of readings they can convey. Their analysis reveals the frequency of various CR forms, with *reprise* (50%), non-reprise (12%) and conventional types (30%) of all. Moreover, 60% of these reprise CRs were of *reprise fragment* type i.e. *elliptical literal reprise* [Ginzburg and Sag, 2000] of a fragment of the lead-in utterance. From our experience with conversational robots, we come to an observation that similar CR forms can be employed in a human-robot conversation. However, since we aim to build autonomous conversational robots that can self-initiate a clarification dialogue, we do not always have a preceding utterance that is being clarified (as is the case in the BNC), and from which the clarification form can be derived. Under such circumstances we need alternative approaches to formulate the CRs in our system.

Since intonation of CRs is our main focus, we follow Steedman's theory of *information structure* (IS) [Steedman, 2000a]. In Steedman's view, the IS of an utterance is composed of two parts. One of these parts links the utterance to the current discourse context (the *theme*), and the other part contributes information (the *rheme*). As an illustration, refer to the bracketing in (1b) and (2b) which indicates the theme-rheme segmentation in view of the respective questions. However, like most of the existing work on IS, Steedman's theory makes predictions about the intonation of statements. Some preliminary hypotheses concerning the IS in questions are formulated in [Prevost and Steedman, 1994a], but these only concern information-seeking wh-questions.

Among the existing investigations into the IS of questions, [Lambrecht and Michaelis, 1998] have discussed in detail accent placement in information questions. Their discussion addresses predominantly questions requesting information, but they also mention some examples of CRs, also including echo questions (also called reprise-questions in the literature, [Bolinger, 1989], [Engdahl, 2006]). [Engdahl, 2006] in her work on information packaging in questions highlights the role of question under discussion (QUD) [Ginzburg, 1996] in providing the right locus to account for focus-ground articulation of utterances. When a speaker produces an utterance with a particular information packiging, this provides information about their information state, what s/he knows and what s/he wants to achieve at this point.

With regard to practical applications, early work on controlling the intonation of synthesized speech w.r.t. context concerned mainly accenting open-class items on first mention, and deaccenting previously mentioned or otherwise "given" items [Hirschberg, 1993], [Monaghan, 1994]. But such algorithms based on givenness fail to account for certain accentuation patterns, such as marking explicit contrast among salient items. Givenness alone also does not seem sufficient to motivate accent type variation.

In [Prevost, 1996] contrastive accent patterns and some accent type variation are modeled using Steedman's approach to IS in English. In one application he handles question-answer pairs where the question intonation analysis in IS terms is used to motivate the IS of the corresponding answer, realized through intonation. Another application concerns intonation in generation of short descriptions of objects, where Theme/Rheme partitioning is motivated on text progression grounds, and Background/Focus partitioning distinguishes between alternatives in context.

In [Kruijff-Korbayová et al., 2003] and [Baker et al., 2004], a similar IS-based approach is applied to assign contextually appropriate intonation to the output of an actual end-to-end dialogue system (German and English, respectively). The reported evaluation results show that this leads to qualitative improvements.

The intonation of questions, and CRs in particular, has so far been largely neglected in dialogue systems. The practical applications mentioned above all concentrated on the assignment of intonation in statements. However, a series of production and perception experiments around the HIGGINS dialogue system [Edlund et al., 2004], shows that fragmentary grounding utterances in Swedish differ in prosodic features depending on their meaning (acknowledgment vs. clarification of understanding or perception), and that subjects differentiate between the meanings accordingly, and respond differently [Edlund et al., 2005], [Skanze et al., 2006].

In a study of a corpus of German task-oriented human-human dialog, [Rodríguez and Schlangen, 2004] also found that the use of intonation seemed to disambiguate clarification types, with rising boundary tones used more often to clarify acoustic problems than to clarify reference resolution.

Our work extends the use of information structure to control the intonation of dialogue system output beyond answers to information-seeking questions: we include acknowledgments as well as clarification requests, and ultimately other types of questions. We include both fragmentary grounding feedback and full utterances, and address varying placement of pitch accents depending on context and communicative intention.

III. FACTORS IN CR FORMULATION

We now discuss a range of contextual factors that we take into account as shaping the content and surface form of grounding feedback in our system. Currently we concentrate on feedback that grounds entities and their properties; we leave grounding of actions for future work. That is, a feedback utterance concerns a referent (currently just a single one) and is clarifying some of its properties.

a) Competing referents in verbal and visual context: We distinguish situations where the context prior to the CR utterance is empty or non-empty. Empty preceding verbal context means that no entities have been mentioned prior to the CR utterance. Empty preceding visual context means that no entities have been made visually available. In the first scenario described in the introduction (example (3), when the first object is placed onto an empty tabletop and nothing has been said yet, the verbal context is empty and the visual context contains only the one object, the referent that the CR addresses. We conjectured that in this situation, the speaker is quite free in formulating their utterance to reflect their assumptions about the common ground and their communicative goal(s), e.g., signaling by intonation what they consider most important about the referent, as in example (5). As soon as either verbal or visual context is non-empty, the speaker needs to take into account similar entities available in the context, and properly distinguish the intended referent from them, as in example (8). It is well known that salience plays an important role in this case.

b) Salience: In this paper we have been explicitly focusing on a scenario where a H(uman) places and object on a tabletop in front of the R(obot). This makes this object inherently salient. We conjecture that even if the visual and/or verbal context is non-empty, the just placed object is the most salient one, and thus allows deictic reference (e.g., "it", "this", "that", "the/this box"). For other cases we will employ existing algorithms for generation of referring expressions, particularly ones that take verbal and visual salience into account [Krahmer and Theune, 2002], [Kelleher and Kruijff, 2006], [Zender et al., 2009].

c) Source of problem: Existing work on CRs has investigated the relationship between the form of a CR and the source of trouble that trigers it, inspired by the *action ladder* [Clark, 1996]. Thus, [Rodríguez and Schlangen, 2004] distinguish between problems in channel, acoustic or lexical recognition, parsing, reference resolution and intention recognition. [Rieser and Moore, 2005] extend this repertoire with ambiguity refinement and belief confirmation. This work has addressed CRs with verbal antecedents, that is, CRs addressing problems in natural language understanding (typically speech). In our system, visual recognition

constitutes an additional potential problem source. In this paper, it is the latter that we concentrate on. As part of the CogX project [Kruijff and Janíček, 2009] formulate an abduction-based approach to planning CRs following [Stone and Thomason, 2002], [Stone and Thomason, 2003], [Thomason et al., pear] and extending this approach to include misunderstanding or lack of understanding relative to a situated context.

d) Multiple communicative goals: As we have pointed out in the introduction, when there are multiple communicative goals pertaining to one referent, we consider ways of formulating an utterance that satisfies these goals simultaneously. Example (5) illustrated a case of acknowledging the type of the referent and verifying its color in one utterance.

e) Clarification issue: While the range of issues under clarification is potentially very broad, we currently focus just on CRs pertaining to either the type or visually recognizable attributes of a referent. That is, for example, whether the referent is a box (or a ball, etc.) and/or what is its color, shape, size and potentially also location w.r.t. other objects.

f) Clarification hypothesis: The system either recognizes the type or attribute(s) of an object with some degree of certainty, or it is at a loss. When it has multiple hypotheses, they may be competing (i.e., comparably good) or there may be a single best one. Correspondingly, the system is then able to formulate CRs with increasing degree of specificity:

(9) What is it? (no hypothesis)
What COLOR does it have?

(10) Is it a CONE or a PYRAMID? (competing hyp.)
Is it RED or BROWN?

(11) Is it a CONE? (single hypothesis)
Is the box RED?

g) Conflicting expectations: In the examples we discussed so far, the system was recognizing objects "out of the blue", that is, it had no expectations or information from other sources about the object's properties. But we also want to consider situations when a CR arises because recognition results could not be integrated with R's beliefs for some reason. This arises for example when visual recognition and linguistic interpretation give conflicting results:

(12) H: (shows a red box) This box is red.
R: (does not recognize the color as red)
WHAT color is the box?

Alternatively, the wh-phrase can be left *in situ*, realized with noticeable rising intonation [Engdahl, 2006]:

(13) The box has WHAT color?

The placement of the pitch accent on the wh-word is appropriate in a context where some specific value has just been mentioned. Bolinger coined the term *reprise questions* for questions that 'replay' a (part of a) previous utterance [Bolinger, 1989], [Engdahl, 2006]. But this intonation pattern seems suitable in a range of situations when the recognition results are incompatible with contextually established expectations, and it is not straightforward to see the CR as a reprise of a previous move:[3]

(14) H: I will now show you some red objects.
R: Okay!
H: (shows a red ball)
R: (recognizes a red ball) A red ball.
H: (shows a red box)
R: (recognizes a box of a different color)
WHAT color is the box?

In both these examples, the intonation indicates that R has trouble matching the recognized color to the expected color. This is different from not being able to recognize the color at all, as in example (9). In both (12) and (14), R could alternatively formulate the CRs as a fragment:

(15) Red?
Is it red?
A RED box?

Moreover, if R is quite sure that the object has another color, it could also say:

(16) Is the box not BLUE?
Isn't the box blue?

(17) Is the box BLUE?

While R's responses in (14) and (15) only indicate unvillingness to accept that the color of the object is red, (16) in addition proposes what R thinks is the correct color. We conjecture that the decision to take initiative and make an alternative proposal depends on the presence of a hypothesis (in this case: color) of which R is certain above some threshhold.

h) Re-raising an issue: When R poses a question but does not get a response that it can interpret as an answer, it needs to reiterate its question. We need to keep track of this in order to ensure that reiteration is not a repetition. Besides an altogether different formulation of the question, what particularly interests us is the intonation of a repeated request. There is the option of using the same intonation pattern, but more pronounced accentuation in this case, e.g., a pitch accent with higher intensity. Another possibility suggested in [Engdahl, 2006] is using a different intonation pattern, particularly, different placement of the nuclear accent.

(18) What COLOR is the box?

(19) WHAT color the box?

Engdahl suggests that by accenting an initial wh-word in an information question the speaker may signal that "the issue she is introducing is one that has already been raised in the converstion but not been resolved" [Engdahl, 2006][p.101].

[3]We believe that the expectation could also arise by inference generalizing from previous actions, thus not on the basis of any verbal clue. For example, if R is shown one red object after another, it may form the expectation that more red objects will follow. When an object arives that R does not recognize as red, we believe it can also pose the CR as in (12) or (13).

i) Authority/responsibility for an issue: On the basis of corpus-based experiments in Swedish reported in [Gustafson-Čapkpová, 2005], [Engdahl, 2006] points out that, contrary to common claims in the literature about declarative questions, declarative utterances are often interpreted as questions even without any formal signal, such as a rising prosodic gesture. She summarizes that statements about the addressee are commonly understood as requests for confirmation (checks). We would like to take this further and claim that what really is at stake here is who has the responsibility to resolve an issue: if it is the speaker, a formally declarative utterance is interpreted as a statement, whereas if it is the hearer, it can be interpreted as a request for confirmation.

The relevance of speaker's vs. hearer's responsibility is also suggested in [Steedman, 2000a] when comparing different theme- and rheme-tunes. In the future we would like to investigate whether these two lines of thought can be brought together.

The contextual factors listed above influence the formulation of CRs in various ways, and therefore need to be available to the utterance planner. We are still in the process of finding out what and how to encode explicitly in the input representations (cf. Section V). We recognize that there are interactions and interdependencies between the factors, and it is part of our ongoing and future work to pin at least some of these down.

IV. Intonation in CRs

Intonation has several aspects. According to the autosegmental phonology approach [Pierrehumbert, 1980], [Ladd, 1996], the intonation contour of an utterance consists of intonation phrases, and these are described in terms of accents and boundary tones of various types. The literature is rife with discussions of the meanings of various intonation contours of statements, including dicussions of the meaning differences ascribed to different accent types and their placement on different words within an utterance. However, much less is available concerning the intonation of questions. Most work addresses boundary tones, but very little research has addressed accent types and placement in questions.

Grounding feedback of course involves both statements, such as acknowledgments, and various types of clarification requests which can be broadly characterized as questions, even though they are not always realized by full sentences, and when they are, they may have interrogative or declarative syntax.

We take Steedman's approach to intonation in English [Steedman, 2000a], [Steedman, 2000b] as a starting point, because it is an approach that (i) tightly couples intonation with grammatical structure; (ii) associates intonation with discourse meaning in terms of information structure; (iii) provides a compositional semantics of English intonation in information-structural terms; (iv) assumes a general IS-sensitive notion of discourse context update; (v) has proved its worth in previous practical applications to control intonation assignment w.r.t. context.

Steedman concentrates on certain aspects of intonation that primarily have to do with *information structure*, as a partitioning of utterance meaning reflecting its relation to the context. He recognizes two independent dimensions of IS: The *Theme-Rheme* partitioning reflects a notion of aboutness: the *theme* maintains a link to the discourse context, i.e., to what has been previously mentioned or can be considered as already established, and the *rheme* contributes some information about the theme. The second partitioning is into *Background* and *Focus*, where focus reflects where a theme or a rheme differs from alternatives in the context. The relation between intonation and IS is, in a nutshell, that (i) themes and rhemes constitute intonation phrases; (ii) main pitch accents are assigned to words realizing the focus within a theme or a rheme; (ii) the type of accent depends on whether the focus is within a theme or a rheme. Regarding boundary tones, Steedman argues that a falling boundary tone signals the speaker's responsibility for (ownership of) the corresponding information unit, whereas a rising boundary tone signals the hearer's responsibility/ownership.

The combination of accents and boundary tones gives rise to theme-tunes and rheme-tunes. As an illustration, observe the intonation contour of the utterances in (1b) and (2b). The rhemes exhibit the **H* L(L%)** tune, where the **L(L%)** tone marks the intonational phrase boundary. The themes exhibit the **L+H* L(H%)** tune.

A. Intonation in statements

Similarly to the earlier work cited in Sec. II, we can straightforwardly apply Steedman's approach to the assignment of intonation in acknowledgments, which are statements of what the robot has perceived or understood with sufficient certainty. Consider the following example, assuming an empty verbal and visual initial context:

(20) a. H places a red box

b. R recognizes a red box with sufficient certainty

c. R: (It is)$_T$ (a RED BOX)$_R$
H* H* LL%

d. H places a blue box

e. R recognizes a blue box with sufficient certainty

f. R: (THIS)$_T$ (is a BLUE box)$_R$
LH* LH% H* LL%

The theme-rheme partitioning of (20c) can be derived by assuming that upon presentation of a new object, the issue of what type and properties hold of this object is a relevant theme to address. This theme can alternatively be left out from the explicit realization, producing the fragmentary acknowledgment "A RED BOX". The rheme contains two new pieces of information (namely, the type of the object and its color), which can both be assigned focus, and thus accent. Given accent-projection rules of English, the overt accent on the adjective is optional in this case.

When the interaction continues in (20cd), the context is no longer empty. While the theme-rheme partitioning of the next

robot's response could be parallel to that of (20c), resulting in "It is a BLUE box", (20f) shows a more fancy alternative. Its theme is construed as contrastive, indicating that the new object is different from another object available in the context (namely, the red box). About this theme it is then asserted that it is of type box and is blue. Within this rheme, it is the color property that distinguishes the current object from the previous one, and therefore is assigned focus, and thus becomes accented.

An open issue in this area is, how to assign focus when the verbal and visual context differ, i.e., not all visually available objects have been verbally mentioned or some verbally introduced referents are not available visually. Do referents recently mentioned verbally or those available visually have a higher priority as contextual alternatives for the purpose of contrast? Or are they all equal? This is a question of deciding between or combining visual and verbal salience, which requires further research.

B. Intonation in questions

We think of questions as including (at least) *information requests* (IRs) and *clarification requests* (CRs) requesting additional information or verification from the hearer.[4] Syntactically, these may be interrogative or declarative sentences or non-sentential fragments. The interrogative sentences may be wh-questions or yes/no-questions.

Our notion of CRs is broad, and potentially hard to delineate from that of IRs because it includes not only requests for verification of a previous verbal utterance (also called *checks* or *check questions* in the literature), but also requests for verification of uncertain or missing information about the visual scene or of the robot's inferences, which may or may not have a verbal antecedent.

Few authors have discussed accent placement in questions and the associated discourse meaning(s).[5] We draw on the approach [Lambrecht and Michaelis, 1998], who studied the formal and pragmatic principles that govern the placement of sentence accent in English *information questions* (IQs). While they do not define what they mean by IQs, their discussion addresses wh-questions used as IRs and, in a few cases, as CRs (request to repeat, i.e., an echo question, or request to identify/clarify a referent).

[Lambrecht and Michaelis, 1998] propose that in IQs the sentence accent does not fall on the focus, i.e., the wh-word (though sometimes it can). Instead they argue that "the sentence accent in IQs represents an independently motivated type: the topic-establishing or -ratifying accents observed in declarative contexts to co-occur with focus accent" [p. 539]. They propose a set of general principles for accent placement. Their analysis of IQs relies on Lambrecht's theory of information structure, and in particular his distinction between *knowledge presupposition* (KP) and *topicality presupposition* (TP) [Lambrecht, 1994]. The KP of an IQ is an open proposition, which may or may not correspond to an established topic. Essentially, accents mark those parts of the KP that are not included in a TP, i.e., not ratified as topics in the discourse. The following example from [Lambrecht and Michaelis, 1998][ex. (41a), p. 525ff] illustrates their analysis:

(21) What cities did you VISIT?

Contexts:

i. I heard you went to France and visited various cities.

ii. I heard you avoided Paris on your trip to France.

Presuppositions:

KP: You visited x cities (in France)
TP: in context (i): 'you' and 'cities' are ratified topics in context (ii): 'you did something with respect to cities (in France)' is ratified

Assertion: x=what?

Focus : what

The accent falls on 'visit' by default, because the other parts of the KP, 'you' and 'cities', are ratified topics in both contexts, and therefore unaccentable. In context (i), the 'you visited some cities' is activated, but not (considered by the speaker) ratified as a TP. In context (ii), 'you visited some cities' is construed (by the speaker) as contrasting with an alternative proposition activated in the context, namely 'you avoided some city'.

We think that there is another possibility, although [Lambrecht and Michaelis, 1998] do not discuss it: namely, that the context in (i) also allows the speaker to consider 'you visited some cities' as a ratified topic, in which case their rules would license accent placement on "what", since "the propositional function in the KP is an already ratified topic" [p. 535], i.e., the KP and the TP coincide:

(22) WHAT cities did you visit?
TP: 'you visited some cities' is a ratified topic

This is a CR to clarify, further *identify a referent*. Another case of accenting the wh-word is for metalinguistic reasons, in *echo questions* [Lambrecht and Michaelis, 1998], [Bolinger, 1989].

(23) A: I went to France and visited $\langle inaudible \rangle$
B: WHAT did you visit?

Presuppositions

KP: You visited x (in France)
TP: 'you visited x' is a ratified topic

Applied to our scenario, the approach of [Lambrecht and Michaelis, 1998] seems to make the following predictions:

(24) Context: empty initial verbal or visual context;
H places a red box;
R recognizes obj_1 to be a box with sufficient certainty, but does not recognize its color

[4] At the moment we do not cosider rhetorical questions.
[5] See [Lambrecht and Michaelis, 1998] for an overview.

Acceptable presuppositions:

KP: obj_1 is a box and has color x

TP: obj_1 is a ratified topic (by virtue of having been ostentatively placed in front of R);
'$obj1$ has some property' is a ratified topic (objects have properties and the ascription of properties to objects is something R does/learns)

Assertion: x=what?

Focus : what

Possible realization:
What COLOR does it have?

The accent lands on "color" because the color property is one of possible object properties. Referring to obj_1 by a pronoun is justified because it is a ratified topic. In order to produce a reference by a full NP "the box", we need to appeal to the fact that although obj_1 is ratified as topic, it has not yet been grounded among the interlocutors that it is a box.

The following example shows a realization that (we believe) is infelicitous in the above context:

(25) WHAT color does it have?

KP: obj_1 has x color
TP: 'obj_1 has x color' is ratified topic

If the given TP holds, then the accent will end up on "what" as the only possibility, because the KP and the TP coincide. It is hard to justify why this TP should be blocked. All physical objects have at least one color, after all, and recognizing and learning objects' colors is one of the things that R does.

We believe that [Engdahl, 2006] helps us to resolve this apparent puzzle. She points out that although sentence-initial wh-phrases are normally not accented in English, it is possible to accent them in certain uses, specifically, when the speaker is introducing an issue that has already been raised in the conversation but not been resolved. The accent on "what" is thus blocked in (25), since the issue of the color of obj_1 has not been previously raised. Note that in (22), the issue 'you visited some cities' has been raised in the context. An issue can be raised explicitly in a preceding utterance but also more indirectly. [Engdahl, 2006] formulates her interpretive account of accent placement on the wh-word or elsewhere in a an IQ, and the placement of the wh-phrase in-situ or sentence initially in terms of *question under discussion* (cf. also [Ginzburg, 1995a], [Ginzburg, 1995b], [Roberts, 1996]). Topic ratification following Lambrecht, theme- and rheme-alternative presupposition resolution following Steedman, or question under discussion accommodation following Engdahl are tightly related, and more research is needed to operationalize them.

Lambrecht's notion of topic and Steedman's notion of Theme are largerly compatible (at least for our practical purposes). Our approach is, therefore, to integrate these two approaches to IS in IQs.

In an early paper on IS partitioning, [Prevost and Steedman, 1994b] suggest that wh-questions of the form illustrated in (26) consist of a theme and a rheme, as follows:

(26) I know what the CAT scan is for, but
(WHICH condition)$_T$ (does URINALYSIS address)$_R$?
LH* LH% H* LL%

The Rheme of the question sets the theme of the corresponding answer:

(27) (URINALYSIS addresses)$_T$ (HEMATURIA)$_R$.
LH* LH% H* LL%

All the questions used in [Prevost and Steedman, 1994b] have two accents of different types. This is indeed an indication of two information units, a theme and a rheme. It is not entirely clear how to analyze questions with just one accent in Steedman's approach. One possibility is that they contain a theme-rheme partitioning, but the theme is prosodically unmarked, because either the theme does not contain a focus, or the focus is not assigned an accent (since accenting theme-focus is optional in the presence of a rheme focus). Another possibility is that such questions constitute just one information unit, either a theme or a rheme – as should be reflected in the type of accent (if Steedman is right about the correspondence between accent type and theme- vs. rheme-focus in English).

[Lambrecht and Michaelis, 1998][p. 531] note that the open proposition presupposed by an IQ (i.e., the KP) constitutes a topic-comment structure, too. Unlike in statements, where the comment is asserted, in IQs it is itself presupposed. In their example (41a) reproduced above as (21), 'cities' is said to constitue a topic and 'did you visit' a comment.

For the time being, we are not excluding any of the three possibilities for theme-rheme partitioning in IQs. Which one obtains in a specific case should depend on the context and the communicative goal of the speaker. The context may constrain, but sometimes does not fully determine the IS of an utterance. This holds of questions as well as statements. The speaker has considerable (but not unlimited) freedom in articulating their utterances.

In order to also deal with *polar questions* (PQs), we our approach combining [Lambrecht and Michaelis, 1998] and [Steedman, 2000a] to them. Thus, a PQ involves a KP concerning the validity of the proposition in question. The proposition itself can be viewed as having an IS partitioning like the corresponding statement would have. Again, it is conceivable that we will find PQs with the full theme-rheme partitioning, as well as ones that only constitute one information unit, a theme or a rheme.

V. IMPLEMENTATION

The production of verbal CRs in the CogX system consists of the following phases: communicative goal planning, utterance planning, surface realization and speech synthesis. Communicative goal planning is described in more detail in [Kruijff and Janíček, 2009]. Below we give more details

about the remaining phases. The utterance planner takes an abstract logical description of a communicative goal as input, and produces one or more logical forms that represent how that goal can be expressed in a contextually appropriate way. The logical forms are then realized into utterances using the OpenCCG realizer [White and Baldridge, 2003], [White, 2006].

To represent communicative intentions, and the corresponding utterance meanings at all levels of processing, we use ontologically rich, relational structures [Kruijff, 2005] based on the Hybrid Logic Dependency Semantics (HLDS)[Kruijff, 2001], [Baldridge and Kruijff, 2003].

A. Utterance planning for CRs

Our implementation of utterance planning for CRs is an extension to [Kruijff, 2005]. We specify the planning grammar as *systemic networks* [Kruijff, 2005], [Mathiessen, 1983], [Bateman, 1997]. These systems take an abstract logical form as input, and enrich them with specifications of the desired linguistic realization, that take available context information into account. There is, of course, a tight relationship between the choices made by the utterance planner and the realization options available in the grammar of the realizer.

Let us illustrate the abstract logical forms used in our system. Since this work is under development, these forms are still subject to modification.

(28) $@_{d1:dvp}$(c-goal
$\wedge \langle SpeechAct \rangle$**question**
$\wedge \langle Relation \rangle$**clarify**
$\wedge \langle Content \rangle (e1 : ascription$
$\wedge \langle Target \rangle (b1 : entity$
$\wedge \langle Salient \rangle$**true**)
$\wedge \langle Property \rangle (b2 : entity \wedge box$
$\wedge \langle Color \rangle (b3 : quality \wedge red \wedge$
$\langle Known \rangle$**uncertain**)))
$\wedge \langle Context \rangle$**empty**)

The relational structure in (28) uses standard operators to model relations between features in HLDS [Kruijff, 2001], [Baldridge and Kruijff, 2003]. $@_n \langle R \rangle m$ implies that there is a relation R between nominals n and m. A nominal is a formula, which is interpreted as a unique reference to a state in the underlying model theory of the logic. Moreover, nominals can be sorted, to indicate the ontological category of the proposition that holds at the state referred to by the nominal. A detailed explanation of the HLDS structures is available in [Kruijff, 2005]. Below we briefly elucidate those aspects that are relevant to the current discussion, i.e., how does this form capture the system's intention to raise a CR, and the necessary contextual details.

This structure specifies: the intention of a communicative act, `SpeechAct`; its relation to the context, `Relation`; the content, `Content`, which consist of a predicate and its arguments to be communicated; and a relevant portion of the current discourse context, `Context`.

We use a shallow classification of speech acts following [Searle, 1975], and incorporating basic insights from the DAMSL [Core and Allen, 1997] classification of *forward-looking functions*. Currently, we distinguish assertions, questions, directives and greetings. Grounding feedback utterances are either assertions (in the case of acknowledgements) or questions (in the case of CRs). The *backward-looking function* of a communicative act is captured as its relation to the context. CUrrently we distinguish between accept, reject, clarify and answer.

The nominal $e1$ of sort *ascribe* in `Content` feature influences the type of main verb in its predicate part of the utterance, and thus its overall structure. In this case it also means that a full sentence will be produced. The relations embedded within the main predicate depend on its type. In the current example, the relational feature `Target` represents an object referred to, to which a property represented under the `Property` relation is being ascribed.

Under `Context`, we currently can include a list of active referents in the verbal and/or situational context, with their relevant properties. (See example (29) below.)

The utterance planner algorithm processes the `Content`-part of the input representation and applies systems of the systemic network to its nominal to enrich this structure with additional syntactic and semantic features relevant for realization. Let us have a brief look at this process.

A systemic network is a collection of systems. Each system has an entry condition checks the presence of specific nominal and/or specific feature values in the input logical form. On having met its entry condition, the chooser associated with the system applies. A chooser is basically a n-ary decision tree. Each node in the tree has a condition, which leads to nodes further down the tree. The leaf nodes in such a tree represent action(s). These actions include adding new relations, features and propositons in the current logical from. As a consequence the abstract logical forms is enriched with additional features. Moreover, these enriched forms can then also become input to another system, and so on.

For example, the presence of the $\langle SpeechAct \rangle question$ feature triggers a system responsible for tense, aspect and mood, which can add the attribute $\langle Mood \rangle int$ (*int*:interrogative), if the decision is to produce an interrogative question (rather than a declarative one). This choice guides the realizer further to decide between a wh-question and a polar one, and to insert the appropriate structure bits. Moreover, since we want to produce spoken output, and we want to control its intonation, we have a system that assigns the type of final boundary tone, in this case, $\langle UtFinalBT \rangle rising$.

Next, the presence of a nominal $e1$ of sort $ascription$ invokes a system, which adds an intransitive verb *be* as the main verb. The presence of this intransitive verb triggers another system to look for its arguments, in this case the scope and restrictor. The arguments of the ascription predicate are represented in the logical from under the relational feature $\langle Target \rangle$ and $\langle Property \rangle$.

For a nominal referring to an entity, further systems

are invoked, which first of all decide whether to use a pronominal or full NP reference. In our example, a pronoun will be used for the entity $b1$ in the $\langle Target \rangle$, because the entity is salient. A full NP will be used for $\langle Property \rangle$. For this, further systems include additional details such as, $\langle Num \rangle sg$, $\langle Quantification \rangle specific$, and $\langle Delimitation \rangle existential$. Finally, the feature $\langle color \rangle$ will trigger systems for adding a modifier to the noun.

The structural decisions discussed so far amount to a plan to realize the question *it it a red box*.

Now let us describe how we plan the information-structure of an utterance. In the implementation developed so far, we work without the theme/rheme distinction and consider the utterances all rheme, for the time being. Fous is assigned either based on contrast w.r.t. alternatives in the context, or, in the absence of alternatives to contrast, to the most informationally prominent part of the utterance. In the latter case, as we suggested earlier, focus can for example be assigned to indicate an uncertain property value. That is what we do in the current example. For the sake of comparison, let us consider another version, where the context is not empty, i.e., the $\langle Context \rangle$ feature contains a list of relevant alternatives/competitors:

$$
\begin{aligned}
(28') \quad & \wedge \langle Context \rangle \textbf{empty}(b4 : de - list \\
& \quad \wedge \langle First \rangle (b5 : entity \wedge box \\
& \quad\quad \wedge \langle Color \rangle (b6 : quality \wedge black)) \\
& \quad \wedge \langle Next \rangle (b7 : entity \wedge box \quad \wedge \\
& \langle Size \rangle (b6 : quality \wedge small))
\end{aligned}
$$

Given that both these other entities are boxes, we assign focus to the color attribute of $b2$. In this particular case the result is the same as it was without the contextt, but that is of course a coincidence.

At the end of the utterance planning process, the final logical form specifying the content of the utterance, enriched with a range of features relevant for a contextually appropriate form of realization has the following shape:[6]

(29)
```
<Content>(e1:ascription
 ^ be ^ <Tense>pres
 ^ <Mood>int ^ <UtFinalBT>rising ^
 <Cop-Restr>(b1:entity ^ context
  ^ <Delimitation>unique ^ <Num>sg
  ^ <Quantification>specific) ^
 <Cop-Scope>(b2:entity ^ box
  ^ <Num>sg
  ^ <Quantification>specific
  ^ <Delimitation>existential
  ^ <Modifier>(b3:quality ^ red
  ^ <Focus>true)) ^
 <Subject>(b1:entity)
)
```

For comparison, let us consider an alternative example, in which a CR request is produced against a context containing red competitors which are not boxes. We have the following input representation:

$$
\begin{aligned}
(30) \quad & @_{d1:dvp}(\text{c-goal} \\
& \wedge \langle SpeechAct \rangle \textbf{question} \\
& \wedge \langle Relation \rangle \textbf{clarify} \\
& \wedge \langle Content \rangle (e1 : ascription \\
& \quad \wedge \langle Target \rangle (b1 : entity \\
& \quad\quad \wedge \langle Salient \rangle \textbf{true}) \\
& \quad \wedge \langle Property \rangle (b2 : entity \wedge box \\
& \quad\quad \wedge \langle Color \rangle (b3 : quality \wedge red \wedge \\
& \quad\quad\quad \langle Known \rangle \textbf{uncertain}))) \\
& \wedge \langle Context \rangle (b4 : de - list \\
& \quad \wedge \langle First \rangle (b5 : entity \wedge ball \\
& \quad\quad \wedge \langle Color \rangle (b6 : quality \wedge red)) \\
& \quad \wedge \langle Next \rangle (b7 : entity \wedge cone \\
& \quad\quad \wedge \langle Size \rangle (b6 : quality \wedge red))))
\end{aligned}
$$

The systemic networks will then accordingly place the focus on the type of the entity in the ascription predicate, namely 'box'. Furthermore, let us assume that this time round the utterance planner decided to produce a demonstrative pronoun in the subject, reflected by the feature $\langle proximity \rangle$`proximal`, and a declarative CR: this changes the mood, while the boundary tone will still be rising. Here is the corresponding logical form:

(31)
```
<Content>(e1:ascription
 ^ be ^ <Tense>pres
 ^ <Mood>decl ^ <UtFinalBT>rising ^
 <Cop-Restr>(b1:entity ^ context
  ^ <Proximity>proximal
  ^ <Delimitation>unique ^ <Num>sg
  ^ <Quantification>specific) ^
 <Cop-Scope>(b2:entity ^ box
  ^ <Focus>true ^ <Num>sg
  ^ <Quantification>specific
  ^ <Delimitation>existential
  ^ <Modifier>(b3:quality ^ red)) ^
 <Subject>(b1:entity)
)
```

These logical form are then provided to the OpenCCG realizer for realization.

B. CR realization

The realizer should produce a (ranked) set of possible utterances and a range of possible intonation contours for each of them. We follow the approach advocated in [Steedman, 2000a], [Steedman, 2000b] of specifying intonation compositionally in the grammar. To integrate intonation within our existing grammar, we employ the multi-level sign approach of [Kruijff and Baldridge, 2004].

The intonation model we use originates with the model of intonation in [Pierrehumbert, 1980]. The core components of this model are *pitch accents*, *phrasal tones* and *boundary tones*. Pierrehumbert identified the ways in which these components can be combined to form the f_0 contours.

Words in English language are associated with lexical stress, which assigns one syllable greater prominence. However, it is the relative prominence of the words in an utterance

[6]The utterance planning process is monotonic, but relations and features that the realizer does not use are pruned in a final "clean-up" step.

that determines its intonation contour. We refer to this type of prominence as *pitch accent*. **H*** and **L*** are the most basic pitch accents. While words bearing **H*** are realized as a tone occurring high in speaker's pitch range, **L*** is realized as a low tone. There also exist bitonal pitch accents such as **L+H***, **L*+H**, **H+L*** and **H*+L** An occurrence of a *phrasal tone* **H** and **L** after a succession of pitch accents delimits a *intermediate* phrase. Phrasal tones thus control the pitch accent between the most recent pitch accent tone and the end of the phrase. Intermediate phrases (or a sequence of these) can be followed by either of the boundary tones **H%** and **L%**. Boundary tones thus form the *intonational phrase*, and describe the general direction of rising or falling of the f_0 contours at the end of intonational contour.

To relate information structure to realization, we adopt Steedman's approach [Steedman, 1991] through [Steedman, 2000a] and [Steedman, 2000b], where the intonational realization of a rheme is with one of a set of possible *rheme-tunes*, such as **H* L(L%)** , **H* LL%** , or **H* LL$** (rheme-marking accents are: H*, L*, H*+L, H+L*), and the theme comes with a *theme-tune*, such as **L+H* L(H%)**, **L+H* LH%** or **L+H* LH$** (theme-marking accents are: L+H*, L*+H). The "%" is marks utterance-medial boundaries, as for phrases set apart by commas, and the "$" for utterance-final boundaries.

The pitch accent, phrasal and boundary tones form the basic lexemes of a prosody grammar. These basic components combine to form intonational phrases. However, this combination is not arbitrary. The combinatorics of a prosody grammar can be summarized as follows (from [Prevost and Steedman, 1994a]):

1) A boundary must combine with at least one pitch accent to its left.
2) A boundary may not combine with another boundary.
3) Constituents which are prosodically unmarked may freely combine with non-boundary constituents bearing prosodic information.
4) Multiple pitch accents may occur in an intonational phrase
5) A complete intonational phrase may combine only with another complete intonational phrase.
6) A constituent of any length bearing no pitch accents can promote itself to a full thematic intonational phrase (Null Theme Promotion Rule)

To incorporate an intonation layer into our grammar, we first add lexical families to indicate pitch accent tunes. For the moment we start with the rheme accent **H***:

```
# Lexicon/Categories for Pitch Accents
family Hs(indexRel="Rheme"){
entry: n<2> [X hs]\* n<2> [X acc] :
                T:pa-unit(<Rheme>X);
}
```

This lexicon entry for the Hs (Hstar) family specifies that it takes an accusative noun as argument and results in a noun, which now has a rheme accent mark (*hs*). Since a rheme accent can mark nouns, adjectives, predicates etc. they are the argument and result of these functional categories. If we want adjectives such as 'red' to be marked with the rheme-accent, we need to add another *entry* in this family that takes an adjective as an argument and results in a rheme-accent bearing modifier. We extend the *signs* with an additional syntactic feature *pitacct* to capture the type of tune marking of a sign. This new syntactic feature further constrains the combination of an unaccented word with any boundary tone (rule 3). On the other hand, this entry prevents two rheme accented nouns from combining. This also implies that the grammar does not support intermediate phrases. The grammar is therefore currently too constrained to accommodate CRs having successive accents.

Pitch accent bearing terms can further combine with boundary tones **LL%**, **LH%** and **HH%** on their left. However, not all of these pitch accent and boundary combinations are allowed. One way to control this is to use a feature in a sign to specify boundary tones that it can combine with. But this will then lead to unnecessary combinations and even a combinatoric explosion if the grammar is large. To avoid this, we introduce boundary tones as separate lexical families. Boundaries are thus defined as functional categories, which take pitch accent bearing lexical items as argument, and result in an intonational phrase. This way only pitch accent bearing words are allowed to combine with a boundary tone, as specified in rule (1). Following is an entry for a boundary that marks an intonational phrase.

```
#Lexical/Categories for Boundary Tones
family LLp(indexRel="CompRheme"){
entry: np<>[X acc cp]\* np<> [X acc ip]:
                T:b-unit(<CompRheme>X);
}
```

The feature *ip* indicates that only an intermediate phrase can combine with a boundary tone **LL%**, to result in a complete rheme phrase i.e an intonational phrase. Only complete theme/rheme phrases are allowed to combine further and result in larger intonational phrases, rule (5). As our current focus is on rheme-only CRs, we haven't incorporated this rule into our grammar yet.

The prosody layer of the grammar is still under development. What we are aiming for are the following realizations including prosodic marking for the examples of logical forms in (29) and (31):

(29′) is it a red `Hs` box `HHp`

(31′) is it a red box `Hs HHp`

The `Hs` indicates rheme-accent **H*** on the word to its left, and `HHp` indicates the utterance final boundary **HH%**.

The final step in the production is speech synthesis. We are using the MARY Text-To-Speech synthesizer [Schröder and Trouvain, 2003][7]. A post-processing step converts the output of the CCG realizer to the MaryXML format, in which a text is annotated with the type and location (word) of pitch accents, and breaks (boundary tones).

VI. EXPERIMENTATION

Since we do not yet have a working integrated system with the desired functionality, we opted to start by experimentally

[7] `mary.dfki.de`

bad	1	2	3	4	5	good
unintelligible	1	2	3	4	5	intelligible
artificial	1	2	3	4	5	lifelike
unnatural	1	2	3	4	5	natural
confusing	1	2	3	4	5	clear
inappropriate	1	2	3	4	5	appropriate

Fig. 1. Semantic differential scales for subjective quality judgments

verifying the crucial assumptions of our approach on a component-basis first. To this end we are currently preparing the first round of experiments addressing intonation assignment in clarification requests. Of the many open issues that call for experimental support, we take as a starting point the question whether differences in the placement of the main accent in confirmation requests are perceivable in synthesized speech and whether visual context licenses contrastive accent placement. The confirmation requests we test have the form "Is it a red box?" for a range of different colors and shapes, where either the noun or the adjective is accented. We synthesize them using the Mary TTS system [Schröder and Trouvain, 2003][8] . Subjects see a photo of a hand placing an object onto a table top in front of a robot, where the table top is either empty or contains a few other objects chosen to make the accent placement either congruent or discongruent with the visual scene. Subjects hear a robot's confirmation request and then a response by a human. These examples illustrate the stimuli:

(32) (the table top already contains a blue box;
H is adding a red box)

a. R: Is it a RED box? — Congruent, correct

b. R: Is it a red BOX? — Discongruent, correct

H: Yes, that's right.

For each such stimulus, we elicit subjective qualitative judgments of the robot's speech using the semantic differential scales shown in Fig. 1.

As the preparation of this first round of experiments is ongoing, we do not yet have results to report.

VII. CONCLUSIONS

In this paper we discussed the importance of dialogue in continuous learning for autonomous robots. We illustrated how spoken dialogue helps to improve grounding in a human-robot conversation. We described our approach to producing grounding feedback with a range of context-dependent forms and with context-dependent intonation. We discussed, in particular, how we assign information structure to utterances to reflect their relation to the context, and how this then determines the intonation of the produced output. We described our approach to implementation employing systemic networks for utterance planning, and a CCG realizer. We outlined how we are going about incorporating a layer describing the intonation structure of English in our existing grammar. We also sketched the setup of experiments the we are preparing to test the validity of our approach to producing contextually appropriate grounding feedback.

The utterance planner and the intonation layer of the grammar are under development, and as we extend them, we further operationalize the theoretical notions used in our approach. Once we start running the experiments, we will be able to work in cycles of specification-extension – empirical validation – implementation – experimental-evaluation.

[8] http://mary.dfki.de

REFERENCES

[Aust et al., 1995] Aust, H., Oerder, M., Seide, F., and Steinbiss, V. (1995). The Philips automatic train timetable information system. *Speech Communications*, 17(3–4):249–262.

[Baker et al., 2004] Baker, R., Clark, R. A. J., and White, M. (2004). Synthetizing contextually appropriate intonation in limited domains. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*.

[Baldridge and Kruijff, 2003] Baldridge, J. and Kruijff, G.-J. M. (2003). Multi-modal combinatory categorial grammmar. In *EACL'03: Proceedings of the 10th conference of the European chapter of the Association for Computational Linguistics*, Budapest, Hungary.

[Bateman, 1997] Bateman, J. A. (1997). Enabling technology for multilingual natural language generation: the kpml development environment. *Journal of Natural Language Engineering*, 3(1):15–55.

[Bolinger, 1989] Bolinger, D. (1989). *Intonation and Its Uses*. Stanford University Press, Stanford, CA.

[Clark, 1996] Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge, UK.

[Clark and Schaefer, 1989] Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13:259–294.

[Core and Allen, 1997] Core, M. and Allen, J. (1997). Coding dialogs with the damsl annotation scheme. In *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Boston, MA.

[Edlund et al., 2005] Edlund, J., House, D., and Skantze, G. (2005). The effects of prosodic features on the interpretation of clarification ellipses. In *Proceedings of Interspeech. Lisbon, Portugal*, pages 2389–2392.

[Edlund et al., 2004] Edlund, J., Skanze, G., and Carlson, R. (2004). Higgins - a spoken dialogue system for investigating error handling techniques. In *Proceedings of ICSLP. Jeju, Korea*, pages 229–231.

[Engdahl, 2006] Engdahl, E. (2006). Infromation packaging in questions. *Empirical Issues in Syntax and Semantics*, 6(1):93–111.

[Ginzburg, 1995a] Ginzburg, J. (1995a). Resolving questions, I. *Linguistics and Philosophy*, 18(5):459–527.

[Ginzburg, 1995b] Ginzburg, J. (1995b). Resolving questions, II. *Linguistics and Philosophy*, 18(6):567–609.

[Ginzburg, 1996] Ginzburg, J. (1996). Interrogatives: Questions, facts and dialogue. In *The Handbook of Contemporary Semantic Theory*, pages 385–422. Blackwell.

[Ginzburg and Sag, 2000] Ginzburg, J. and Sag, I. A. (2000). Interogative investigations: the form, meaning and use of english interrogatives. *CSLI Lecture Notes and Philosophy*, (123).

[Gustafson-Čapkpová, 2005] Gustafson-Čapkpová, S. (2005). *Integrtaing Prosody into an Account of Discourse Structure*. PhD thesis, Stockholm University.

[Hirschberg, 1993] Hirschberg, J. (1993). Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, (63):305–340.

[Kelleher and Kruijff, 2006] Kelleher, J. and Kruijff, G. (2006). Incremental generation of spatial referring expressions in situated dialogue. In *Proc. Coling-ACL-2006*, Sydney, Australia.

[Krahmer and Theune, 2002] Krahmer, E. and Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In van Deemter, K. and R.Kibble, editors, *Information Sharing: Givenness and Newness in Language Processing*. CSLI Publications, Stanford, CA, USA.

[Kruijff, 2005] Kruijff, G. (2005). Context-sensitive utterance planning for ccg. In *Proceedings of the European Workshop on Natural Language Generation*, Aberdeen, Scotland.

[Kruijff and Janíček, 2009] Kruijff, G. and Janíček, M. (2009). Abduction for clarification in situated dialogue. Report in this deliverable.

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

[Kruijff and Baldridge, 2004] Kruijff, G.-J. M. and Baldridge, J. (2004). Generalizing dimensionality in combinatory categorial grammar. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 191, Morristown, NJ, USA. Association for Computational Linguistics.

[Kruijff-Korbayová et al., 2003] Kruijff-Korbayová, I., Ericsson, S., Rodríguez, K. J., and Karagjosova, E. (2003). Producing contextually appropriate intonation is an information-state based dialogue system. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 227–234. ACL.

[Ladd, 1996] Ladd, D. R. (1996). *Intonational Phonology*. Cambridge University Press, Cambridge.

[Lambrecht, 1994] Lambrecht, K. (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge.

[Lambrecht and Michaelis, 1998] Lambrecht, K. and Michaelis, L. A. (1998). Sentence accent in information questions: default and projection. *Linguistics and Philisophy*, 21(5):477–544.

[Larsson, 2002] Larsson, S. (2002). *Issue-Based Dialogue Management*. Phd thesis, Department of Linguistics, Göteborg University, Göteborg, Sweden.

[Mathiessen, 1983] Mathiessen, C. M. I. M. (1983). Systemic grammar in computation: the Nigel case.

[Monaghan, 1994] Monaghan, A. (1994). Intonation accent placement in a concept-to-dialogue system. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 171–174, New Paltz, NY.

[Pierrehumbert, 1980] Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology.

[Prevost and Steedman, 1994a] Prevost, S. and Steedman, M. (1994a). Information based intonation synthesis. In *In Proceedings of the ARPA Workshop on Human Language Technology*, pages 193–198.

[Prevost and Steedman, 1994b] Prevost, S. and Steedman, M. (1994b). Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153.

[Prevost, 1996] Prevost, S. A. (1996). *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. Phd thesis, University of Pennsylvania, Institute for Research in Cognitive Science Technical Report, Pennsylvania, USA.

[Purver, 2004] Purver, M. (2004). *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, King's College, University of London.

[Purver et al., 2003] Purver, M., Ginzburg, J., and Healey, P. (2003). On the means for clarification in dialogue. In Smith, R. and van Kuppevelt, J., editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*, pages 235–255. Kluwer Academic Publishers.

[Rieser and Moore, 2005] Rieser, V. and Moore, J. (2005). Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05), Ann Arbor*.

[Roberts, 1996] Roberts, C. (1996). Information structure in discourse: towards an integrated formal theory of pragmatics. In Yoon, J.-H. and Kathol, A., editors, *Papers in semantics*, volume 46 of *OSU Working papers in linguistics*. Ohio State University, Columbus.

[Rodríguez and Schlangen, 2004] Rodríguez, K. J. and Schlangen, D. (2004). Form, intonation and function of clarification requests in german task oriented spoken dialogues. In *Proceedings of Catalog '04 (The 8th Workshop on the Semantics and Pragmatics of Dialogue, SemDial04), Barcelona, Spain, July* .

[Schröder and Trouvain, 2003] Schröder, M. and Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377.

[Searle, 1975] Searle, J. R. (1975). Indirect speech acts. In Cole, P. and Morgan, J. L., editors, *Speech Acts: Syntax and Semantics, Volume 3*, pages 59–82. Academic Press, New York.

[Skanze et al., 2006] Skanze, G., House, D., and Edlund, J. (2006). User responses to prosodic variation in fragmentary grounding utterances in dialogue. In *Proceedings of Interspeech ICSLP. Pittsburgh PA, USA*, pages 2002–2005.

[Steedman, 1991] Steedman, M. (1991). Structure and intonation. *Language*, pages 260–296.

[Steedman, 2000a] Steedman, M. (2000a). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31:649–689.

[Steedman, 2000b] Steedman, M. (2000b). *The Syntactic Process*. The MIT Press, Cambridge MA.

[Steedman and Kruijff-Korbayová, 2003] Steedman, M. and Kruijff-Korbayová, I. (2003). Discourse and information structure. *Journal of Logic, Language and Information*, 12:249–259.

[Stone and Thomason, 2002] Stone, M. and Thomason, R. (2002). Context in abductive interpretation. In *Proceedings of EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue*.

[Stone and Thomason, 2003] Stone, M. and Thomason, R. (2003). Coordinating understanding and generation in an abductive approach to interpretation. In *Proceedings of DIABRUCK 2003: 7th workshop on the semantics and pragmatics of dialogue*.

[Thomason et al., pear] Thomason, R., Stone, M., and DeVault, D. (to appear). Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. In Byron, D., Roberts, C., and Schwenter, S., editors, *Presupposition Accommodation*.

[White, 2006] White, M. (2006). Efficient realization of coordinate structures in combinatory categorial grammar. *Research on Language and Computation*, 4(1):3975.

[White and Baldridge, 2003] White, M. and Baldridge, J. (2003). Adapting chart realization to CCG. In *Proceedings of the Ninth European Workshop on Natural Language Generation*, Budapest, Hungary.

[Zender et al., 2009] Zender, H., Kruijff, G.-J., and Kruijff-Korbayová, I. (2009). Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence, Pasadena, CA, United States, AAAI*.

Efficient Parsing of Spoken Inputs for Human-Robot Interaction

Pierre Lison and Geert-Jan M. Kruijff

***Abstract*— The use of deep parsers in spoken dialogue systems is usually subject to strong performance requirements. This is particularly the case in human-robot interaction, where the computing resources are limited and must be shared by many components in parallel. A real-time dialogue system must be capable of responding quickly to any given utterance, even in the presence of noisy, ambiguous or distorted input. The parser must therefore ensure that the number of analyses remains bounded at every processing step.**

The paper presents a practical approach to address this issue in the context of deep parsers designed for spoken dialogue. The approach is based on a word lattice parser combined with a statistical model for parse selection. Each word lattice is parsed incrementally, word by word, and a discriminative model is applied at each incremental step to prune the set of resulting partial analyses. The model incorporates a wide range of linguistic and contextual features and can be trained with a simple perceptron. The approach is fully implemented as part of a spoken dialogue system for human-robot interaction. Evaluation results on a Wizard-of-Oz test suite demonstrate significant improvements in parsing time.

I. Introduction

Developing robust and efficient parsers for spoken dialogue is a difficult and demanding enterprise. This is due to several interconnected reasons.

The first reason is the pervasiveness of *speech recognition errors* in natural (i.e. noisy) environments, especially for open, non-trivial discourse domains. Automatic speech recognition (ASR) is indeed a highly error-prone task, and parsers designed to process spoken input must therefore find ways to accomodate the various ASR errors that may (and will) arise. This problem is particularly acute for robots operating in real-world noisy environments and deal with utterances pertaining to complex, open-ended domains.

Next to speech recognition, the second issue we need to address is the *relaxed grammaticality* of spoken language. Dialogue utterances are often incomplete or ungrammatical, and may contain numerous disfluencies like fillers (err, uh, mm), repetitions, self-corrections, etc. Rather than getting crisp-and-clear commands such as "*Put the red ball inside the box!*", we are more likely to hear utterances such as: "*right, now, could you, uh, put the red ball, yeah, inside the ba/ box!*". This is natural behaviour in human-human interaction [1] and can also be observed in several domain-specific corpora for human-robot interaction [2]. Spoken dialogue parsers should therefore be made robust to such ill-formed utterances.

This work was supported by the EU FP7 ICT Integrated Project "CogX" (FP7-ICT- 215181).

Pierre Lison and Geert-Jan M. Kruijff are with the German Research Centre for Artificial Intelligence (DFKI GmbH), Language Technology Lab, Saarbrücken, Germany {pierre.lison},{gj} @ dfki.de

Finally, the vast majority of spoken dialogue systems are designed to operate in *real-time*. This has two important consequences. First, the parser should not wait for the utterance to be complete to start processing it – instead, the set of possible semantic interpretations should be gradually built and extended as the utterance unfolds. Second, each incremental parsing step should operate under strict time constraints. The main obstacle here is the high level of ambiguity arising in natural language, which can lead to a combinatorial explosion in the number of possible readings.

The remaining of this paper is devoted to addressing this last issue, building on an integrated approach to situated spoken dialogue processing previously outlined in [3], [4]. The approach we present here is similar to [5], with some notable differences concerning the parser (our parser being specifically tailored for robust spoken dialogue processing), and the features included in the discriminative model.

An overview of the paper is as follows. We first describe in Section II the cognitive architecture in which our system has been integrated. We then discuss the approach in detail in Section III. Finally, we present in Section IV the quantitative evaluations on a WOZ test suite, and conclude.

II. Architecture

The approach we present in this paper is fully implemented and integrated into a cognitive architecture for autonomous robots. A recent description of the architecture is provided in [6], [7]. It is capable of building up visuo-spatial models of a dynamic local scene, and continuously plan and execute manipulation actions on objects within that scene. The robot can discuss objects and their material- and spatial properties for the purpose of visual learning and manipulation tasks. Figure 1 illustrates the architecture schema for the communication subsystem, limited to the comprehension side.

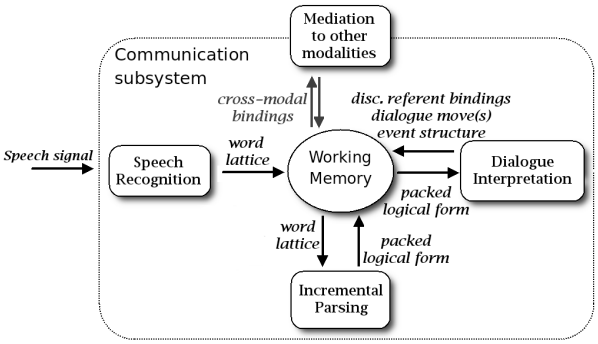


Fig. 1. Architecture schema of the communication subsystem (limited to the comprehension part).

Starting with ASR, we process the audio signal to establish a *word lattice* containing statistically ranked hypotheses about word sequences. Subsequently, parsing constructs grammatical analyses for the given (partial) word lattice. A grammatical analysis constructs both a syntactic analysis of the utterance, and a representation of its meaning. The analysis is based on an incremental chart parser[1] for Combinatory Categorial Grammar [8]. These meaning representations are ontologically richly sorted, relational structures, formulated in a (propositional) description logic – more precisely in the HLDS formalism [9]. The parser itself is based on a variant of the CKY algorithm [10].

Once all the possible (partial) parses for a given (partial) utterance are computed, they are filtered in order to retain only the most likely interpretation(s). This ensures that the number of parses at each incremental step remains bounded and avoid a combinatorial explosion of the search space. The task of selecting the most likely parse(s) among a set of possible ones is called *parse selection*. We describe it in detail in the next section.

At the level of dialogue interpretation, the logical forms are then resolved against a dialogue model to establish coreference and dialogue moves.

Linguistic interpretations must finally be associated with extra-linguistic knowledge about the environment – dialogue comprehension hence needs to connect with other subarchitectures like vision, spatial reasoning or planning. We realise this information binding between different modalities via a specific module, called the "binder", which is responsible for the ontology-based *mediation* accross modalities [11].

A. Context-sensitivity

The combinatorial nature of language provides virtually unlimited ways in which we can communicate meaning. This, of course, raises the question of how precisely an utterance should then be understood as it is being heard. Empirical studies have investigated what information humans use when comprehending spoken utterances. An important observation is that interpretation *in context* plays a crucial role in the comprehension of the utterance as it unfolds [12]. During utterance comprehension, humans combine linguistic information with scene understanding and "world knowledge" to select the most likely interpretation.

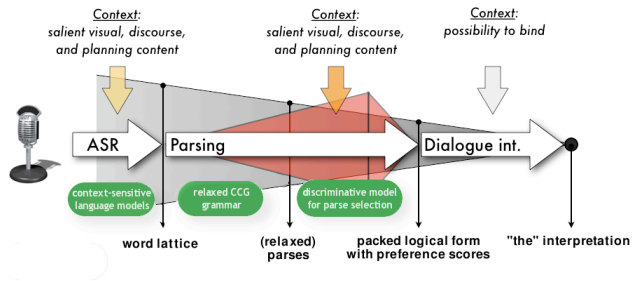


Fig. 2. Context-sensitivity in processing situated dialogue understanding

[1]Built using the OpenCCG API: http://openccg.sf.net

Several approaches in situated dialogue for human-robot interaction have made similar observations [13], [14], [15], [7]: A robot's understanding can be improved by relating utterances to the situated context. By incorporating contextual information into our model, our approach to robust processing of spoken dialogue seeks to exploit this important insight. At each processing step (speech recognition, word lattice parsing, dialogue-level interpretation and cross-modal binding), contextual information is used to prime the utterance comprehension, as shown in the Figure 2.

III. Approach

As we just explained, the parse selection module is responsible for selecting at each incremental step a subset of "good" parses. Once the selection is made, the best analyses are kept in the parse chart, while the others are discarded and pruned from the chart.

A. The parse selection task

To achieve this selection, we need a mechanism to discriminate among the possible parses. This is done via a (discriminative) statistical model covering a large number of features.

Formally, the task is defined as a function $F : \mathcal{X} \rightarrow \mathcal{Y}$ where the domain $\mathcal{X}$ is the set of possible inputs (in our case, $\mathcal{X}$ is the set of possible *word lattices*), and $\mathcal{Y}$ the set of parses. We assume:

1) A function $\mathbf{GEN}(x)$ which enumerates all possible parses for an input x. In our case, the function represents the admissibles parses according to the CCG grammar.
2) A d-dimensional feature vector $\mathbf{f}(x,y) \in \Re^d$, representing specific features of the pair (x,y). It can include various acoustic, syntactic, semantic or contextual features which can help us discriminate between the various parses.
3) A parameter vector $\mathbf{w} \in \Re^d$.

The function F, mapping a word lattice to its most likely parse, is then defined as:

$$F(x) = \underset{y \in \mathbf{GEN}(x)}{\operatorname{argmax}} \mathbf{w}^T \cdot \mathbf{f}(x,y) \qquad (1)$$

where $\mathbf{w}^T \cdot \mathbf{f}(x,y)$ is the inner product $\sum_{s=1}^{d} w_s \ f_s(x,y)$, and can be seen as a measure of the "quality" of the parse. Given the parameters $\mathbf{w}$, the optimal parse of a given utterance x can be therefore easily determined by enumerating all the parses generated by the grammar, extracting their features, computing the inner product $\mathbf{w}^T \cdot \mathbf{f}(x,y)$, and selecting the parse with the highest score.

The task of parse selection is an example of a *structured classification problem*, which is the problem of predicting an output y from an input x, where the output y has a rich internal structure. In the specific case of parse selection, x is a word lattice, and y a logical form.

B. Training data

In order to estimate the parameters $\mathbf{w}$, we need a set of training examples. Unfortunately, no corpus of situated dialogue adapted to our task domain is available to this day, let alone semantically annotated. The collection of in-domain data via Wizard of Oz experiments being a very costly and time-consuming process, we followed the approach advocated in [16] and *generated* a corpus from a hand-written task grammar.

To this end, we first collected a small set of WoZ data, totalling about a thousand utterances related to a simple scenario of object manipulation and visual learning. This set is too small to be directly used as a corpus for statistical training, but sufficient to capture the most frequent linguistic constructions in this particular context. Based on it, we designed a domain-specific context-free grammar covering most of the utterances. Each rule is associated to a semantic HLDS representation. Weights are automatically assigned to each grammar rule by parsing our corpus, hence leading to a small *stochastic context-free grammar* augmented with semantic information.

Once the grammar is specified, it is randomly traversed a large number of times, resulting in a larger set (about 25.000) of utterances along with their semantic representations. Since we are interested in handling errors arising from speech recognition, we also need to "simulate" the most frequent recognition errors. To this end, we *synthesise* each string generated by the domain-specific grammar, using a text-to-speech engine[2], feed the audio stream to the speech recogniser, and retrieve the recognition result.

Via this technique, we are able to easily collect a large amount of training data. Because of its relatively artificial character, the quality of such training data is naturally lower than what could be obtained with a genuine corpus. But, as the experimental results will show, it remains sufficient to train the perceptron for the parse selection task, and achieve significant improvements in accuracy and robustness. In a near future, we plan to progressively replace this generated training data by a real spoken dialogue corpus adapted to our task domain.

C. Perceptron learning

The algorithm we use to estimate the parameters $\mathbf{w}$ using the training data is a **perceptron**. The algorithm is fully online - it visits each example in turn, in an incremental fashion, and updates $\mathbf{w}$ if necessary. Albeit simple, the algorithm has proven to be very efficient and accurate for the task of parse selection [5], [17].

The pseudo-code for the online learning algorithm is detailed in [**Algorithm 1**].

It works as follows: the parameters $\mathbf{w}$ are first initialised to some arbitrary values. Then, for each pair (x_i, z_i) in the training set, the algorithm searchs for the parse y' with the highest score according to the current model. If this parse happens to match the best parse which generates z_i (which we shall denote y^*), we move to the next example. Else, we perform a simple perceptron update on the parameters:

$$\mathbf{w} = \mathbf{w} + \mathbf{f}(x_i, y^*) - \mathbf{f}(x_i, y') \quad (2)$$

The iteration on the training set is repeated T times, or until convergence. The most expensive step in this algorithm is the calculation of $y' = \text{argmax}_{y \in \mathbf{GEN}(x_i)} \mathbf{w}^T \cdot \mathbf{f}(x_i, y)$ - this is the *decoding* problem.

Algorithm 1 Online perceptron learning

Require: - Set of n training examples $\{(x_i, z_i) : i = 1...n\}$
- For each incremental step j with $0 \leq j \leq |x_i|$, we define the partially parsed utterance x_i^j and its gold standard semantics z_i^j
- T: number of iterations over the training set
- GEN(x): function enumerating possible parses for an input x, according to the CCG grammar.
- GEN(x, z): function enumerating possible parses for an input x and which have semantics z, according to the CCG grammar.
- $L(y)$ maps a parse tree y to its logical form.
- Initial parameter vector $\mathbf{w_0}$

```
% Initialise
w ← w_0
% Loop T times on the training examples
for t = 1 ... T do
  for i = 1 ... n do
    % Loop on the incremental parsing steps
    for j = 0...|x_i| do
      % Compute best parse according to model
      Let y' = argmax_{y∈GEN(x_i^j)} w^T · f(x_i^j, y)
      % If the decoded parse ≠ expected parse, update the parameters of the model
      if L(y') ≠ z_i^j then
        % Search the best parse for the partial utterance x_i^j with semantics z_i^j
        Let y* = argmax_{y∈GEN(x_i^j, z_i^j)} w^T · f(x_i^j, y)
        % Update parameter vector w
        Set w = w + f(x_i^j, y*) − f(x_i^j, y')
      end if
    end for
  end for
end for
return parameter vector w
```

It is possible to prove that, provided the training set (x_i, z_i) is separable with margin $\delta > 0$, the algorithm is assured to converge after a finite number of iterations to a model with zero training errors [5]. See also [18] for convergence theorems and proofs.

D. Features

As we have just seen, the parse selection operates by enumerating the possible parses and selecting the one with the highest score according to the linear model parametrised by the weights $\mathbf{w}$.

[2]We used MARY (http://mary.dfki.de) for the text-to-speech engine.

The accuracy of our method crucially relies on the selection of "good" features $\mathbf{f}(x,y)$ for our model - that is, features which help *discriminating* the parses. In our model, the features are of four types: semantic features, syntactic features, contextual features, and speech recognition features.

1) Semantic features: What are the substructures of a logical form which may be relevant to discriminate the parses? We define features on the following information sources: the nominals, the ontological sorts of the nominals, the dependency relations (following [19]), and the sequences of dependency relations.

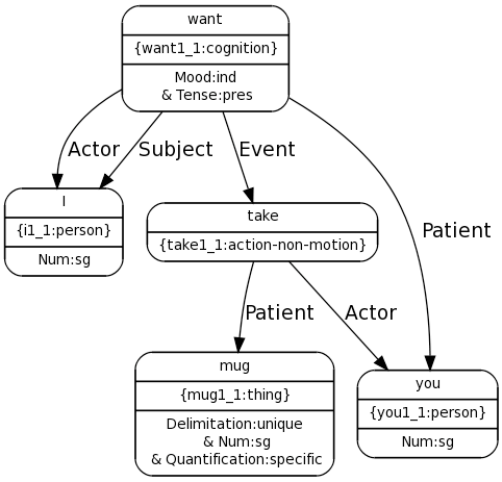


Fig. 3. HLDS logical form for "*I want you to take the mug*".

The features on nominals and ontological sorts aim at modeling (aspects of) *lexical semantics* - e.g. which meanings are the most frequent for a given word -, whereas the features on relations and sequence of relations focus on *sentential semantics* - which dependencies are the most frequent.

These features therefore help us handle various forms of lexical and syntactic ambiguities.

2) Syntactic features: Syntactic features are features associated to the *derivational history* of a specific parse. Alongside the usual CCG rules (application, composition and type raising), our parser also uses a set of non-standard rules designed to handle disfluencies, speech recognition errors, and combinations of discourse units by selectively relaxing the grammatical constraints (see [4] for details). In order to "*penalise*" to a correct extent the application of these non-standard rules, we include in the feature vector $\mathbf{f}(x,y)$ new features counting the number of times these rules are applied in the parse. In the derivation shown in the Figure 4, the rule *corr* (correction of a speech recognition error) is for instance applied once.

These syntactic features can be seen as a *penalty* given to the parses using these non-standard rules, thereby giving a preference to the "normal" parses over them.

This ensures that the grammar relaxation is only applied

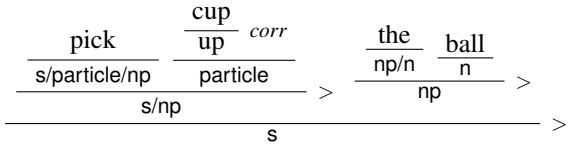


Fig. 4. CCG derivation of *"pick cup the ball"*.

"as a last resort" when the usual grammatical analysis fails to provide a parse.

3) Contextual features: As we have already outlined in the background section, one striking characteristic of spoken dialogue is the importance of *context*. Understanding the visual and discourse contexts is crucial to resolve potential ambiguities and compute the most likely interpretation(s) of a given utterance.

The feature vector $\mathbf{f}(x,y)$ therefore includes various features related to the context:

- *Activated words*: our dialogue system maintains in its working memory a list of contextually activated words (cfr. [20]). This list is continuously updated as the dialogue and the environment evolves. For each context-dependent word, we include one feature counting the number of times it appears in the utterance string.
- *Expected dialogue moves*: for each possible dialogue move, we include one feature indicating if the dialogue move is consistent with the current discourse model. These features ensure for instance that the dialogue move following a QuestionYN is a Accept, Reject or another question (e.g. for clarification requests), but almost never an Opening.

4) Speech recognition features: Finally, the feature vector $\mathbf{f}(x,y)$ also includes features related to the *speech recognition*. The ASR module outputs a set of (partial) recognition hypotheses, packed in a word lattice. One example is given in Figure 5.

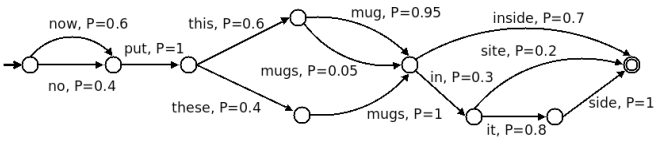


Fig. 5. Example of word lattice

We want to favour the hypotheses with high confidence scores, which are, according to the statistical models incorporated in the ASR, more likely to reflect what was uttered. To this end, we introduce in the feature vector several acoustic features measuring the likelihood of each recognition hypothesis.

E. Incremental chart pruning

In the previous subsections, we explained how the parse selection was performed, and on basis of which features.

	Beam width	Size of word lattice	Average parsing time (in s.)	Exact-match			Partial-match		
				Precision	Recall	F_1-value	Precision	Recall	F_1-value
(Baseline)	*(none)*	*10*	*10.1*	*40.4*	*100.0*	*57.5*	*81.4*	*100.0*	*89.8*
	120	10	5.78	40.9	96.9	57.5	81.9	98.0	89.2
	60	10	**4.82**	41.1	92.5	56.9	81.7	94.1	87.4
	40	10	4.66	39.9	88.1	54.9	79.6	91.9	85.3
	30	10	4.21	41.0	83.0	54.9	80.2	88.6	84.2
	20	10	4.30	40.1	80.3	53.5	78.9	86.5	82.5
(Baseline)	*(none)*	*5*	*5.28*	*40.0*	*100.0*	*57.1*	*81.5*	*100.0*	*89.8*
	120	5	6.62	40.9	98.4	57.8	81.6	98.5	89.3
	60	5	5.28	40.5	96.9	57.1	81.7	97.1	88.7
	40	5	4.26	40.9	91.0	56.5	81.7	92.4	86.7
	30	5	**3.51**	40.7	92.4	56.5	81.4	93.9	87.2
	20	5	2.81	36.7	87.1	51.7	79.6	90.7	84.8

TABLE I

EVALUATION RESULTS (IN SECONDS FOR THE PARSING TIME, IN % FOR THE EXACT- AND PARTIAL-MATCH).

This parse selection is used at each incremental step to discriminate between the "good" parses that needs to be kept in the parse chart, and the parses that should be pruned in order to keep a limited number of interpretations, and hence avoid a combinatory explosion of analyses.

To achieve this, we introduce a new parameter in our parser: the *beam width*. The beam width defines the maximal number of analyses which can be kept in the chart at each incremental step. If the number of possible readings exceeds the beam width, the analyses with a lower parse selection score are removed from the chart.

Practically, this is realised by removing the top signs associated in the chart with the set of analyses to prune, as well as all the intermediate signs which are included in these top signs *and* are not used in any of the "good" analyses retained by the parse selection module.

A simple backtracking mechanism is also implemented in the parser. In case the beam width happens to be too narrow and renders the utterance unparsable, it is possible to reintroduce the signs previously removed from the chart and restart the parse at the failure point.

The combination of incremental parsing and incremental chart pruning provides two decisive advantages over classical, non-incremental parsing techniques: first, we can start processing the spoken inputs as soon as a partial analysis can be outputted by the speech recogniser. Second, the pruning mechanism ensures that each incremental parsing step remains time-bounded. Such a combination is therefore ideally suited for the real-time spoken dialogue systems used in human-robot interaction.

IV. EVALUATION

We performed a quantitative evaluation of our approach, using its implementation in a fully integrated system (cf. Section II). To set up the experiments for the evaluation, we have gathered a Wizard-of-Oz corpus of human-robot spoken dialogue for our task-domain (Figure 6), which we segmented and annotated manually with their expected semantic interpretation. The data set contains 195 individual utterances [3] along with their complete logical forms.

The results are shown in the Table I. We tested our approach for five different values of the beam width parameter, and for two sizes of the word lattice. The results are compared against a baseline, which is the performance of our parser without chart pruning. For each configuration, we give the average parsing time, as well as the exact-match and partial-match results (in order to verify that the performance increase is not cancelled by a drop in accuracy). The most important observation we can make is that the choice of the beam width parameter is crucial. Above 30, the chart pruning mechanism works very efficiently – we observe a notable decrease in the parsing time without significantly affecting the accuracy performance. Below 30, the beam width is too small to retain all the necessary information in the chart, and the recall quickly drops.

Figure 7 illustrates the evolution of the ambiguity level (in terms of number of alternative semantic interpretations) during the incremental parsing. We observe that the chart pruning mechanism acts as a *stabilising factor* within the parser, by limiting the number of ambiguities produced after every incremental step to a reasonable level.

Fig. 6. Wizard-of-Oz experiments for a task domain of object manipulation and visual learning

[3] More precisely, word lattices provided by the speech recogniser. These word lattices can contain a maximum of 10 recognition hypotheses.

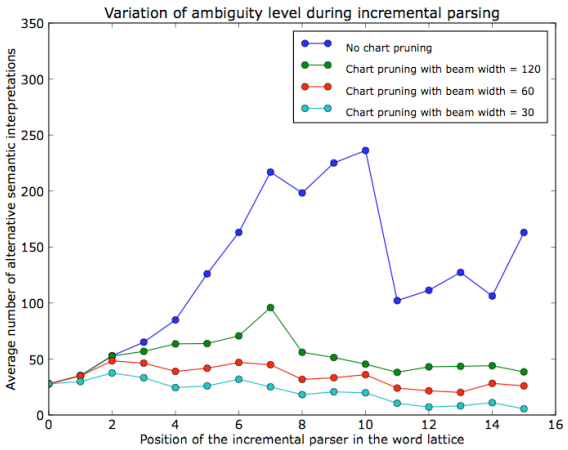


Fig. 7. Variation of ambiguity level during incremental parsing, with and without chart pruning (on word lattices with NBest 10 hypotheses).

V. Conclusions

We presented in this paper an original mechanism for efficient parsing of spoken inputs, based on a combination of incremental *parsing* (to start the processing as soon as a partial speech input is recognised) and incremental *chart pruning* (to limit at every step the number of analyses retained in the parse chart).

The incremental parser is based on a fine-grained Combinatory Categorial Grammar, and takes ASR word lattices as input. It outputs a set of partial semantic interpretations ("logical forms"), which are progressively refined and extended as the utterance unfolds.

Once the partial interpretations are computed, they are subsequently pruned/filtered to keep only the most likely hypotheses in the parse chart. This mechanism is based on a *discriminative model* exploring a set of relevant semantic, syntactic, contextual and acoustic features extracted for each parse. At each incremental step, the discriminative model yields a score for each resulting parse. The parser then only retains in its chart the set of parses associated with a high score, the others being pruned.

The experimental evaluation conducted on a Wizard-of-Oz test suite demonstrated that the aforementioned approach was able to significantly improve the parser performance .

As forthcoming work, we shall examine the extension of our approach in new directions, such as the introduction of more refined contextual features, the extension of the grammar relaxation rules, or the use of more sophisticated learning algorithms such as Support Vector Machines.

References

[1] R. Fernández and J. Ginzburg, "A corpus study of non-sentential utterances in dialogue," *Traitement Automatique des Langues*, vol. 43, no. 2, pp. 12–43, 2002.

[2] E. A. Topp, H. Hüttenrauch, H. Christensen, and K. Severinson Eklundh, "Bringing together human and robotic environment representations – a pilot study," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, October 2006.

[3] P. Lison, "Robust processing of situated spoken dialogue," Master's thesis, Universität des Saarlandes, Saarbrücken, 2008, http://www.dfki.de/∼ plison/pubs/thesis/main.thesis.plison2008.pdf.

[4] P. Lison and G.-J. M. Kruijff, "An integrated approach to robust processing of situated spoken dialogue," in *Proceedings of the International Workshop on Semantic Representation of Spoken Language (SRSL'09)*, Athens, Greece, 2009, (to appear).

[5] M. Collins and B. Roark, "Incremental parsing with the perceptron algorithm," in *ACL '04: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2004, p. 111.

[6] N. A. Hawes, A. Sloman, J. Wyatt, M. Zillich, H. Jacobsson, G.-J. M. Kruijff, M. Brenner, G. Berginc, and D. Skocaj, "Towards an integrated robot with multiple cognitive functions." in *Proc. AAAI'07*. AAAI Press, 2007, pp. 1548–1553.

[7] G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes, "Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction," in *Language and Robots: Proceedings from the Symposium (LangRo'2007)*, Aveiro, Portugal, December 2007, pp. 55–64.

[8] M. Steedman and J. Baldridge, "Combinatory categorial grammar," in *Nontransformational Syntax: A Guide to Current Models*, R. Borsley and K. Börjars, Eds. Oxford: Blackwell, 2009.

[9] J. Baldridge and G.-J. M. Kruijff, "Coupling CCG and hybrid logic dependency semantics," in *ACL'02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA: Association for Computational Linguistics, 2002, pp. 319–326.

[10] T. Kasami, "An efficient recognition and syntax analysis algorithm for context free languages," Air Force Cambridge Research Laboratory, Bedford, Massachussetts, Scientific Report AF CRL-65-758, 1965.

[11] H. Jacobsson, N. Hawes, G.-J. Kruijff, and J. Wyatt, "Crossmodal content binding in information-processing architectures," in *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands, March 12–15 2008.

[12] P. Knoeferle and M. Crocker, "The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking," *Cognitive Science*, 2006.

[13] D. Roy, "Semiotic schemas: A framework for grounding language in action and perception," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 170–205, 2005.

[14] D. Roy and N. Mukherjee, "Towards situated speech understanding: visual context priming of language models," *Computer Speech & Language*, vol. 19, no. 2, pp. 227–248, April 2005.

[15] T. Brick and M. Scheutz, "Incremental natural language processing for HRI," in *Proceeding of the ACM/IEEE international conference on Human-Robot Interaction (HRI'07)*, 2007, pp. 263 – 270.

[16] K. Weilhammer, M. N. Stuttle, and S. Young, "Bootstrapping language models for dialogue systems," in *Proceedings of INTERSPEECH 2006*, Pittsburgh, PA, 2006.

[17] L. S. Zettlemoyer and M. Collins, "Online learning of relaxed CCG grammars for parsing to logical form," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 678–687.

[18] M. Collins, "Parameter estimation for statistical parsing models: theory and practice of distribution-free methods," in *New developments in parsing technology*. Kluwer Academic Publishers, 2004, pp. 19–55.

[19] S. Clark and J. R. Curran, "Log-linear models for wide-coverage ccg parsing," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 97–104.

[20] P. Lison and G.-J. M. Kruijff, "Salience-driven contextual priming of speech recognition for human-robot interaction," in *Proceedings of the 18th European Conference on Artificial Intelligence*, Patras (Greece), 2008.

An Integrated Approach to Robust Processing of Situated Spoken Dialogue

Pierre Lison
Language Technology Lab,
DFKI GmbH,
Saarbrücken, Germany
pierre.lison@dfki.de

Geert-Jan M. Kruijff
Language Technology Lab,
DFKI GmbH,
Saarbrücken, Germany
gj@dfki.de

Abstract

Spoken dialogue is notoriously hard to process with standard NLP technologies. Natural spoken dialogue is replete with disfluent, partial, elided or ungrammatical utterances, all of which are difficult to accommodate in a dialogue system. Furthermore, speech recognition is known to be a highly error-prone task, especially for complex, open-ended domains. The combination of these two problems – ill-formed and/or misrecognised speech inputs – raises a major challenge to the development of robust dialogue systems.

We present an integrated approach for addressing these two issues, based on an incremental parser for Combinatory Categorial Grammar. The parser takes word lattices as input and is able to handle ill-formed and misrecognised utterances by selectively relaxing its set of grammatical rules. The choice of the most relevant interpretation is then realised via a discriminative model augmented with contextual information. The approach is fully implemented in a dialogue system for autonomous robots. Evaluation results on a Wizard of Oz test suite demonstrate very significant improvements in accuracy and robustness compared to the baseline.

1 Introduction

Spoken dialogue is often considered to be one of the most natural means of interaction between a human and a robot. It is, however, notoriously hard to process with standard language processing technologies. Dialogue utterances are often incomplete or ungrammatical, and may contain numerous disfluencies like fillers (err, uh, mm), repetitions, self-corrections, etc. Rather than getting crisp-and-clear commands such as "*Put the red ball inside the box!*", it is more likely the robot will hear such kind of utterance: "*right, now, could you, uh, put the red ball, yeah, inside the ba/ box!*". This is natural behaviour in human-human interaction (Fernández and Ginzburg, 2002) and can also be observed in several domain-specific corpora for human-robot interaction (Topp et al., 2006).

Moreover, even in the (rare) case where the utterance is perfectly well-formed and does not contain any kind of disfluencies, the dialogue system still needs to accomodate the various speech recognition errors thay may arise. This problem is particularly acute for robots operating in real-world noisy environments and deal with utterances pertaining to complex, open-ended domains.

The paper presents a new approach to address these two difficult issues. Our starting point is the work done by Zettlemoyer and Collins on parsing using relaxed CCG grammars (Zettlemoyer and Collins, 2007) (ZC07). In order to account for natural spoken language phenomena (more flexible word order, missing words, etc.), they augment their grammar framework with a small set of non-standard combinatory rules, leading to a *relaxation* of the grammatical constraints. A discriminative model over the parses is coupled with the parser, and is responsible for selecting the most likely interpretation(s) among the possible ones.

In this paper, we extend their approach in two important ways. First, ZC07 focused on the treatment of ill-formed input, and ignored the speech recognition issues. Our system, to the contrary, is able to deal with both ill-formed and misrecognized input, in an integrated fashion. This is done by augmenting the set of non-standard combinators with new rules specifically tailored to deal with speech recognition errors.

Second, the only features used by ZC07 are syntactic features (see section 3.4 for details). We significantly extend the range of features included

in the discriminative model, by incorporating not only *syntactic*, but also *acoustic*, *semantic* and *contextual* information into the model. As the experimental results have shown, the inclusion of a broader range of linguistic and contextual information leads to a more accurate discrimination of the various interpretations.

An overview of the paper is as follows. We first describe in Section 2 the cognitive architecture in which our system has been integrated. We then discuss the approach in detail in Section 3. Finally, we present in Section 4 the quantitative evaluations on a WOZ test suite, and conclude.

2 Architecture

The approach we present in this paper is fully implemented and integrated into a cognitive architecture for autonomous robots. A recent version of this system is described in (Hawes et al., 2007). It is capable of building up visuo-spatial models of a dynamic local scene, and continuously plan and execute manipulation actions on objects within that scene. The robot can discuss objects and their material- and spatial properties for the purpose of visual learning and manipulation tasks.

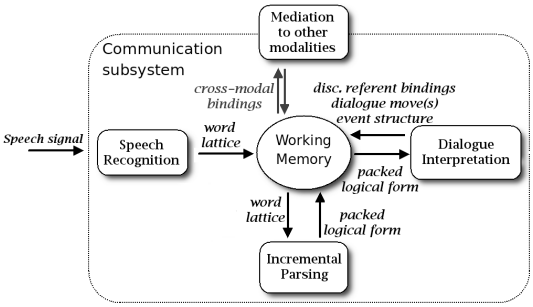


Figure 1: Architecture schema of the communication subsystem (only for comprehension).

Figure 2 illustrates the architecture schema for the communication subsystem incorporated in the cognitive architecture (only the comprehension part is shown).

Starting with ASR, we process the audio signal to establish a *word lattice* containing statistically ranked hypotheses about word sequences. Subsequently, parsing constructs grammatical analyses for the given word lattice. A grammatical analysis constructs both a syntactic analysis of the utterance, and a representation of its meaning. The analysis is based on an incremental chart parser[1] for Combinatory Categorial Grammar (Steedman and Baldridge, 2009). These meaning representations are ontologically richly sorted, relational structures, formulated in a (propositional) description logic, more precisely in the HLDS formalism (Baldridge and Kruijff, 2002). The parser compacts all meaning representations into a single *packed logical form* (Carroll and Oepen, 2005; Kruijff et al., 2007). A packed LF represents content similar across the different analyses as a single graph, using over- and underspecification of how different nodes can be connected to capture lexical and syntactic forms of ambiguity.

At the level of dialogue interpretation, a packed logical form is resolved against a SDRS-like dialogue model (Asher and Lascarides, 2003) to establish co-reference and dialogue moves.

Linguistic interpretations must finally be associated with extra-linguistic knowledge about the environment – dialogue comprehension hence needs to connect with other subarchitectures like vision, spatial reasoning or planning. We realise this information binding between different modalities via a specific module, called the "binder", which is responsible for the ontology-based *mediation* across modalities (Jacobsson et al., 2008).

2.1 Context-sensitivity

The combinatorial nature of language provides virtually unlimited ways in which we can communicate meaning. This, of course, raises the question of how precisely an utterance should then be understood as it is being heard. Empirical studies have investigated what information humans use when comprehending spoken utterances. An important observation is that interpretation *in context* plays a crucial role in the comprehension of utterance as it unfolds (Knoeferle and Crocker, 2006). During utterance comprehension, humans combine linguistic information with scene understanding and "world knowledge".

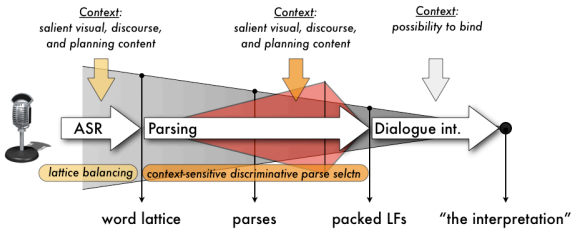


Figure 2: Context-sensitivity in processing situated dialogue understanding

[1] Built using the OpenCCG API: http://openccg.sf.net

Several approaches in situated dialogue for human-robot interaction have made similar observations (Roy, 2005; Roy and Mukherjee, 2005; Brick and Scheutz, 2007; Kruijff et al., 2007): A robot's understanding can be improved by relating utterances to the situated context. As we will see in the next section, by incorporating contextual information into our model, our approach to robust processing of spoken dialogue seeks to exploit this important insight.

3 Approach

3.1 Grammar relaxation

Our approach to robust processing of spoken dialogue rests on the idea of **grammar relaxation**: the grammatical constraints specified in the grammar are "relaxed" to handle slightly ill-formed or misrecognised utterances.

Practically, the grammar relaxation is done via the introduction of *non-standard CCG rules* (Zettlemoyer and Collins, 2007). In Combinatory Categorial Grammar, the rules are used to assemble categories to form larger pieces of syntactic and semantic structure. The standard rules are application ($<, >$), composition ($\mathbf{B}$), and type raising ($\mathbf{T}$) (Steedman and Baldridge, 2009).

Several types of non-standard rules have been introduced. We describe here the two most important ones: the *discourse-level composition rules*, and the *ASR correction rules*. We invite the reader to consult (Lison, 2008) for more details on the complete set of grammar relaxation rules.

3.1.1 Discourse-level composition rules

In natural spoken dialogue, we may encounter utterances containing several independent "chunks" without any explicit separation (or only a short pause or a slight change in intonation), such as

(1) "yes take the ball no the other one on your left right and now put it in the box."

Even if retrieving a fully structured parse for this utterance is difficult to achieve, it would be useful to have access to a list of smaller "discourse units". Syntactically speaking, a discourse unit can be any type of saturated atomic categories - from a simple discourse marker to a full sentence.

The type-changing rule $\mathbf{T}_{du}$ allows the conversion of atomic categories into discourse units:

$$\mathsf{A} : @_i f \Rightarrow \mathsf{du} : @_i f \qquad (\mathbf{T}_{du})$$

where A represents an arbitrary saturated atomic category (s, np, pp, etc.).

The rule $\mathbf{T}_C$ is a type-changing rule which allows us to integrate two discourse units into a single structure:

$$\mathsf{du} : @_a x \Rightarrow \mathsf{du} : @_c z \;/\; \mathsf{du} : @_b y \qquad (\mathbf{T}_C)$$

where the formula $@_c z$ is defined as:

$$\begin{aligned} @_{\{c:\text{d-units}\}}(&\mathbf{list} \wedge \\ &(\langle \text{FIRST} \rangle\ a \wedge x) \wedge \\ &(\langle \text{NEXT} \rangle\ b \wedge y)) \end{aligned} \qquad (2)$$

3.1.2 ASR error correction rules

Speech recognition is a highly error-prone task. It is however possible to partially alleviate this problem by inserting new error-correction rules (more precisely, new lexical entries) for the most frequently misrecognised words.

If we notice e.g. that the ASR system frequently substitutes the word "wrong" for the word "round" during the recognition (because of their phonological proximity), we can introduce a new lexical entry in the lexicon in order to correct this error:

$$round \vdash \mathsf{adj} : @_{attitude}(\mathbf{wrong}) \qquad (3)$$

A set of thirteen new lexical entries of this type have been added to our lexicon to account for the most frequent recognition errors.

3.2 Parse selection

Using more powerful grammar rules to relax the grammatical analysis tends to increase the number of parses. We hence need a mechanism to discriminate among the possible parses. The task of selecting the most likely interpretation among a set of possible ones is called *parse selection*. Once all the possible parses for a given utterance are computed, they are subsequently filtered or selected in order to retain only the most likely interpretation(s). This is done via a (discriminative) statistical model covering a large number of features.

Formally, the task is defined as a function $F : \mathcal{X} \rightarrow \mathcal{Y}$ where the domain $\mathcal{X}$ is the set of possible inputs (in our case, $\mathcal{X}$ is the set of possible *word lattices*), and $\mathcal{Y}$ the set of parses. We assume:

1. A function $\mathbf{GEN}(x)$ which enumerates all possible parses for an input x. In our case, this function simply represents the set of parses of x which are admissible according to the CCG grammar.

2. A d-dimensional feature vector $\mathbf{f}(x, y) \in \Re^d$, representing specific features of the pair (x, y). It can include various acoustic, syntactic, semantic or contextual features which can be relevant in discriminating the parses.

3. A parameter vector $\mathbf{w} \in \Re^d$.

The function F, mapping a word lattice to its most likely parse, is then defined as:

$$F(x) = \underset{y \in \mathbf{GEN}(x)}{\operatorname{argmax}} \mathbf{w}^T \cdot \mathbf{f}(x, y) \qquad (4)$$

where $\mathbf{w}^T \cdot \mathbf{f}(x, y)$ is the inner product $\sum_{s=1}^{d} w_s \ f_s(x, y)$, and can be seen as a measure of the "quality" of the parse. Given the parameters $\mathbf{w}$, the optimal parse of a given utterance x can be therefore easily determined by enumerating all the parses generated by the grammar, extracting their features, computing the inner product $\mathbf{w}^T \cdot \mathbf{f}(x, y)$, and selecting the parse with the highest score.

The task of parse selection is an example of a *structured classification problem*, which is the problem of predicting an output y from an input x, where the output y has a rich internal structure. In the specific case of parse selection, x is a word lattice, and y a logical form.

3.3 Learning

3.3.1 Training data

In order to estimate the parameters $\mathbf{w}$, we need a set of training examples. Unfortunately, no corpus of situated dialogue adapted to our task domain is available to this day, let alone semantically annotated. The collection of in-domain data via Wizard of Oz experiments being a very costly and time-consuming process, we followed the approach advocated in (Weilhammer et al., 2006) and *generated* a corpus from a hand-written task grammar.

To this end, we first collected a small set of WoZ data, totalling about a thousand utterances. This set is too small to be directly used as a corpus for statistical training, but sufficient to capture the most frequent linguistic constructions in this particular context. Based on it, we designed a domain-specific CFG grammar covering most of the utterances. Each rule is associated to a semantic HLDS representation. Weights are automatically assigned to each grammar rule by parsing our corpus, hence leading to a small *stochastic CFG grammar* augmented with semantic information.

Once the grammar is specified, it is randomly traversed a large number of times, resulting in a larger set (about 25.000) of utterances along with their semantic representations. Since we are interested in handling errors arising from speech recognition, we also need to "simulate" the most frequent recognition errors. To this end, we *synthesise* each string generated by the domain-specific CFG grammar, using a text-to-speech engine[2], feed the audio stream to the speech recogniser, and retrieve the recognition result. Via this technique, we are able to easily collect a large amount of training data[3].

3.3.2 Perceptron learning

The algorithm we use to estimate the parameters $\mathbf{w}$ using the training data is a **perceptron**. The algorithm is fully online - it visits each example in turn and updates $\mathbf{w}$ if necessary. Albeit simple, the algorithm has proven to be very efficient and accurate for the task of parse selection (Collins and Roark, 2004; Collins, 2004; Zettlemoyer and Collins, 2005; Zettlemoyer and Collins, 2007).

The pseudo-code for the online learning algorithm is detailed in [**Algorithm 1**].

It works as follows: the parameters $\mathbf{w}$ are first initialised to some arbitrary values. Then, for each pair (x_i, z_i) in the training set, the algorithm searchs for the parse y' with the highest score according to the current model. If this parse happens to match the best parse which generates z_i (which we shall denote y^*), we move to the next example. Else, we perform a simple perceptron update on the parameters:

$$\mathbf{w} = \mathbf{w} + \mathbf{f}(x_i, y^*) - \mathbf{f}(x_i, y') \qquad (5)$$

The iteration on the training set is repeated T times, or until convergence.

The most expensive step in this algorithm is the calculation of $y' = \operatorname{argmax}_{y \in \mathbf{GEN}(x_i)} \mathbf{w}^T \cdot \mathbf{f}(x_i, y)$ - this is the *decoding* problem.

It is possible to prove that, provided the training set (x_i, z_i) is separable with margin $\delta > 0$, the

[2] We used `MARY (http://mary.dfki.de)` for the text-to-speech engine.

[3] Because of its relatively artificial character, the quality of such training data is naturally lower than what could be obtained with a genuine corpus. But, as the experimental results will show, it remains sufficient to train the perceptron for the parse selection task, and achieve significant improvements in accuracy and robustness. In a near future, we plan to progressively replace this generated training data by a real spoken dialogue corpus adapted to our task domain.

algorithm is assured to converge after a finite number of iterations to a model with zero training errors (Collins and Roark, 2004). See also (Collins, 2004) for convergence theorems and proofs.

Algorithm 1 Online perceptron learning

Require: - set of n training examples $\{(x_i, z_i) : i = 1...n\}$
- T: number of iterations over the training set
- $\texttt{GEN}(x)$: function enumerating possible parses for an input x, according to the CCG grammar.
- $\texttt{GEN}(x, z)$: function enumerating possible parses for an input x and which have semantics z, according to the CCG grammar.
- $L(y)$ maps a parse tree y to its logical form.
- Initial parameter vector $\mathbf{w_0}$

% Initialise
$\mathbf{w} \leftarrow \mathbf{w_0}$
% Loop T times on the training examples
for $t = 1...T$ **do**
  for $i = 1...n$ **do**
    % Compute best parse according to current model
    Let $y' = \text{argmax}_{y \in \mathbf{GEN}(x_i)} \mathbf{w}^T \cdot \mathbf{f}(x_i, y)$
    % If the decoded parse $\neq$ expected parse, update the parameters
    if $L(y') \neq z_i$ **then**
      % Search the best parse for utterance x_i with semantics z_i
      Let $y^* = \text{argmax}_{y \in \mathbf{GEN}(x_i, z_i)} \mathbf{w}^T \cdot \mathbf{f}(x_i, y)$
      % Update parameter vector $\mathbf{w}$
      Set $\mathbf{w} = \mathbf{w} + \mathbf{f}(x_i, y^*) - \mathbf{f}(x_i, y')$
    end if
  end for
end for
return parameter vector $\mathbf{w}$

3.4 Features

As we have seen, the parse selection operates by enumerating the possible parses and selecting the one with the highest score according to the linear model parametrised by $\mathbf{w}$.

The accuracy of our method crucially relies on the selection of "good" features $\mathbf{f}(x, y)$ for our model - that is, features which help *discriminating* the parses. They must also be relatively cheap to compute. In our model, the features are of four types: semantic features, syntactic features, contextual features, and speech recognition features.

3.4.1 Semantic features

What are the substructures of a logical form which may be relevant to discriminate the parses? We define features on the following information sources:

1. *Nominals*: for each possible pair $\langle prop, sort \rangle$, we include a feature f_i in $\mathbf{f}(x, y)$ counting the number of nominals with ontological sort $sort$ and proposition $prop$ in the logical form.

2. *Ontological sorts*: occurrences of specific ontological sorts in the logical form.

3. *Dependency relations*: following (Clark and Curran, 2003), we also model the *dependency structure* of the logical form. Each dependency relation is defined as a triple $\langle sort_a, sort_b, label \rangle$, where $sort_a$ denotes the sort of the incoming nominal, $sort_b$ the sort of the outgoing nominal, and $label$ is the relation label.

4. *Sequences of dependency relations*: number of occurrences of particular sequences (ie. bigram counts) of dependency relations.

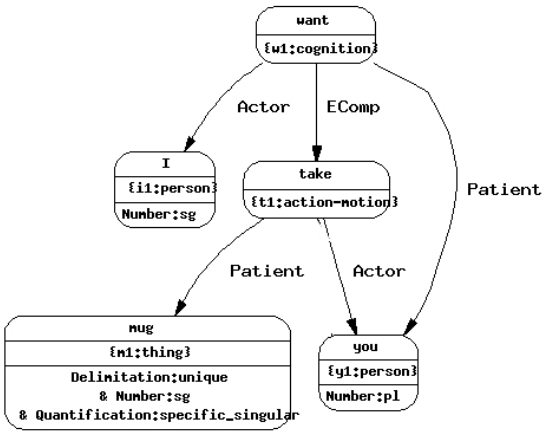


Figure 3: graphical representation of the HLDS logical form for "*I want you to take the mug*".

The features on nominals and ontological sorts aim at modeling (aspects of) *lexical semantics* - e.g. which meanings are the most frequent for a given word -, whereas the features on relations and sequence of relations focus on *sentential semantics* - which dependencies are the most frequent. These features therefore help us handle lexical and syntactic ambiguities.

3.4.2 Syntactic features

By "syntactic features", we mean features associated to the *derivational history* of a specific parse. The main use of these features is to *penalise* to a

correct extent the application of the non-standard rules introduced into the grammar.

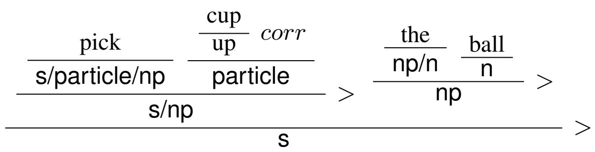


Figure 4: CCG derivation of *"pick cup the ball"*.

To this end, we include in the feature vector $\mathbf{f}(x, y)$ a new feature for each non-standard rule, which counts the number of times the rule was applied in the parse.

In the derivation shown in the figure 4, the rule *corr* (correction of a speech recognition error) is applied once, so the corresponding feature value is set to 1. The feature values for the remaining rules are set to 0, since they are absent from the parse.

These syntactic features can be seen as a *penalty* given to the parses using these non-standard rules, thereby giving a preference to the "normal" parses over them. This mechanism ensures that the grammar relaxation is only applied "as a last resort" when the usual grammatical analysis fails to provide a full parse. Of course, depending on the relative frequency of occurrence of these rules in the training corpus, some of them will be more strongly penalised than others.

3.4.3 Contextual features

As we have already outlined in the background section, one striking characteristic of spoken dialogue is the importance of *context*. Understanding the visual and discourse contexts is crucial to resolve potential ambiguities and compute the most likely interpretation(s) of a given utterance.

The feature vector $\mathbf{f}(x, y)$ therefore includes various features related to the context:

1. *Activated words*: our dialogue system maintains in its working memory a list of contextually activated words (cfr. (Lison and Kruijff, 2008)). This list is continuously updated as the dialogue and the environment evolves. For each context-dependent word, we include one feature counting the number of times it appears in the utterance string.

2. *Expected dialogue moves*: for each possible dialogue move, we include one feature indicating if the dialogue move is consistent with the current discourse model. These features ensure for instance that the dialogue move following a QuestionYN is a Accept, Reject or another question (e.g. for clarification requests), but almost never an Opening.

3. *Expected syntactic categories*: for each atomic syntactic category in the CCG grammar, we include one feature indicating if the category is consistent with the current discourse model. These features can be used to handle *sentence fragments*.

3.4.4 Speech recognition features

Finally, the feature vector $\mathbf{f}(x, y)$ also includes features related to the *speech recognition*. The ASR module outputs a set of (partial) recognition hypotheses, packed in a word lattice. One example of such a structure is given in Figure 5. Each recognition hypothesis is provided with an associated confidence score, and we want to favour the hypotheses with high confidence scores, which are, according to the statistical models incorporated in the ASR, more likely to reflect what was uttered.

To this end, we introduce three features: the *acoustic confidence score* (confidence score provided by the statistical models included in the ASR), the *semantic confidence score* (based on a "concept model" also provided by the ASR), and the *ASR ranking* (hypothesis rank in the word lattice, from best to worst).

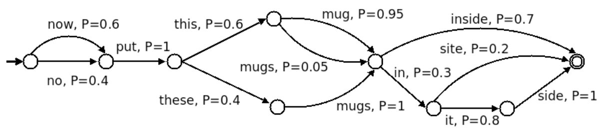


Figure 5: Example of word lattice

4 Experimental evaluation

We performed a quantitative evaluation of our approach, using its implementation in a fully integrated system (cf. Section 2). To set up the experiments for the evaluation, we have gathered a corpus of human-robot spoken dialogue for our task-domain, which we segmented and annotated manually with their expected semantic interpretation. The data set contains 195 individual utterances along with their complete logical forms.

4.1 Results

Three types of quantitative results are extracted from the evaluation results: *exact-match*, *partial-*

	Size of word lattice (number of NBests)	Grammar relaxation	Parse selection	Precision	Recall	F_1-value
(Baseline)	1	No	No	40.9	45.2	**43.0**
.	1	No	Yes	59.0	54.3	56.6
.	1	Yes	Yes	52.7	70.8	60.4
.	3	Yes	Yes	55.3	82.9	66.3
.	5	Yes	Yes	55.6	84.0	66.9
(Full approach)	10	Yes	Yes	55.6	84.9	**67.2**

Table 1: Exact-match accuracy results (in percents).

	Size of word lattice (number of NBests)	Grammar relaxation	Parse selection	Precision	Recall	F_1-value
(Baseline)	1	No	No	86.2	56.2	**68.0**
.	1	No	Yes	87.4	56.6	68.7
.	1	Yes	Yes	88.1	76.2	81.7
.	3	Yes	Yes	87.6	85.2	86.4
.	5	Yes	Yes	87.6	86.0	86.8
(Full approach)	10	Yes	Yes	87.7	87.0	**87.3**

Table 2: Partial-match accuracy results (in percents).

match, and *word error rate*. Tables 1, 2 and 3 illustrate the results, broken down by use of grammar relaxation, use of parse selection, and number of recognition hypotheses considered.

Each line in the tables corresponds to a possible configuration. Tables 1 and 2 give the precision, recall and F_1 value for each configuration (respectively for the exact- and partial-match), and Table 3 gives the Word Error Rate [WER].

The first line corresponds to the baseline: no grammar relaxation, no parse selection, and use of the first NBest recognition hypothesis. The last line corresponds to the results with the full approach: grammar relaxation, parse selection, and use of 10 recognition hypotheses.

Size of word lattice (NBests)	Grammar relaxation	Parse selection	WER
1	No	No	**20.5**
1	Yes	Yes	19.4
3	Yes	Yes	16.5
5	Yes	Yes	15.7
10	Yes	Yes	**15.7**

Table 3: Word error rate (in percents).

4.2 Comparison with baseline

Here are the comparative results we obtained:

- Regarding the exact-match results between the baseline and our approach (grammar relaxation and parse selection with all features activated for NBest 10), the F_1-measure climbs from 43.0 % to 67.2 %, which means a relative difference of $\mathbf{56.3}$ %.
- For the partial-match, the F_1-measure goes from 68.0 % for the baseline to 87.3 % for our approach – a relative increase of $\mathbf{28.4}$ %.
- We observe a significant decrease in WER: we go from 20.5 % for the baseline to 15.7 % with our approach. The difference is statistically significant (p-value for t-tests is 0.036), and the relative decrease of $\mathbf{23.4}$ %.

5 Conclusions

We presented an *integrated* approach to the processing of (situated) spoken dialogue, suited to the specific needs and challenges encountered in human-robot interaction.

In order to handle disfluent, partial, ill-formed or misrecognized utterances, the grammar used by the parser is "relaxed" via the introduction of a set of *non-standard combinators* which allow for the insertion/deletion of specific words, the combination of discourse fragments or the correction of speech recognition errors.

The relaxed parser yields a (potentially large) set of parses, which are then packed and retrieved by the parse selection module. The parse selection is based on a discriminative model exploring a set of relevant semantic, syntactic, contextual and acoustic features extracted for each parse. The parameters of this model are estimated against an automatically generated corpus of ⟨utterance, logical form⟩ pairs. The learning algorithm is an perceptron, a simple albeit efficient technique for parameter estimation.

As forthcoming work, we shall examine the potential extension of our approach in new directions, such as the exploitation of parse selection for *incremental* scoring/pruning of the parse chart, the introduction of more refined contextual features, or the use of more sophisticated learning algorithms, such as Support Vector Machines.

6 Acknowledgements

This work was supported by the EU FP7 ICT Integrated Project "CogX" (FP7-ICT- 215181).

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

J. Baldridge and G.-J. M. Kruijff. 2002. Coupling CCG and hybrid logic dependency semantics. In *ACL'02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Philadelphia, PA. Association for Computational Linguistics.

T. Brick and M. Scheutz. 2007. Incremental natural language processing for HRI. In *Proceeding of the ACM/IEEE international conference on Human-Robot Interaction (HRI'07)*, pages 263 – 270.

J. Carroll and S. Oepen. 2005. High efficiency realization for a wide-coverage unification grammar. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP'05)*, pages 165–176.

S. Clark and J. R. Curran. 2003. Log-linear models for wide-coverage ccg parsing. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 97–104, Morristown, NJ, USA. Association for Computational Linguistics.

M. Collins and B. Roark. 2004. Incremental parsing with the perceptron algorithm. In *ACL '04: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, page 111, Morristown, NJ, USA. Association for Computational Linguistics.

M. Collins. 2004. Parameter estimation for statistical parsing models: theory and practice of distribution-free methods. In *New developments in parsing technology*, pages 19–55. Kluwer Academic Publishers.

R. Fernández and J. Ginzburg. 2002. A corpus study of non-sentential utterances in dialogue. *Traitement Automatique des Langues*, 43(2):12–43.

N. A. Hawes, A. Sloman, J. Wyatt, M. Zillich, H. Jacobsson, G.-J. M. Kruijff, M. Brenner, G. Berginc, and D. Skocaj. 2007. Towards an integrated robot with multiple cognitive functions. In *Proc. AAAI'07*, pages 1548–1553. AAAI Press.

H. Jacobsson, N. Hawes, G.-J. Kruijff, and J. Wyatt. 2008. Crossmodal content binding in information-processing architectures. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, The Netherlands, March 12–15.

P. Knoeferle and M.C. Crocker. 2006. The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*.

G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N.A. Hawes. 2007. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Language and Robots: Proceedings from the Symposium (LangRo'2007)*, pages 55–64, Aveiro, Portugal, December.

P. Lison and G.-J. M. Kruijff. 2008. Salience-driven contextual priming of speech recognition for human-robot interaction. In *Proceedings of the 18th European Conference on Artificial Intelligence*, Patras (Greece).

P. Lison. 2008. Robust processing of situated spoken dialogue. Master's thesis, Universität des Saarlandes, Saarbrücken. http://www.dfki.de/ plison/pubs/thesis/main.thesis.plison2008.pdf.

D. Roy and N. Mukherjee. 2005. Towards situated speech understanding: visual context priming of language models. *Computer Speech & Language*, 19(2):227–248, April.

D. Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

M. Steedman and J. Baldridge. 2009. Combinatory categorial grammar. In Robert Borsley and Kersti Börjars, editors, *Nontransformational Syntax: A Guide to Current Models*. Blackwell, Oxford.

E. A. Topp, H. Hüttenrauch, H.I. Christensen, and K. Severinson Eklundh. 2006. Bringing together human and robotic environment representations – a pilot study. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Beijing, China, October.

K. Weilhammer, M. N. Stuttle, and S. Young. 2006. Bootstrapping language models for dialogue systems. In *Proceedings of INTERSPEECH 2006*, Pittsburgh, PA.

L. S. Zettlemoyer and M. Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pages 658–666.

L. S. Zettlemoyer and M. Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687.